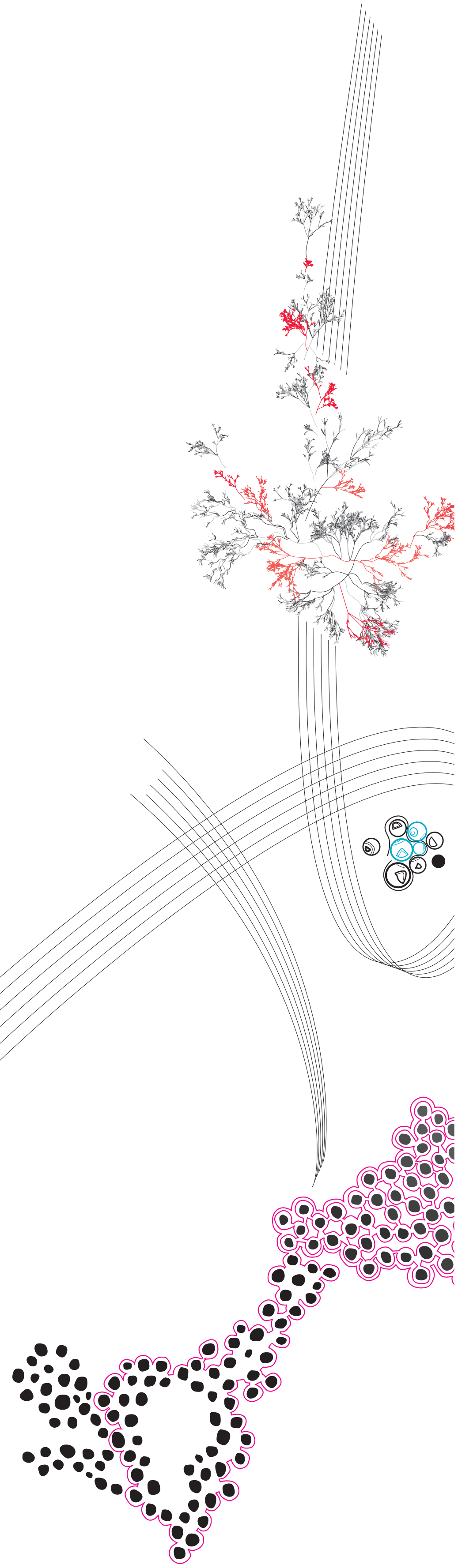
Msc Thesis Industrial Engineering and Management



Forecasting Transportation Volumes at Farm Trans

Hendrik Anton van Ramshorst

Supervisors: D.R.J. Prak and E.A. Lalla

August 22, 2022

Faculty of Behavioural, Management and Social Sciences (BMS)

UNIVERSITY OF TWENTE.

Preface

This master thesis is the result of the final part of my master study Industrial Engineering and Management at the University of Twente. I got the opportunity to do the internship at CAPE Groep and conduct the research at a client of theirs, Farm Trans, from the beginning of February 2022 to the end of August 2022.

Since this marks the end of my student life, I would like to take this opportunity to express my gratitude to everyone who has helped me during my research, and student life in general.

Starting with my first supervisor from the University Dennis Prak for all the time, support and feedback during the research. The feedback was always very helpful and this certainly helped me a lot. I also want to thank my second supervisor Eduardo Lalla for the new insights gained towards the end of the research. Besides the academic conversations, we also had nice chats about the personal life which I am grateful for.

Secondly, I want to thank the supervisors from both CAPE Groep and Farm Trans for not only providing me with the right information and feedback, but also for the fun time I have had. In particular I want to thank Youri Verveld for the supervision on the project and implementation in the end and Mart Busger op Vollenbroek for information about innovative aspects in the implementation phase. Furthermore I would like to thank Sander Kock and Jan-Marijn Vink from Farm Trans for all the input, feedback and supervision throughout the project. I have always felt welcome and free to ask for the information needed.

Lastly, I would like to thank all the graduate students at CAPE Groep. In particular Bouke Reitsma, Sven Koning ter Heege and Ivo Zwienenberg. We organised two-weekly meetings where we could discuss the projects and give each other feedback. Especially at the start of the research this was really helpful.

Kind regards,

Rick van Ramshorst

Management summary

This research for the master thesis is performed at a client of CAPE Groep: Farm Trans. Farm Trans is a transportation company in the agriculture sector specialised in bulk and conditioned transport. Farm Trans' business is growing with which more data and information becomes available and therefore also the need to use it increases. From this point of view the question originated how to get a better insight into the future demand, i.e., transportation volumes. This is important since the better this insight, the better the business can be adjusted to the requirements of the customers resulting in less costs and a better customer satisfaction.

The main research question is: How can Farm Trans predict the transportation volumes within the conditioned business unit with the available information? To answer this question, the research is divided into six parts. First the problem is identified. Then the current system is analysed and a literature study is performed to get acquainted with the situation and possible solutions. Then the solutions are being formulated and a final decision is made. Lastly the solution is implemented. The forecast of transportation volumes supplements the capacity decision making process which on the short term, 12 weeks, can be adjusted by hiring capacity. On this term a weekly forecast is requested. On the long term, 13 period or 1 year, new capacity may be acquired resulting in the need for a forecast per period. Next to that it is important to know in which region the transport is required, focus lies on the Benelux, Germany and the United Kingdom. These temporal and geographic aggregations add up to 8 forecast levels in total which opens the opportunity to use hierarchical forecasting methods. The current practise of fleet capacity planning is based on an estimation of the revenue, the budget. However since revenue is subject to change by for example fuel prices or inflation in general, a forecast of transportation volumes is desired to better supplement the decision of fleet capacity. Together with average volume transported per truck and trailer, an estimate of the required capacity is made.

After performing a literature study on what forecasting models and methods are suitable, how to tune these models and how to evaluate the performance, different models and methods are tested and evaluated. Based on their forecasting performance the best forecasting model for Farm Trans for each aggregation level is chosen. The goal of a model is to predict an outcome based on input. This input is mainly historical information of the demand itself but also the external factors: an indicator of covid-19, gross domestic product of the Netherlands and the budget created by Farm Trans are used as predictors. Furthermore, due to the hierarchical component of the data that we have available, we can use hierarchical forecasting methods to further increase accuracy. Many models are available to perform forecasting. Therefore, a selection is made based on the literature and current situation. This results in testing ARIMA, exponential smoothing and random forest methods. These models will be combined with the explanatory variables and the hierarchical methods and this makes up for the approaches that are tested in this research. The hyper parameters, input variables for model functioning, are optimized using a cross validation approach. The resulting models are tested on a test data set to assess performance on data the model has not yet seen before.

Overall the best individual model to use is ARIMA, for the most aggregation levels this provides the highest accuracy and otherwise the difference with the best model is small. For some aggregation levels using explanatory variables or hierarchy is positive for per-

formance. Especially for Germany when including the covid-19 indicator performance is greatly increased by 40% and 25% for the period and week respectively. By using hierarchy, earlier generated forecasts are exploited at other aggregation levels which may lead to increased accuracy. This is for example due to different demand patterns or external influences depending on the aggregation level which are then exploited at the other levels. This has a positive effect on the Benelux period and UK week forecasts. In Table 1, an overview of the best models per level with their performance is shown. This is the performance based on the test data which ranges from period 6 in 2021 to period 5 in 2022. Farm Trans achieved a MAPE of 10.5% with their own budget from 2018 to 2021 and 4.6% in the test data timeframe. The forecast accuracy is not better than the performance of the budget but it does give a clear view of what the business can expect namely in terms of transport volume instead of revenue. Also, the forecasts are aggregated among the 8 levels. The budget is not a forecast but a guideline for the business. Therefore, actions are taken to match the budget as closely as possible, especially when revenue is below budget, resulting in a better accuracy.

Aggregation	Best model	MAPE (%)
Global period	ARIMA	6.7
Global week	ARIMA	7.6
Benelux period	ARIMA hierarchy	4.6
Germany period	ARIMA variables	13.5
UK period	RF	13.1
Benelux week	ARIMA	6.5
Germany week	ARIMA variables	17.6
UK week	RF hierarchy	12.1

Table 1: Overview best models per aggregation

Implementation of this research is two fold. On the one hand a prototype of the best forecasting model is created for Farm Trans. This model is set up in a generic way, meaning that it can also be used by other business units of Farm Trans. Also, one of the goals of CAPE Groep is to be a strategic partner of their clients. This research contributes to that end since the model can be used for any time series to be forecasted and therefore it can be implemented at other clients. The forecast for Farm Trans is incorporated into their weekly reports. Together with historical information about average volumes, an estimation of the required capacity is made by Farm Trans. This is input for the capacity decision process, where first there were no data driven estimations, now there is a forecast and estimation on which the discussion is based. Together with explicit knowledge of the employees and the forecast a better decision can be made. Farm trans indicated that this can be a difficult process and that the forecast helps in creating more insight in the future demand.

Further research can be done one the fleet capacity planning problem. Where now we make a simple estimation of the required capacity, problem like the fleet size and composition vehicle routing problem (FSCVRP) can be solved to more accurate make a decision on the fleet capacity. Furthermore, the implementation can be further exploited by automating the process of generating forecasts each week and period. Next to that the model can be implemented for the other business units of Farm Trans. This does require that the configuration of the model is adjusted since a completely new time series is forecasted.

Contents

1 Introduction

This thesis is conducted for completion of the master Industrial Engineering and Management which is part of the faculty of Behavioral, Management and Social Sciences at the University of Twente. In this chapter an introduction is given, starting with background information about the organisation. Then in Section 1.2, the problem is identified using a problem cluster resulting in a core problem. Section 1.3 provides the research problem and goal following from the core problem. In the final section the research questions are formulated. These serve as an outline of the chapters of the thesis.

1.1 Background information

The research is done at a client of CAPE Groep. CAPE Groep is an IT consultancy firm who realises digital innovation and transformation for multiple customers, one of which is Farm Trans. Farm Trans is a transportation company in the agriculture sector and specialized mainly in bulk and conditioned food transport. Food transport needs to be safe, hygienic, reliable and fast. Farm Trans accomplishes this by innovative service, a modern and extensive fleet and years of experience in international food transport. They naturally opt for the most sustainable solutions. Bulk concerns the transport from farmers to food processing factories. From the factories most products need to be transported fresh and/or frozen (conditioned) to a distribution centre, this is the responsibility of the conditioned business unit. Other business units are Eastern Europe and Connected Services. The main office is located in Moerdijk, but there are also locations in Belgium and the UK. In total Farm Trans has around 220 trucks available for all business units. Farm Trans has one large customer which is responsible for around 60% of the yearly revenue. In total around 10% of the customers are responsible for 90% of the revenue. The request for this research originates from the analysis and control team of Farm Trans of which the focus is to analyse data according to the needs of the business.

1.1.1 Assignment from the organisation

Because Farm Trans' business is growing, there is more data and information available and therefore also a need to use it. From this point of view, the question originated what Farm Trans can do with this data. One of the areas is gaining insight in future demand. While they are getting better in analysing the past, predicting the future demand is difficult. The trends towards digitization continues but the transportation and logistics sector is relatively traditional. Around 72% of the companies in the logistic sector do not consider themselves as advanced in digitization (Daum et al., 2021) (PWC, 2017). The question from the organisation therefore is how to get insight into the future demand, i.e. transportation volumes, so that better operational decisions can be made. One the most prominent decisions is determining how much capacity is needed to fulfill demand.

1.1.2 Main process

The focus of the research is on fresh and frozen (F&F), also called conditioned, transport of Farm Trans. F&F is responsible for the Benelux, Germany and the UK. The order process starts when a customer requires transport. They send their transport order to Farm Trans, either directly into the control tower or to the customer service. This order consists of information about the product type, product quantity, load and unload locations and time windows. In the control tower, the orders are send to the transport management

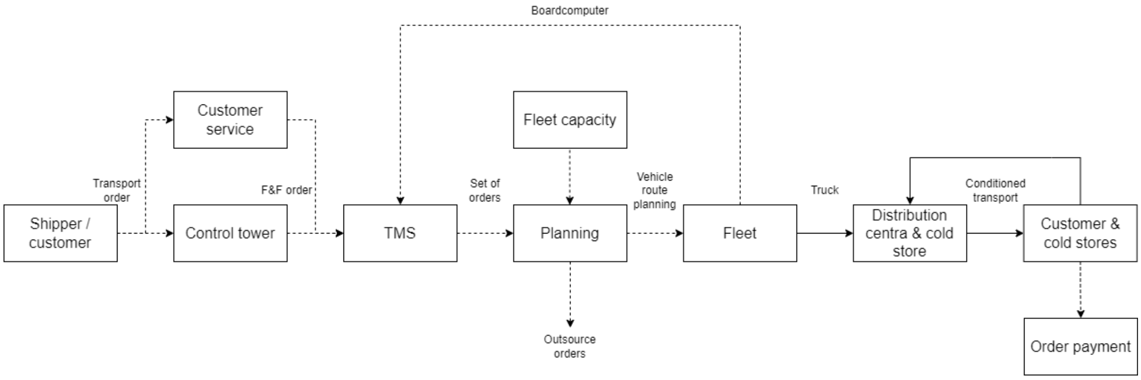


Figure 1: The flow of transportation orders for Farm Trans fresh and frozen

system (TMS) of the right carrier, i.e. business unit. The conditioned transport in the Benelux, Germany and the UK is send to F&F. About 40% comes directly from the control tower into the TMS, the remaining is handled via the customer service. Then the orders are sent to the planners who construct routes for the vehicles, thereby using the available fleet capacity. When there are not enough trucks available in-house, routes need to be outsourced. Next the truck drivers are informed to carry out the specified activities. These activities are done on daily bases, before 2 pm the orders for the next day may be sent. The process is summarized in Figure 1, where the dotted lines present a flow of information and the solid lines a material flow. Farm Trans has a couple of distribution centres. This means that a lot of orders need to be picked up and brought to those locations, from which they are distributed further. One of the largest is located in Lommel, which is responsible for about a third of the demand.

1.1.3 Capacity decision process

To be able to fulfill the demand, fleet capacity is required. Determining this fleet capacity is a prominent decision for Farm Trans where forecasts are required to support this process. Figure 2 shows the process of the capacity decision. The current way of forecasting is creating budgets, revenue expectations, done by the sales department. These are made each year and are based on historical sales data and are altered with the expected change this year, which is an estimate of the sales employees. They have contact with the customers to accomplish this. The budget gives a broad overview of the revenue expected per customer each period (4 weeks). This is then further divided into revenue achieved with own material and outsourcing and is based on historical percentages of the revenue. This process takes up a lot of time of the sales employees. It is determined by management that outsourcing around 35% is a good percentage in terms of the combination between profit and service. The revenue by own material is of importance for the decision about the fleet capacity. This is combined with the historical revenue per truck to determine the number of trucks needed. To determine the number of trailers needed a different process is done. We need at least one trailer for each truck, but extra trailers are for overseas transport, loading and unloading processes or chartering. Because of this there are around 3 times as many trailers as trucks.

Altering fleet size can be done on the medium and long term. Three periods in advance medium term capacity can be altered. One period is equal to 4 weeks, this is the time unit Farm Trans uses for reporting. These tactical decisions include for example chartering, i.e., the hiring of transport including driver, truck and/or trailer. This is a relatively long

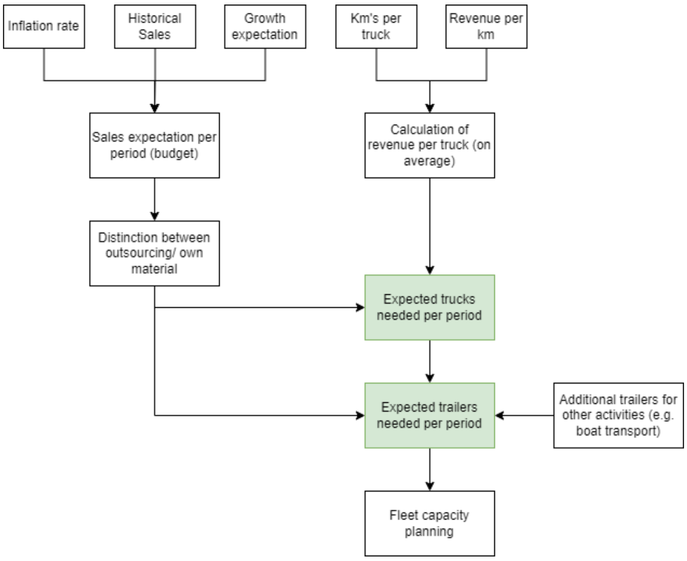

Inflation rate
Historical Sales
Growth expectation
Km's per truck
Revenue per km
Sales expectation per period (budget)
Calculation of revenue per truck (on average)
Distinction between outsourcing/ own material
Expected trucks needed per period
Expected trailers needed per period
Additional trailers for other activities (e.g. boat transport)
Fleet capacity planning

Figure 2: Process of capacity determination. The green boxes indicate the result of the process

period and is due to the current market situation where it is difficult to hire charters because of high demand. It is possible to alter fleet size on a short term in emergency situations but this is not desirable, orders are then rather outsourced. Buying trailers is a strategical decision and currently has a lead time of a year. This is also a relatively long period which is due to the computer chip shortages. Furthermore in the long term (12-24 months) a rough estimate can help the fleet planning department to account for possible large fluctuations in the future.

Currently the fleet capacity is determined by consensus between the sales and operational departments. Fleet planning then checks if that planning is doable and needs to assure that the requested capacity is available. When there is more demand than capacity, transportation is outsourced. First at the inter-company organisations, i.e., the other business units within Farm Trans. When there is also no capacity available, these orders are outsourced externally. Since internal capacity is used to create trips as efficiently as possible, orders which are not easily incorporated into the schedule are outsourced, this may even result in higher profits. It is, therefore, not always bad to outsource the demand to other carriers. There are, however, multiple reasons why to in-house transport, such as the quality of service and therefore customer relationships. For example, some customers request the same Farm Trans truck each time. This customer relation is important because Farm Trans has a relative small number of customers which is accountable for most of the revenue.

1.2 Problem identification

The fleet capacity decision making process can be improved by better knowing what to expect in the future, mainly because the current decisions are based on the knowledge of the employees and not on data driven forecasts. Therefore it is difficult to make well-founded decisions. There are weekly meetings to make these decisions which take up a large amount of time and cause discussion within the organisation on what to decide. Having a data driven estimate of future demand can help in making the capacity decision. It also provides information to feed the discussion and arise awareness on the problem. By knowing 12 weeks up front if demand is expected to change, the capacity can be adjusted accordingly by chartering. Acquiring new capacity has a lead time of a year and is therefore seen as a long term decision. There is little knowledge available within Farm Trans on how to forecast their demand which is the cause of this research.

In a business-to-business environment, forecasting future demand is quite crucial as the entire production and supply process depends on these forecasts (Rohaan et al., 2022). Currently it can happen that demand arises without having the capacity to fulfil it, or that not enough demand arises as was anticipated on. It is then possible to outsource the orders, but when demand is high outsourcing becomes more expensive due to market forces and there may not be enough capacity on the market to fulfill all demand. Therefore management has set a target that around 65% of the demand should be fulfilled by own capacity. This number has proven to be effective in recent years, but the optimal balance between in- and outsourcing is unknown. The trade-off is between profit and customer service/ quality.

Furthermore, orders arrive very last-moment, i.e. orders come in today for tomorrow. The business is agile for this reason. If it is busy we want to outsource the orders that take a long time to complete. But if it is not busy we can do the long orders ourselves so that capacity is fully utilized. In this time span capacity cannot be altered. We therefore do not focus on this short term horizon. It is also noticed that some general knowledge about demand patterns is missing within the organisation. The budgets that are available are not communicated throughout the organisation. It has happened that during busy weeks some drivers where on vacation for example. It could also be the case that standard days in the week or year have a high demand. In the past, little research was conducted on the forecasting topic but no major steps were made on this subject, therefore no previous knowledge is available. The problems described above are summarised in a problem cluster shown in Figure 3 using the methodology from Heerkens and van Winden (2012).

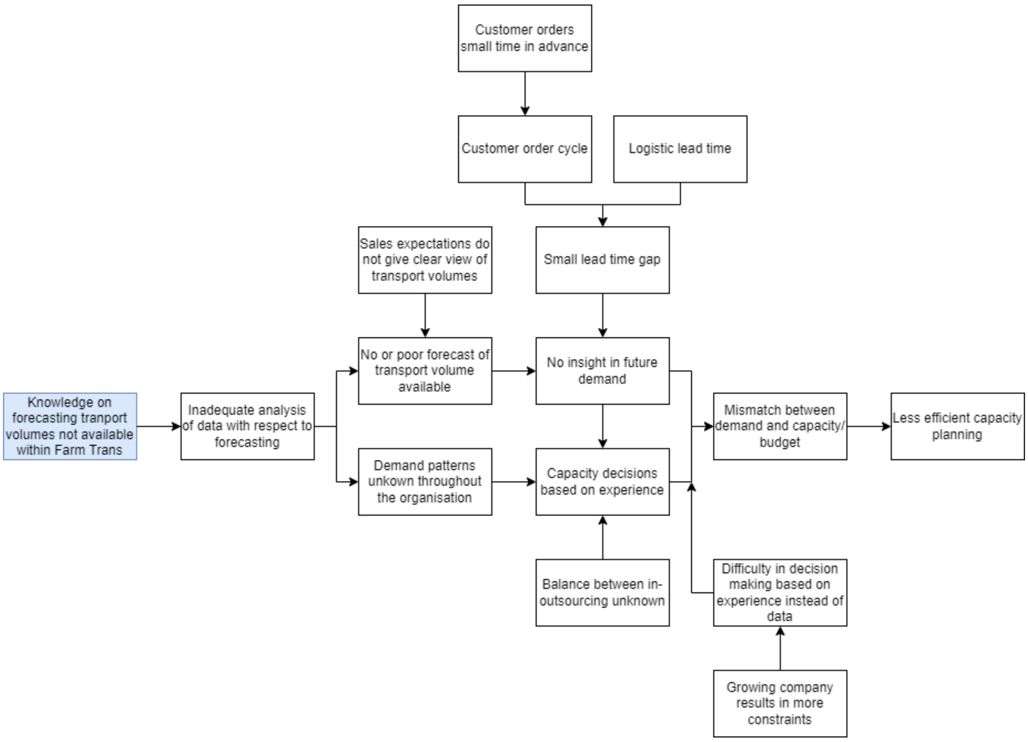

Customer orders small time in advance
Customer order cycle
Logistic lead time
Sales expectations do not give clear view of transport volumes
Small lead time gap
Knowledge on forecasting tranport volumes not available within Farm Trans
Inadequate analysis of data with respect to forecasting
No or poor forecast of transport volume available
No insight in future demand
Mismatch between demand and capacity/ budget
Less efficient capacity planning
Demand patterns unkown throughout the organisation
Capacity decisions based on experience
Balance between in- outsourcing unknown
Difficulty in decision making based on experience instead of data
Growing company results in more constraints

Figure 3: Problem cluster

1.2.1 Research motivation

Finding the right balance between customer requirements and the capabilities of a supply chain is referred to as demand management. With the right process, management can proactively match supply and demand such that that a company can be proactive to anticipated demand and more reactive to unanticipated demand. This process is not limited to forecasting. Increasing flexibility and reducing variability is also part of the process. The goal of demand management is to meet customer demand in the most effective way (Croxton et al., 2002).

We however focus on forecasting which is about predicting the future as accurately as possible. According to Profillidis and Botzoris (2019), transport demand forecasts are a prerequisite for almost all activities related to transport. Any decision related to planning, investment and operation of infrastructure and transport modes can benefit from an as accurate forecast as possible. Any operation of a transport service company which is not based on the most accurate forecast runs the risk to turn into an economic adventure or disaster. For example the fleet of a transport company is a direct consequence of the future demand to be served. Therefore an increase in forecasting accuracy leads to improvement on the business performance.

Matching demand forecast to the supply chain capabilities is referred to as sales and operations planning (Croxton et al., 2002). Creating more insight into future demand affects the decision making of Farm Trans in multiple ways. It creates general knowledge within the organisation on when to expect high peaks in demand. In these periods all possible capacity must be available. This needs to be incorporated in for example vacations of employees and the maintenance of trailers. One of the main usages of the forecast is determining the fleet capacity needed. Capacity reflects to the amount of trailers and trucks which is called operation of infrastructure by Profillidis and Botzoris (2019). A forecast based on volumes and having concrete, substantiated numbers can be used to improve on this process such that a better estimate can be made. Furthermore, forecasts also have an impact on the sales department. They monitor the current sales with respect to the goal but they don't know how this reflects throughout the year. If the sales are behind on schedule and the forecast does not make up for it, acquisitions can be made. To summarise, there is a general need for a more accurate forecast of future transportation volumes throughout the organization.

1.3 Research problem

Following the problem cluster of Figure 3, we define the core problem to be: *There is no knowledge available within Farm Trans on how to use available data to gain insight in future demand.* The focus lies on gaining and using the knowledge in making forecasts because this has the highest impact on the problems identified. Currently orders arrive today for tomorrow. Bringing the order moment forward would not have an impact on determining fleet capacity because this time span would be to short, as described earlier. We therefore do not focus on that part of the problem.

1.3.1 Research question

The main research question is defined as follows: *How can Farm Trans predict the transportation volumes within the conditioned business unit with the available information?* The goal of the forecast is to create a better match between demand and capacity by creating forecasts such that decision making improves. We decided to focus on the F&F business unit, instead of for example bulk, because the demand in this segment is less vulnerable for external factors, like the weather, which is quite unpredictable. Also, since no earlier research is done on the subject of forecasting and demand management we narrowed the scope. In this way we acquire knowledge and test if this approach is applicable for Farm Trans, before rolling out the project over the entire company. A wish is to develop a general approach which could be applied to the other business units of Farm Trans or other customers of CAPE Groep.

1.3.2 Scope

The scope of the forecast is to predict the transportation volumes per week, for 12 weeks ahead of Farm Trans F&F. Because the business is agile, there is less need for knowing how much demand arises the next day or week. Capacity for this short term is already determined and excess demand can, most of the time, be outsourced. It is therefore of more importance to get the right balance between in and outsourcing by determining fleet capacity on the tactical and strategical level. Fleet capacity can be adjusted 12 weeks in front by chartering. Internal capacity needs to be ordered a year upfront, therefore this time period is also taken into account in generating forecasts. However for this period we do not need a forecast on week basis, a forecast for each period is sufficient to determine long term fleet capacity. The transportation volumes determine the number of trucks/ trailers needed by combining it with average volume transported per truck and trailer. A second level of aggregation, besides the week and period, is the region in which the capacity is needed, and thus forecasting information is necessary. These regions are the Benelux, Germany and the United Kingdom. The regions are based on the final destination of transport, where most transport originates in the Netherlands. Information about capacity usage differs in these regions which is why it is important to separate them. Dealing with these levels of hierarchy may be challenging but also results in extra opportunities for model development as hierarchical forecasting methods can be used. in Table 2 an overview of the decisions per aggregation level is shown.

Table 2: Forecast aggregation levels and decisions

Forecast	Decision 12 weeks ahead	Decision more than 12 weeks ahead	Decision 1 year ahead
Global per period	-	Chartering	Purchase of new capacity
Global per week	Chartering	Chartering	Purchase of new capacity
Region per period	-	Chartering per region	Purchase of new capacity per region
Region per week	Chartering per region	Chartering per region	Purchase of new capacity per region

In Section 1.4.3 an approach for generating the forecast is given. The forecasts is be updated each week, since the closer to the current time unit the more accurate predictions can be made and a new forecast for the 12 weeks ahead must be made. The individual customer level is not taken into account because this does not have a direct influence on the capacity needed. Customers with a steady and predictable demand are already identified by Farm Trans.

1.4 Research approach

To answer the main research question, multiple sub questions are formulated. These are divided in separate parts which act as chapters throughout the thesis. Answering these questions leads to answering the main question and thereby solving the core problem.

1.4.1 Current system analysis

First, it is important to understand the current situation of Farm Trans. Before we can start building a model it is important to exactly know what the requirements of the forecast are. Also according to Profillidis and Botzoris (2019), the first step is to clearly understand which are the factors or driving forces that affect the transport demand. Also, the first steps of the CRISP-DM, a process model for data mining projects, are to understand the business and the data (Schröer et al., 2021). Next to that, we are interested in the model requirements. The research questions are as follows:

- What is the current forecasting practise and how are forecasts used in the decision making process?
- What data is available and usable for model development?
- What variables influence the demand?
- What are requirements for the forecasting model?

1.4.2 Literature study

Knowledge needs to be gained in order to build a forecasting model in practise. We need to search for forecasting models in literature such that we can decide upon which is best applicable for Farm Trans. Here we mainly look for models which had a good result in predicting demand, i.e. transportation volumes. It is also important to know how a model can be evaluated so that we can compare the different models and eventually choose the best. Some models require a decision on which hyper-parameters to use, how the these parameters can be tuned is also investigated. Finally since we are dealing with hierarchy in our data, literature on hierarchical forecasting is explored. Therefore we formulated the following research questions:

- What forecasting methods are available to forecast demand in the transportation segment?
- How can the forecasting model performance be evaluated?
- How can the forecasting model hyper-parameters be tuned?
- What hierarchical forecasting methods may be beneficial?

1.4.3 Formulating solutions

When we know the current situation of Farm Trans and what theory is usable, we can start with formulating different solutions, i.e. building forecasting models. These are tested and evaluated. The best model is selected to be used as final approach. In total there are 4 levels of forecasting which are being considered, this is due to the aggregation levels at which the decisions are taken. According to Zotteri et al. (2005), the choice of the appropriate level of aggregation depends on the decision making process the forecast is expected to support. The forecasts are developed in the same top down order. It is summarized in Table 2, where a period refers to 4 weeks. In Chapter 4 the forecast design approach is presented. The following research questions will be answered:

- Which forecasting method, or approach, is best applicable at Farm Trans?
- What is the performance of the forecasting models tested?

1.4.4 Decision

Lastly, we need to decide upon what forecast method to use and how to use it. The impact of the new forecast is discussed as is the way it is incorporated into the decision making process. Farm Trans needs to be able to use the tool themselves and therefore at least a working prototype needs to be developed. This prototype needs to fulfill the set model requirements. The research questions are defined as follows:

- What forecasting method is best for predicting demand?
- How can Farm Trans utilize the forecast and what will be the benefit?
- How should decision making processes be altered such that the forecast is incorporated?

1.4.5 Research model

As guideline for the project, first a project plan is set up consisting of the problem identification and research questions, this is discussed in Chapter 1. The research questions serve as the main chapters of the thesis and lead to solving the core problem. They are divided into four project phases being: identifying the current situation, literature study, formulating solutions and making a decision and implementation. These phases are based on the managerial problem solving method (Heerkens and van Winden, 2012) as well as the CRISP-DM model. When the project plan is approved by the stakeholders, we can begin with answering the research questions and thereby solving the core problem. A general overview of the activities is shown in Figure 4. This research model shows in what order activities need to be done and how the knowledge and information gained from each activity is used further upon the project. With the end goal to improve the decision making processes by creating a forecasting tool such that more insight into future demand is gained.

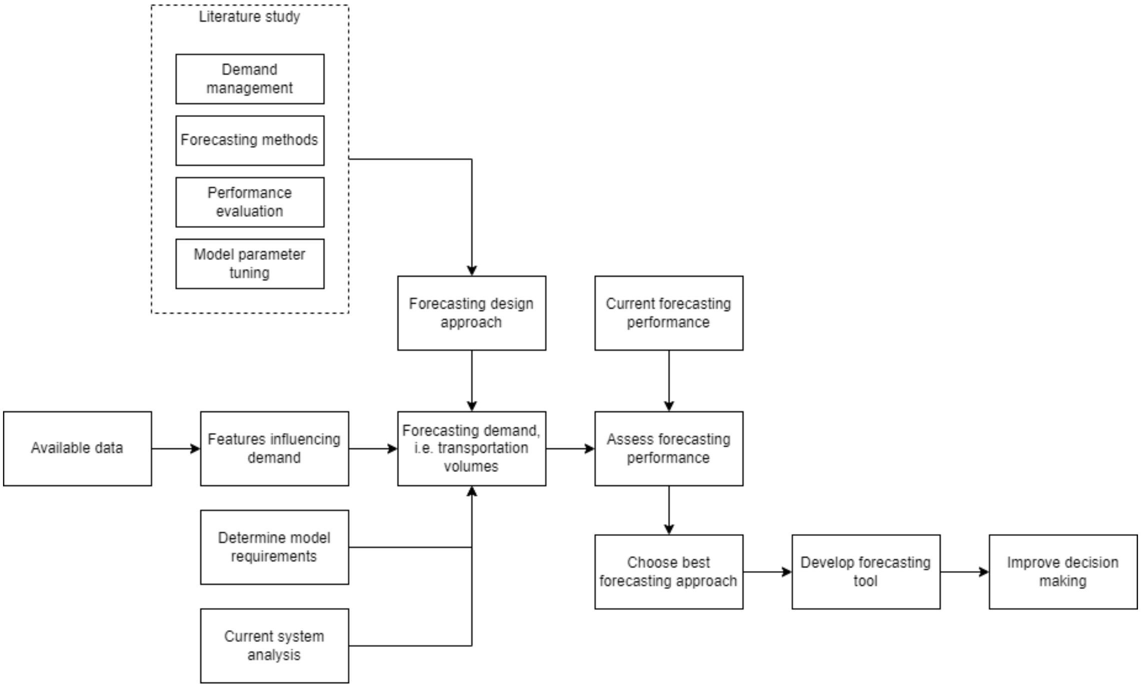

Literature study
Demand management
Forecasting methods
Performance evaluation
Model parameter tuning
Forecasting design approach
Current forecasting performance
Available data
Features influencing demand
Forecasting demand, i.e. transportation volumes
Assess forecasting performance
Determine model requirements
Choose best forecasting approach
Develop forecasting tool
Improve decision making
Current system analysis

Figure 4: Research model

2 Current system analysis

In this chapter, the current system is analysed. First some more information about the demand and capacity of the fresh and frozen department is given. To get acquainted with the demand, the patterns are analyzed. We are looking for seasonality or trends in the data, which have an impact on what forecasting models to use. We also look for events that have an impact on changes in demand. Then in Section 2.2 the current forecasting process of Farm Trans is explained. What data is available to develop a forecasting model is described in Section 2.3. Lastly, the requirements of the model are discussed in Section 2.4.

2.1 Demand and capacity

The research is focused on the fresh and frozen business of Farm Trans. In total they own around 45 trucks and 100 trailers, but this varies throughout the year. The biggest industry in which fresh and frozen is active is food service, for example frozen fries, the second biggest is retail. This can have an impact on the demand since in food service on average the first quarter of the year has less demand than the others (cbs, 2022*b*). We take a look at the year 2021 and give an overview of the performance and capacity used to fulfil the demand. In Table 3 we see the capacity used in 2021. Here we only state the capacity that is fulfilled by Farm Trans F&F themselves, so no outsourcing. Regarding the demand and revenue, it can be noticed that the revenue over volume ratio increases throughout the year, not shown in the table due to confidentiality. This means that more money is earned for every truckload. This is explained by the increase of fuel prices throughout the year, which is one of the main reasons why revenue is not straightforward to use when deciding on fleet capacity, as is the current practise. The fuel costs are charged to the clients and therefore counted as revenue. This can be seen as a trend but since we focus on forecasting transportation volumes, we do not take this into account. When we look at the capacity we can see that this differs throughout the year. An increase in truckload also leads to an increase in trailers used. The number of trucks is more stable throughout the year because this is a long term decision and cannot easily be altered on the short term. On average in 2021 the volume we transported per truck in one period slowly decreases throughout the year. This can be explained by the impact of covid-19 in the last quarter of 2021 where there was less demand due to the lockdowns and also the increase of trucks in general.

In Figure 5, an overview of the demand throughout the year 2021 is shown. In total 404,771 meters of load are fulfilled in 2020 compared to 373,572 in 2021, which is a 7.7% decrease. It can be noticed that demand is not stable throughout the year. For example, in 2020 there is a high demand in January, February and March after which the demand declines. This is explained by the lockdowns due to covid-19. This has a large impact since a part of the customers of Farm Trans supplies food to the catering industry, and these were closed during the lockdown. It is therefore important to determine how to use the available data with regards to model development and how to deal with the influence of covid-19. This is explained in Sections 2.3 and 3.7.

2.1.1 Demand patterns

With demand patterns we mean the existence of a trend or seasonality in the data. A trend is a constant increase or decrease of the demand over time. Seasonality occurs if in a certain period (day, week or month) there is always more or less demand than average. If there is a trend or seasonal pattern present, it is said that the time series is not stationary. We

Table 3: An overview of the capacity per period in 2021

Period	Trailers	Trucks
1	99	38
2	87	39
3	102	39
4	99	37
5	78	37
6	81	38
7	79	38
8	81	43
9	81	41
10	118	45
11	143	47
12	109	45
13	87	45

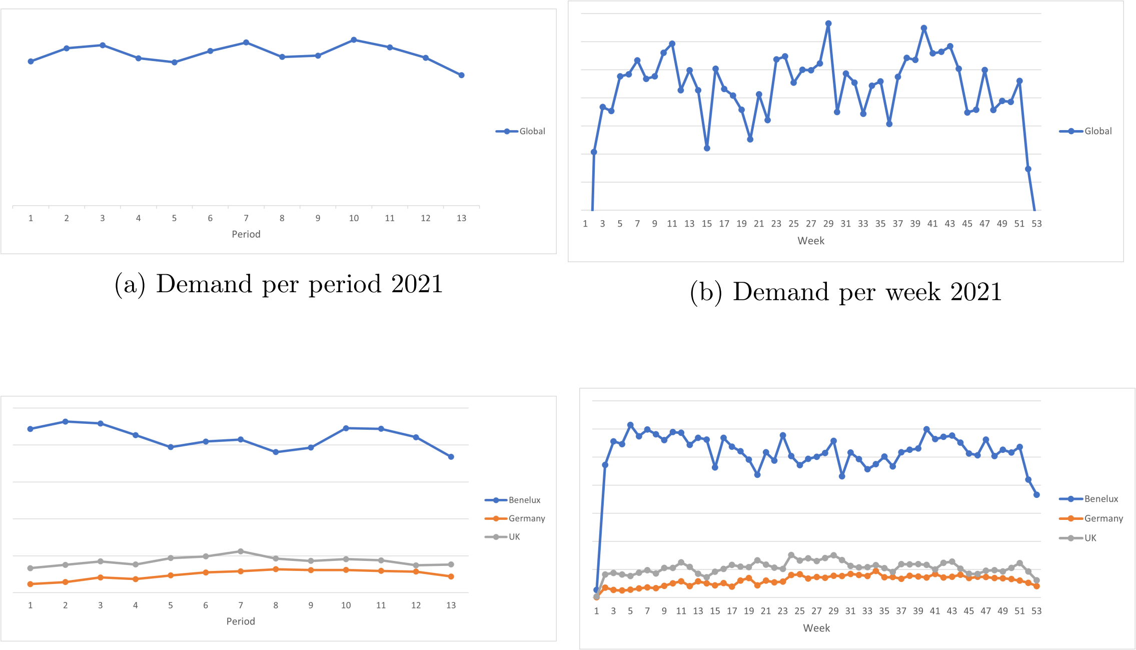
can test stationarity with a Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski et al., 1992). Here the null-hypothesis is that the data are stationary and we search for evidence that this is false. The test is performed over the data of 2018 until 2021, the results are summarised in Table 4. The null hypothesis is: the data is stationary around a constant. We test this against a significance level of 0.05.

Table 4: p-values as result of a KPSS test

	2018-2022	2018-2020	2020-2022
Period	0.039	0.061	0.1
Period Benelux	0.1	0.1	0.1
Period UK	0.01	0.1	0.1
Period GER	0.1	0.1	0.1
Week	0.01	0.053	0.084
Week Benelux	0.1	0.1	0.0378
Week UK	0.01	0.1	0.1
Week GER	0.1	0.1	0.049

When the p-value is lower than our significance level, we reject our null hypothesis. This is the case for the time series aggregated over the period, period UK, week and week UK. Therefore we can say that those time series are not stationary and include some trend or seasonality (Hyndman and Athanasopoulos, 2018). It may be interesting to separate the data in two parts, before and after 2020. Because from 2020 onward covid-19 has had an impact on the demand, which can cause the test to detect non-stationarity. These results are also shown in Table 4. We see that for example per period there is a stationary time series when we separate the data, so the result changes. In the UK there is a sudden increase in demand in 2020, which can be the reason for the non-stationarity. This is also solved by separating the data set. This information can be taken into account when developing the forecasting models since some models require stationary data to perform well. In most cases we are dealing with stationary data.

Next, we take a closer look at the demand in 2021 specified over all aggregation levels (period, week, global and regional). In Figure 5 the demand in 2021 of these different aggregation levels is shown. This is the demand for which we want to generate a forecast.



(a) Demand per period 2021

(b) Demand per week 2021

(c) Demand per period and region 2021

(d) Demand per week and region 2021

Figure 5: Demand aggregation levels

It is interesting how the hierarchy can be used in generating the forecast later on. According to the graphs there is no trend present in the data, at least not in 2021. In 2019 we see a consistent increase in truckload throughout the year. The question is if this is halted due to covid-19, or that the business has become more stable in general. The goal of the sales department is to increase the demand and therefore it can be said that a trend may be present and the growth is halted due to covid-19.

To determine if there is seasonality we need to look at multiple years of data. The average demand over the years 2018 until 2022 is higher in week 10 and 11, which fall in period 3, than in the rest of the year (about 107% compared to the average per period). Also we see that in the first period of the year there is the least demand, about 85% of the average demand per period. Therefore we can say that there is some seasonality present in the time series, but it is little.

When we decompose the time series into a trend-cycle, seasonal and remainder component we can check whether there exists a trend or seasonal pattern. We use STL (Seasonal and Trend decomposition using Loess) decomposition to decompose the time series as suggested by Hyndman and Athanasopoulos (2018). In Figure 6 the result is shown. We can again see the clear trend in 2019, and the decrease in 2020 due to covid-19. What is more interesting is the seasonal component. This varies a bit throughout the year, but on average the seasonal component is the highest in the third period and the lowest in the first period. Which is the same conclusion as when we looked at the average demand.

Looking at the daily demand throughout the first month of January 2022, it is noticed that demand on Wednesday is always higher than on Tuesday. On average the demand on Monday and Tuesday is around 1050 meters, where on the other days it is around 1200. So towards the weekend it becomes a bit more busy. This however does not have an impact on our forecasting process since we do not consider demand per day, but for the organisation this may be important to take into account in their daily activities.

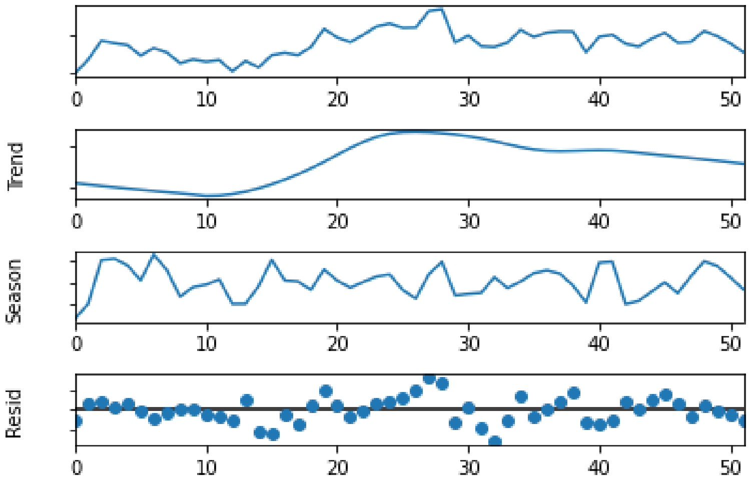


Figure 6: Time series periodly decomposition using STL

2.1.2 External influences

Transport demand is said to be derived demand, i.e., it depends on the need of service of other organisations. Therefore, there may be other factors, next to the historical demand itself, that have an influence or predictive value on the demand that we want to forecast. These variables can also be taken into account when creating forecasts to increase the performance. According to van Wee and Annema (2012), the gross domestic product (GDP) has a strong explanatory power for freight transport. The transport sector is responsible for 3.9% of the GDP in the Netherlands and therefore a growth in GDP implies a growth in transportation volume. It may be expected that when this increases, the demand at Farm Trans also increases. Since the Benelux, and mainly the Netherlands, is responsible for the most part of the demand, the GDP of the Netherlands is taken into account. Data is gathered from cbs (2022*c*), representing the GDP per quarter. Unfortunately GDP per period or week is not available. Since our timeframe is periods and weeks, the GDP per quarter is copied and used for 13 weeks and 3 periods of that quarter. Since we have 13 periods and 53 weeks, the last quarter of the year is used for the last 14 weeks and 4 periods. This does not exactly match the start and end of each timeframe, but it gives a good indication if the GDP has any predictive power. The GDP of the Netherlands is also used for the UK and Germany since most transport originates in the Netherlands. In order to generate a forecast, the expected growth of the GDP is used (cbs, 2022*a*). This is a growth measure year on year of 3.6% so this is multiplied by the GDP of the quarter of the previous year. Furthermore, we have seen that the lockdowns due to covid-19 have had an impact on the transportation volumes. This was mainly due to closing of the catering industry. To cope with this an indicator is generated, showing whether there were (1) no covid measures, (2) limited opening measures or (3) only food deliveries possible (Rijksoverheid, 2022). This is also the information from the Netherlands. Regarding the future it is currently assumed that no more lockdowns will be present. Whether this is expected or not is not predictable but this indicator can help to explain certain drops in demand. Third, the budget of Farm Trans is taken into account. This variable includes information from the sales employees like expected new customers and this may be a good indicator of transport volume. Since the accuracy of the budget is accurate to around 10%, we make use of this to improve accuracy. The budget can also be used to manually give direction to the forecast. Budgets are created for each period and to get the budget indicator per week we copy the information per period for the 4 representing weeks. Another, indirect, influence may be the weather. A large fraction of the customers are related to the catering industry which has higher demand when temperatures are higher. This can indicate an increase in demand for Farm Trans, but it is not taken into account since weather forecasts are not accurate more than a week in the future. Therefore this information can not be used to generate forecasts.

2.2 Current forecasting

Currently almost none forecasting activities are performed at Farm Trans. The only available prediction of the future is the expected revenue, also called budgets. The sales department develops these budgets at the beginning of the year. For each customer a sales expectation is developed based on the past year and/or a forecast from the customer itself. These are adjusted for expected growth and inflation rates which is done based on employee experience and information gained from the customer. Also new, potential, customers are taken into account. Sales employees share information about when and how much demand they expect from certain customers. This yearly information is then divided over the weeks, each customer is expected to request their orders in a different period due to the seasonality of some customer segment. Fries and ice cream have a higher demand in the summer periods for example. The result is a general expectation of revenues in each week which is expressed in euros. Throughout the year these are modified when the expectations change. This process depends heavily on the experience of the employees which is not desirable since the expectations can differ between employees and this may lead to discussions within the teams. It is also a time consuming activity. The main use of the budgets is to control the business, i.e., compare the budget to the realised revenue, and determine what actions need to be performed.

In Figure 7 the budget, revenue and truckload of Farm Trans are plotted. The exact numbers are left out due to confidentiality. The revenue and budget roughly follow the same pattern, especially towards the end of 2022. The accuracy of the budget is therefore considered high, on average the budget has an error of 10.5% over the real revenue. However, the budget is also something which is used as guideline for the business. If they are not on track to meet the budget, extra actions are taken to improve the sales. Therefore, we cannot directly compare it to the forecasts that are generated in this research but it gives an indication of the current performance of the budget activities. Therefore, if the forecast of the transport volumes achieves an error of around 10%, we achieve the same performance. It, however, has the addition that transport volumes are easier to translate into business requirements using average transported volume per truck and trailer. The budgets are more difficult to translate to required fleet capacity. For example, the budget is at its highest in the last period of 2022 while the truckload did not exceed the peak in 2020. We can therefore conclude that more revenue is made with less truckload being transported. With the current practise of determining fleet capacity, using average revenue per truck, this leads to an increase in trucks while not necessarily more truckload is expected. Therefore this approach is not viable in the future which is why truckload forecasting, and therefore this research, becomes more important.

2.2.1 From transportation volumes to fleet capacity

The main usage of the forecast is to support the fleet capacity decision process. However, to translate the forecast of transportation volumes to required capacity is not the main goal of this research. Therefore, an indication of the required trucks and trailers is given

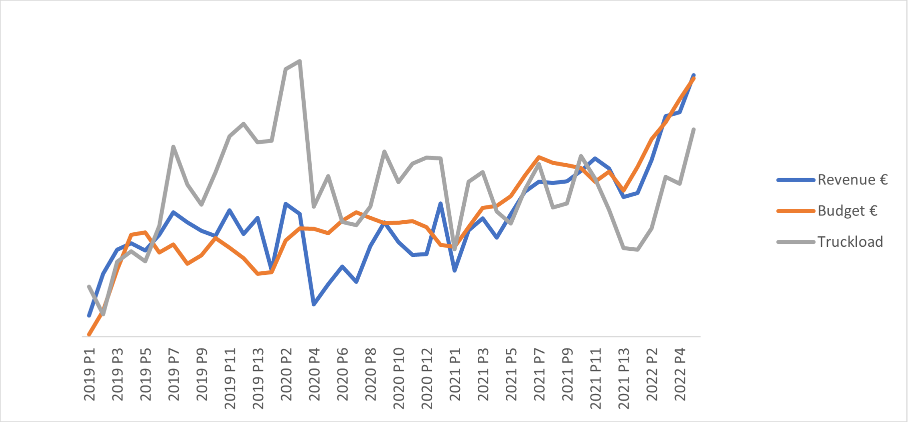


Figure 7: Budget, revenue and truckload per period of Farm Trans fresh and frozen from 2021 to 2022

based on the historical data. The average volume transported per truck and trailer is calculated per aggregation level, that is per period, week and region. We only take into account the trucks and trailers which are owned by Farm Trans fresh and frozen themselves for this calculation. For the UK and Germany, most of the times the transport is done by other carriers or hired transport. Therefore, for those regions we take into account the number of unique trucks and trailers used by all creditors. In reality the unique number of trucks used may be a bit lower because when outsourcing different trucks can be used for each transport and this is not taken into account in the data available. The average volumes per truck and trailer will be higher in reality because of this. However, the priority lies on the global capacity since this is related to the decision about purchasing capacity for the whole department. With this approach we achieve the same level of accuracy as is currently achieved with the decision making process using the budgets and average revenue per truck.

To get a more accurate representation of the required capacity a more thorough analysis can be made by calculating average truckloads and turnover rates. Alternatively, research can be done to optimal route planning by including the number of trucks and trailers required to fulfill demand in a certain period or week. For example, simulation methods can be used to determine the capacity required to fulfill a certain demand. This is out of scope for this research.

The exact values are left out due to confidentiality. We calculated the average truckload per week and period and then averaged this over the entire year to get an estimate. The forecast of volume to transport can then by divided by this number to get the approximation of the total capacity required. It is then up to the management to decide on how to fill in the required capacity, i.e., purchase, hire or outsourcing. In Section 2.4 it is further elaborated what the impact is of forecasting accuracy on capacity decision.

2.3 Available data

To develop the forecasting model, we make use of the available data. Farm Trans makes use of a transport management system (TMS), I-Teq, which is designed for the logistic sector. In this system the main activities like the planning are performed. The information is stored in a data warehouse which we can access through the use of SQL (Structured Query Language). We can assume the data is reliable since the same source is used to generate weekly reports. However since we import the data and modify it, it is important to check if we are using the correct numbers. To ensure that we did the correct modifications, the information used in the forecasts is checked with the information from the TMS and the weekly reports.

In the research we want to make a prediction about the transportation volumes. Therefore this is the most important data to collect. The demand can be found back in a log consisting of all loads, also called order lines, that have been transported. It may happen that multiple loads are part of one order and multiple orders may be part of a route, depending on order size. Each order line consists of 53 columns of information. The most important information is the number of pallets, weight, load and unload location with date, time windows and corresponding identification numbers for the order and route. This however is not the information which can be used to explain the demand, it is needed to perform the transport and the information is known when the order has already arrived. We can therefore not use this data as explanatory variables in our forecasting model.

The data has to be modified in order to fit our needs. For example, to create a weekly forecast we require a time series considering the weeks. This also yields for the period and whether we consider the demand in a region or globally. A period is equal to 4 weeks. At the start and end of a year the number of days in a week and period can differ a bit since there are not exactly 52 weeks in one year. The period at the start of the year 2022 contains 29 days instead of 28 for example. When this is the case the demand is normalized such that all periods contain the same number of days, 28. This is done to create consistency throughout the year. Since most of the data is entered into the system through the customer service, it can contain human errors. For example the maximum truckload is 13.6 meters but sometimes a higher truckload is entered. These data points are removed from the data set as they are considered as outliers. We only consider orderlines with a truckload between 0 and 13.6 meters.

To predict the transport volumes historical data may be used, but also other information can be used to make predictions. For example the sales budgets of each year are also available. This can be used as an explanatory variable for predicting the transport volume. Using the budget may be beneficial since this is generated based on sales expectations, which also include new customers. It needs to be investigated if adding these variables to a model has a positive impact on the forecasting accuracy.

In chapter 1 we have seen the demand throughout 2020 and we recognized a significant decrease due to covid-19. This had a significant impact on the transportation volumes which we want to forecast. At the moment of writing there are no more restrictions regarding covid-19 and therefore it can be questioned whether the historical data from 2020 and 2021 can accurately represent the current demand. This are however factors which are unpredictable and therefore we first work with the actual numbers. It may be possible to include the impact of covid-19 as a explanatory variable, but then we still have the problem that this is unpredictable in the future.

2.4 Model requirements

The requirements for the model are related to the scope of the research and the forecasting levels to be developed. The goal is to forecast on a weekly basis, 12 weeks ahead and on period basis one year ahead. This is a rolling forecast, so it is updated each week. Since in general the forecast becomes more accurate the closer it is to the current date this implementation is chosen. First the transportation volumes of the whole fresh and frozen department is forecast on period basis, then on weekly basis. A further distinction is then made over the regions (Benelux, Germany and UK). These four levels of aggregation are linked to the decisions that can be made on short and long term concerning the fleet capacity. This hierarchy is very important since the capacity needed to fulfill demand varies per region. When transporting to the UK from the Netherlands, more capacity is needed due to the use of boat transport, resulting in a longer turnover rate per truck/ trailer. Also, the trailers are put on the boats and picked up by a truck in the UK. This means that there are two trucks needed for the same order where in the Benelux or Germany just one truck is needed. Another important requirement for the forecast is that the input and output is verified. Garbage in is garbage out is a well known phrase within model development. Without the right inputs the model will not perform well and it is therefore of at most importance that the data accurately represents the reality. This is described in Section 2.3. Also the output of the model need to be tested so that we know the accuracy. More information about this is discussed in Section 3.6.

Next to that, the forecast needs to be incorporated into the weekly report of Farm Trans. This is important because these reports are discussed each week and form the basis of the decision making. Also the forecasts are seen by the employees, creating awareness at the employees. The reports are developed through the use of Microsoft Power BI which in turn uses information from the SQL database. Farm Trans needs to be able to access and run the forecast model themselves. Therefore the generated forecasts need to be easily accessible, for example through uploading it into the data warehouse. This is under the condition that assess to the database is granted. Otherwise the forecast are generated locally and it needs to be updated manually. Also the model needs to be run weekly to update the forecasts. Therefore some knowledge needs to be transferred, but more importantly the model needs to be easy to use.

Regarding forecasting performance the goal is to, closely, match the current accuracy of the budget compared to the revenue which achieved a mean absolute percentage error, MAPE, of 10.5%. The MAPE and other performance metrics are introduced in Section 3.6. However, according to Farm Trans it is not a hard requirement to match the performance of the budget. Ofcourse, the lower the MAPE, the better, but the capacity decision does not boil down to a truck more or less. It is however important to achieve an accurate forecast since otherwise to much capacity will be acquired which results in a lower profit for the business. Using the average volume transported we can determine the impact of a certain error. Since the transport volume has an impact on this we use the average volume per period in 2021. In that situation an average error of 10% is equivalent to 10 trucks and 24 trailers. For the average weekly demand it boils down to 9 trucks and 13 trailers, the numbers are rounded to the nearest integer. Therefore, an increase in error of one percent means one extra truck and two extra trailers are needed on the period term. Assuming that 65% of demand is fulfilled in house, as this is the general factor determined by management, this means that for every 1.5% percent that our forecast improves one truck and 2.3 trailers less needs to be acquired. The global period forecast is of most importance since this determines the total capacity needed for the operation.

Not a hard requirement, but a more of a wish, is to generalize the forecasting model. This can be of importance because other departments within Farm Trans could then also use the model. New models need to be trained for this and the accuracy needs to be assessed before it can be used. In this fashion, also CAPE Groep can make use the forecast model and the knowledge gained during this research is not lost. Since they are involved in the development of it related systems at Farm Trans they can incorporate it if their is a need for it at the other departments. This are taken into account when there is more time available then there was anticipated on. Furthermore the generated forecast may be updated manually by the employees. Since they have knowledge about the market circumstances. Demand may increase due to new customers for example.

2.5 Conclusion

The goal of this chapter was to understand the current situation. First, we looked at the demand patterns. We have seen that there is no trend present in the demand, except in 2020. This trend, however, does not extend in the following years. On average in the first period of the year there is less demand than in the rest of the year, about 85%. Also, in the third period there is a 107% demand compared to the average over the year 2018-2022. Therefore, we can say that there is a weak seasonality present, but this only yields for two periods. This is likely not beneficial to use in model development since it is a small effect.

The current way of forecasting is in the form of sales expectations called budgets. Developing a data driven forecast of the transport volumes can help Farm Trans in their decision making. Since budgets are vulnerable to increase in fuel prices, the current way of fleet capacity planning, using revenue generated per truck, results in a higher number of trucks required while transport volumes may not increase. Transport volumes together with truckloads and turnover rates therefore better reflect the capacity needed to fulfill demand.

To develop the forecasting model, data about the transportation volumes is required. This can be obtained through the order data which are stored in a database. Next to that external factors like budget, covid-19 and GDP are defined to be good explanatory variables for predicting transportation demand. In Section 4.3.1 more analysis is done on the effect of these variables on the demand.

Finally, the model requirements are identified. The goal is to generate forecasts on a weekly basis 12 weeks ahead and on a period (4 week) basis one year ahead. This rolling forecast is updated each week such that the most accurate forecasts are developed. A further distinction is made between the regions such that decision making can be aggregated to the regions where capacity is required. The model will be implemented via a prototype such that forecasts are accessible and are incorporated into the weekly reports.

3 Literature study

The goal of the literature study is to get familiar with theories about forecasting and gather the required knowledge to develop a forecasting model for Farm Trans and how the hierarchy should be handled. First some general knowledge about forecasting and the activities surrounding it are identified, including fleet capacity planning. Next we discuss what forecasting methods are applicable in this research, for example methods that have earlier performed well in predicting demand, which in case of Farm Trans are the transportation volumes. We do not go into much detail about the mathematics behind the forecasting models since the focus is on choosing well performing models for Farm Trans. Also most models nowadays are available as "black box". Furthermore the forecasts need to be evaluated in order to determine the accuracy such that we can choose the most accurate model. Some models require hyper-parameters as input. Tuning these is essential in order to obtain a solid forecasting model an optimize performance. Section 3.5 goes into detail about that. In the final section the hierarchical aspect of forecasting is taken into account.

3.1 General

Reduced to its basic essence, the goal of supply chain management is to try to match supply and demand. This is a difficult task due to uncertainty on both sides. Traditionally, this has been achieved through forecasting ahead of demand and creating inventory against that forecast. Demand management is the term that has come to be used to describe the various tools and procedures that enable a more effective balancing of supply and demand to be achieved through a deeper understanding of the causes of demand volatility. Activities in this process are aimed at reducing variability and increasing flexibility. Demand planning is the translation of our understanding of what the real requirement of the market is into a fulfilment program, i.e. making sure that products can be made available at the right times and place (Christopher, 2011). Forecasting is a part of this process and the goal is to predict the future as accurately as possible to decrease variability (Croxton et al., 2002). Accuracy can be described as the difference between the model outcome and the reality. Based on the forecasts, decisions are made to form the capabilities of the supply chain such that it best matches the customer requirements. Therefore the more accurate the forecast, the more accurate decisions can be made. But it has to be said that accuracy, and forecasting, is not an objective, it is a mean to achieve other objectives such as service levels (Zotteri and Kalchschmidt, 2007). For Farm Trans, this can for example be the determination of the in-house capacity needed.

Zotteri and Kalchschmidt (2007) also show also that forecasting has an impact on company performance though the impact depends on what the forecast is used for. Companies should therefore carefully consider how to use their forecasts according to their competitive priorities. Moreover they show that improving forecasts plays a significant role in improving company performance.

3.1.1 Fleet planning process

A decision support system for the fleet planning process is conceptualised in Figure 8 (Couillard, 1993). This approach could be used by Farm Trans for the decisions regarding capacity. The goal is to generate, evaluate and select a plan for the fleet size and composition by using information like the forecasts of demand. The forecast is thus an essential piece of information which is needed for the process, indicating why this research is essen-

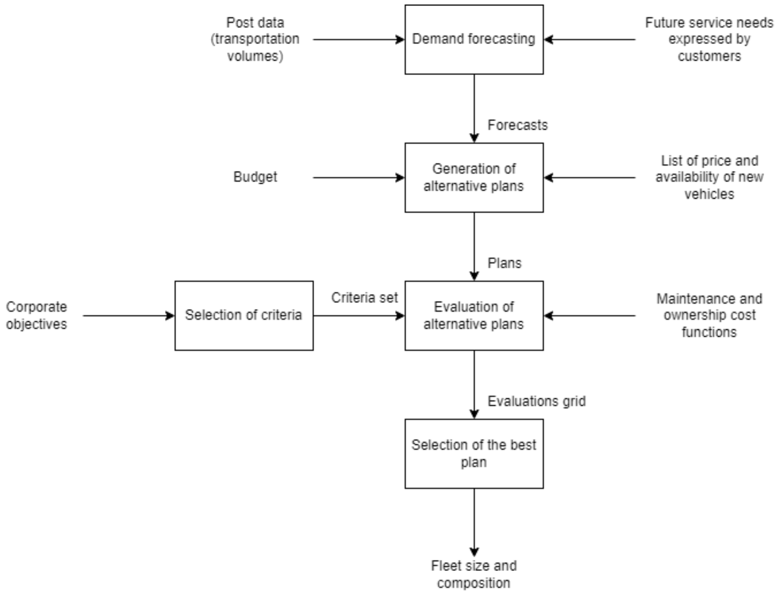

Post data (transportation volumes)
Demand forecasting
Future service needs expressed by customers
Forecasts
Budget
Generation of alternative plans
List of price and availability of new vehicles
Plans
Corporate objectives
Selection of criteria
Criteria set
Evaluation of alternative plans
Maintenance and ownership cost functions
Evaluations grid
Selection of the best plan
Fleet size and composition

Figure 8: Conceptual model of vehicle fleet planning (Couillard, 1993)

tial to conduct. Together with the budget and vehicle information alternative plans are generated. This is the main use of the forecasts that are generated in this research. Farm Trans can make use of the forecast by combining the information together with vehicle load and turnover rates to create an estimate of the capacity needed. These can be seen as the generation of alternative plans. This information differs between the different regions which is why that distinction is made in the first place. The support system itself is not be further applied in the research. Baykasoğlu et al. (2019) give a review of fleet planning problems in transportation systems.

3.2 Hierarchical forecasting

Our time series is disaggregated by geographic location (cross sectional aggregation) and time (temporal aggregation), this is also referred to as a grouped time series. This brings certain challenges and opportunities with it. In Figure 9 the hierarchy is shown. We can work with this aggregation in multiple ways which can improve forecasting performance. The forecasts add up the in the same way as the data, for example the demand at the geographic locations add up to the demand of the whole department. The traditional way to cope with this are the bottom up and top down approaches. We can forecast the bottom level series and sum them up to gain a forecast per week or period, the bottom up approach. This also works the other way around, top down. Forecasting the week or period series and then dividing it over the regions according to the historical proportions (Petropoulos, 2022). The same applies for time aggregation.

Athanasopoulos et al. (2017) show that forecasting with temporal hierarchies increases accuracy over conventional forecasting, particularly under increased modelling uncertainty. The most ideal solution is to combine the accurate aspects of the forecasts from each level. In this way we make use of medium and long term dynamics in the data, by forecasting the week and period level. It is shown in literature that it is a general phenomenon that forecast combinations outperform the individual forecasts (Andrawis et al., 2011). This also yields for cross sectional hierarchies. Combinations can be made based on a multiple techniques, Andrawis et al. (2011) mention 15 methods in total. Using a simple average

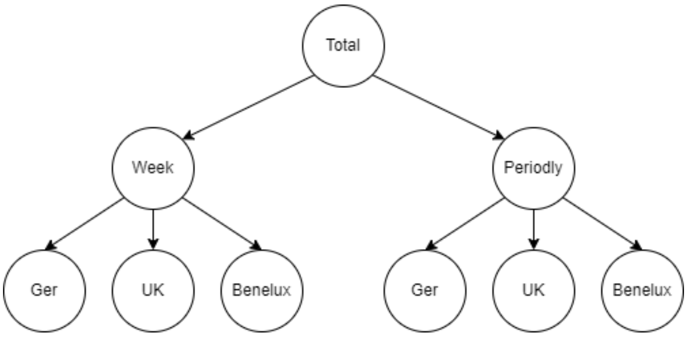


Figure 9: Forecasting hierarchy

is shown to be a robust and simple method, having comparable performance to the other methods who are more comprehensive. In our case we can combine different models or aggregation levels. When combining aggregation levels we also need to make use of a bottom up or top down approach to get the same aggregation. Since the focus is on the hierarchical component, we focus on combining the aggregation levels.

Another way we can use aggregation is through the use of a hybrid model, as also stated in Section 3.4. By first developing a global forecast of the demand per period and using this as explanatory variable in the other aggregated levels. This can be useful since generally a forecast at the highest aggregation is the most accurate. Here we can also use different methods, for example Zhang (2003) used an ARIMA and fed the outcome into an ANN. Daum et al. (2021) uses a macro-micro approach where a macro level is forecast by an ARIMA and fed into a ML algorithm that forecasts at micro level. The approach we use is related to the hierarchy, forecasting the highest aggregation level and feeding it to the lower levels. In Chapter 4 the forecasting approach is given where this is further ellaborated.

3.3 Forecasting methods

A forecasting model has the objective to predict the future. To predict an outcome, or response, Y with predictors X, we can assume that there is some kind of relationship. Which is written in the general form:

$$Y = f(X) + e \tag{1}$$

$f(X)$ is a function and e is the error term. When creating a model, we try to estimate $f(X)$ as accurately as possible to make a prediction about Y. Statistical learning refers to a set of approaches for estimating $f(X)$. Most statistical learning problems fall in two categories: supervised and unsupervised learning. Training forecasting methods falls in the supervised learning category: "we wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations" (James et al., 2021).

There are multiple categories in which forecasting models can be divided. A main distinction can be made between subjective and objective forecast methods. Objective methods make forecasts based on data and are therefore also referred to as quantitative methods. A further distinction can be made between the use of one or multiple explanatory variables (univariate and multivariate). Subjective methods are more qualitative based, they rely on judgemental approaches and educated guesses of professionals (Archer, 1980). These methods are appropriate where data is insufficient or inadequate. Examples are surveys, consensus of experts (Delphi technique) or scenario analysis, a combination may also be used. Initially quantitative based forecasts are reviewed and may be adjusted by the company's demand planners. Taking into account exceptional circumstances expected over the planning horizon or correcting perceived inadequacies in the system forecast (Fildes et al., 2009).

3.3.1 Time series models

A time series is a set of observations, each being recorded at a specific time (Brockwell and Davis, 2016). Time series models make use of historical data patterns and analyse the variable to be forecast. Based on these patterns a prediction of the future demand can be made. This is like extrapolation of data and it therefore assumes that the historical data is representative for future demand.

One of the simplest methods is a moving average. Where the last n observations are summed up and divided by n. When a new observation becomes available, the oldest in the sequence is dropped and replaced by the new one. This approach is fairly basic since no weights, trends or cyclical information is taken into account (Archer, 1980). A more meaningful approach is exponential smoothing, these methods are well known for forecasting time series. The methods originated from the 1950s and 1960s with the work of Brown (1959), Holt (1957) and Winters (1960). These methods have evolved over time including more types of trend and seasonality. The taxonomy of Hyndman et al. (2002) provides a help full categorization for describing the methods. There are five types of trend (none, additive, multiplicative, damped additive and damped multiplicative) and three types of seasonality (none, additive and multiplicative). Damped indicates that a trend or seasonality factor in the forecast is reduced when the length of a forecast horizon increases. These combinations result in 15 different models of which the most well known are for example simple exponential smoothing (SES) with no trend and no seasonality or Holt-Winters' additive method with additive trend and seasonality. A level demand is calculated and used in each model, this can be seen as the steady demand. Then, depending on the method used, a trend and seasonality is included. A trend indicates a steady growth in demand throughout the years. Seasonality is included if more or less demand arises in a certain period, for example every January. Which model to use depends on what demand pattern is present, i.e. trend and seasonality.

3.3.1.1 ARIMA ARIMA stands for Auto Regressive Integrated Moving Average and is a univariate method. It is a class of models that explain a time series based on its own past values (auto regression) and forecast errors (moving average). An ARIMA model is characterized by three terms: (p, d, q). The order of the auto regression and moving average are represented by p and q respectively, it represents the number of lagged values used. A pure auto regression model is like a multiple regression model but with lagged values of the outcome as predictors. d is the number of times differencing is required to make the time series stationary, i.e., the time series properties do not depend on the time at which they are observed. Since ARIMA models use linear regression, which work best when the predictors are not correlated, we require stationary time series. A stepwise approach can be used to search multiple combinations of the input parameters (p, d, q) and chooses the model that performs the best based on what accuracy measure is selected. The model parameters are estimated using maximum likelihood estimation (Hyndman and Athanasopoulos, 2018). ARIMA models can also incorporate seasonality. This is done by including additional seasonal terms $P, D, Q)$ in the model. They are similar to the standard terms but involve a backshift of the seasonal period.

3.3.2 Causal models

Where time series focus on analysis of the variable to be forecast, causal models are based on analysis of the relationship between explanatory variables and the one of interest, also called response variable (Archer, 1980). These explanatory variables are used to predict the response variable. When as explanatory variables the historic demand is used we get close to the time series, or univariate, models. Linear regression is a straightforward approach on predicting a response based on one or multiple predictor variables. It assumes a linear relationship between the two. Using the least squares approach, the best linear relationship is found which can then be used to predict the outcome. Many statistical learning approaches can be seen as generalizations or extensions of linear regression (James et al., 2021).

3.3.2.1 Neural networks Computer science has led the way with methods such as neural networks and other types of machine learning, which are getting a great deal of attention from forecasters and decision makers (Petropoulos, 2022). A neural network takes an input vector of p variables $X = (X_1, X_2, ..., X_p)$ and builds a nonlinear function $f(X)$ to predict the response Y. This idea is the same as described earlier, only the structure of the model is different.

Data which is sequential in nature, like our time series, can benefit from the use of recurrent neural networks (RNN). The input object X is in this case a sequence and the RNN is designed to accommodate and take advantage of the sequential nature of such input objects. A special case is the long-short term memory model (LSTM). It has the capability of remembering the values from earlier stages and to use that in the future. Siami-Namini et al. (2018) show that LSTM are superior to traditional ARIMA methods in terms of accuracy. Since a LSTM is a relatively new method it is interesting to assess performance of this method for predicting the transportation volumes.

3.3.2.2 tree-based methods The basis version of a tree-based method is called a decision tree, and in our case a regression tree since we try to predict a numerical value (transportation volumes). A regression tree consists of a series of splitting rules. These are based on the predictor variables and result in regions. The mean value of the response variables in a region represents the prediction when an observation lies in that region. The goal is to define the regions such that the RSS (residual sum of squares) is minimized (James et al., 2021).

A single tree generally does not have the same level of accuracy as some of the other regression models. However by generating many decision trees using bagging, random forest or boosting, the accuracy can be substantially improved. Random forests provide an improvement over bagging by decorrelating the trees. At each split in the tree, a random forest is not allowed to consider all available predictors, where bagging considers all available predictors. In this way when there is one strong predictor, the bagged trees use this predictor at the top split. Resulting in many similar trees which are correlated. Random forests overcome this problem by forcing each split to consider a subset of predictors of size m, an input for the random forest. Random forests are therefore the preferred method to use for forecasting (James et al., 2021).

3.4 Model and variable selection

According to Engle and Brown (1986), there are two basic types of model selection procedures. One is to perform a sequence of diagnostic tests between competing models to determine the smallest acceptable model. The other is to fit several models and select the model with the best accuracy according to some statistical evaluation measure. The evaluation measures are described in Section 3.6. The latter is most used in theory.

Despite the excitement surrounding the newer methods, older methods such as ARIMA and exponential smoothing are still valuable. The simple approaches are quite robust and not as prone to overfitting as the more complex, newer, methods (Petropoulos, 2022). ARIMA models are traditionally used very often, for example by Andreoni and Postorino (2006) to forecast air transport demand and by Rudakov et al. (2017) for forecasting volumes of railway goods transportation. It is also possible to use a combination of models. Zhang (2003) applied an ARIMA model and then used its forecast in an ANN (artificial neural network). Comparing the standalone ARIMA and ANN models with the hybrid ARIMA-ANN approach showed that short and long term predictions achieved higher accuracy with the hybrid approach. A hybrid approach forecasts the same variable first by a classical method and then feed its prediction to an ML method for forecasting the same variable. Building on this rationale, we can use a classical method like ARIMA to predict one variable and use that prediction as explanatory variable in another model. We can use this principle in the hierarchy by predicting the high aggregation and feeding it to another model predicting the lower level.

When there are a lot of variables which can be incorporated into the model, it may be the case that not all of them contribute to an accurate model. The most extensive approach for selecting which variable to use is best subset selection. A step wise approach to selecting the predictor variables is forward or backward selection. These methods are computationally efficient since much less models are created. Also shrinkage methods can be used for variable selection (James et al., 2021).

3.5 Model tuning

Hyper-parameters are the intrinsic parameters of a (machine learning) model which are set by the developer. Hyper-parameter tuning is therefore important in choosing a set of values that yields the most accurate model. It also determines the extent to which the model under- or overfits the data (James et al., 2021). Unfortunately the relationship between the performance of the machine learning algorithms and hyper-parameters is unclear. Therefore in practise we need to test different combinations of values (Wu et al., 2019). In selecting the hyper-parameters we can use the same approaches as defined in section 3.4. We however need to determine the ranges in which the parameters may vary, since otherwise the search space is infinite. So we must determine a set of values for the hyper parameters that we want to test. Grid search exhaustively trains a model for all combinations of hyper-parameter values. We can then choose the model with the best performance. Other popular methods are random search or bayesian optimization. In random search we define a statistical distribution for each hyper-parameter from which values are sampled. We can limit the number of combinations which can decrease the time needed for tuning. Bayesian optimization is a sequential model based optimization algorithm, it uses results from previous iterations in deciding the next hyper-parameter values. These methods are often more efficient because promising results are prioritized (Hutter et al., 2018).

Automatic forecasting algorithms can determine an appropriate model, estimate the parameters and compute the forecasts. So instead of choosing the hperparameters our self, we can use a pre programmed method to do this for us. Popular automatic forecasting algorithms are exponential smoothing or ARIMA models (Hyndman and Khandakar, 2008). In the standard ARIMA model we need to specify what values of p, q, d to use. But, with the use of auto ARIMA, the optimal values are determined according to a provided information criteria such as AIC or BIC. The parameters p and q can be iteratively searched by the auto ARIMA. The differencing term, d, requires a special set of tests of stationarity to estimate. Auto ARIMA uses KPSS tests for stationarity and Canova-Hansen test for seasonality.

3.6 Model evaluation

When developing a new forecasting model, it is common to evaluate the performance according to some measure of forecast accuracy and compare it with other forecasting methods (Petropoulos, 2022). To achieve this the data is separated into two parts, training and test data. The training data is used to develop and train the model such that it best fits the training data according to a measure. This model is then used to create forecasts for the test data and here its accuracy is evaluated. Because the test data is not used in determining the forecasts, it provides an indication of the performance on new data. The test set should be at least as large as the maximum forecast horizon required (Hyndman and Athanasopoulos, 2018). It can be the case that a model can accurately describe the training data, but perform poor on the test data. This is referred to as overfitting the data and it can happen because the statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance. There are multiple methods described in literature to measure this accuracy (Hyndman and Koehler, 2006) (James et al., 2021). The most known are mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). These are calculated according to Equations 2, 3, 4 and 5 respectively. y_i is a single observation, $\overline{y}$ represents the sample mean and $\hat{y_i}$ is the estimation. When we assess the performance of the training set we should refer to these measures as training measure, for example training MSE.

According to Hyndman and Koehler (2006), the RMSE is preferred to the MSE as it is on the same scale as the data but it is sensitive for outliers. Therefore this measure is used for comparing different models on the same data set. When we want to compare between different data sets we can best use percentage errors, like MAPE, since these are scale independent. MAPE is of the most importance for this research because of the better managerial insight. When we for example forecast 13 periods ahead on the tuning or test set, we can calculate the MAPE over this period. Looking at Equation 5, n is set to 13 and we calculate the absolute percentage error for each of the 13 periods and then take the average of these percentage errors to obtain the MAPE.

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y_i})^2 \tag{2}$$

$$RMSE = \sqrt{MSE} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i}^{n}(|y_i - \hat{y_i}|) \tag{4}$$

$$MAPE = \frac{1}{n}\sum_{i}^{n}\frac{|y_i - \hat{y_i}|}{y_i} \tag{5}$$

These measures only focus on the error of the forecast compared to reality. To account for overfitting the data we can add a penalty for the number of predictors used. Therefore this is often used as measure when selecting variables and tuning hyper-parameters. Akaike's Information Criteria (AIC) does this by estimating the test MSE, as can be seen in equation 6 where $RSS = \sum_i^n (y_i - \hat{y_i})^2$, d is the amount of variables in the model and σ^2 is an estimate of the variance of the error.

$$AIC = \frac{1}{n}(2d\sigma^2 + RSS) \tag{6}$$

Another measure is the Bayesian Information Criteria (BIC) and can be calculated according to equation 7. BIC replaces $2d\sigma^2$ by $log(n)d\sigma^2$. For $n > 7$, $log(n) > 2$ this means that BIC places a heavier penalty on a large number of variables then the AIC. So when using BIC we generally select models with less variables (James et al., 2021).

$$BIC = \frac{1}{n}(log(n)d\sigma^2 + RSS) \tag{7}$$

We however opt for using MAPE and RMSE since this gives better interpretable information, making the research more comprehensible. Also we are not dealing with choosing among a lot of predictor variable, like in classification models often is the case.

3.6.1 Validation

Next to separating the data into a training and test test, called the validation set approach as stated in Section 3.6, there are other methods which can be used for model validation. This is referred to as resampling methods, they repeatedly draw samples from a training set and refit the model to obtain additional information (James et al., 2021). The two most used methods are cross validation and the bootstrap. It can be computationally expensive since multiple models are fit but since more tests are done we get a better view on the models accuracy. Cross validation can be used in multiple forms. The general concept divides the data into k parts, also called folds, and assigns one fold as test set and the rest as training set. In this fashion k models are fit, each time using different data as training and test set. We then calculate an accuracy measure and take the average over all the models. Because we are dealing with time series we cannot randomly generate k folds, this would break the time series. For evaluation of a time series we can make use of a rolling origin evaluation. Forecasts for a fixed horizon are performed by sequentially moving values from the test set to the training set, thereby each time increasing the training set. When we keep the training set the same length it is referred to as rolling window evaluation.

Therefore we make use of time series cross validation, which looks a lot like leave one out cross validation. We take a series of a certain length and the last point in the series is left out as test observation. We fit the model to the training data and use it to make a prediction for the last observation. Then we can move ahead one period in time and repeat the process. How many times we can do this depends on the number of observations available and the number of lagged values that are used for training the model.

3.7 Impact of covid-19 on demand

As described earlier, covid-19 has had an impact on the transportation volumes of Farm Trans. Such a sudden change over time is also called a structural break. It denotes the moment in a time series when trends and patterns among variables change. A structural break can be one of the major reasons for poor forecasting performance (Bauwens et al., 2015). One way to cope with such event is by including extra variables. By including a binary variable that indicates whether covid-19 has had an impact on the demand in a period or not, this event can be explained. The proposed methods can then include this variable when training the model and thereby include the fact that covid-19 was present or not. Another option is to consider the period before or after the structural break.

3.8 Conclusion

In this chapter we investigated what forecasting methods are available to use for forecasting the transportation volumes at Farm Trans, answering the research questions stated for this chapter. We started with general literature about the bigger picture, supply chain management, in which forecasting plays a major role. There are a lot of possibilities in what kind of models and methods can be used. Ranging from simple exponential smoothing methods to more complex machine learning approaches using hierarchy. Because of the hierarchy in the time series that we want to forecast, we can make use of hierarchical forecasting techniques. This can be of great importance since it is proven by (Athanasopoulos et al., 2017) that it can increase accuracy over individual models. The approach of all models and methods stays the same: trying to predict an outcome with some predictors by making use of a mathematical function. This function is estimated as accurately as possible by making use of training data and is evaluated by making use of test data. Evaluation is done based on MAPE and RMSE. The MAPE is suitable to compare between different data sets, and the RMSE gives an indication of the standard deviation and is on the same level as the data.

Automatic forecasting algorithms are an attractive choice to incorporate into the research since these can determine an appropriate model and estimate the parameters themselves. Therefore, the auto ARIMA model is tested. Furthermore with the rise of machine learning methods and the good results, a random forest and LSTM are also tested. However the more traditional ARIMA and exponential smoothing methods provide often good results which is why those are also tested. The models require tuning of the hyper parameters. Here, also multiple approaches are suitable, but we opt for using grid search. Because we do not have a lot of observations per forecasting level, we expect the models to run fairly quickly and thus grid search is a viable option.

These methods can furthermore be combined with approaches such as hierarchical forecasting. A top down and bottom up approach are used to experiment with the accuracy and to see what the benefit is, compared to the individual models. Also, the forecast is fed into the models of the lower aggregation levels as explanatory variable to assess performance, this is referred to as the hybrid approach. Finally also combinations of individual forecast can made to generate forecasts. In this fashion we can determine the impact of the hierarchy on creating the forecasts since this plays a big role in the research.

4 Formulating Solutions

4.1 Introduction

In this chapter we formulate the possible solutions to solve the problem. First, the forecasting design is explained, giving more insight into the steps that are to be taken to generate the forecasts. It is explained how the available data is handled and prepared in order to train, tune and test the models. From Section 4.3 to Section 4.6 the forecasting models for all the aggregation levels are tuned and the results interpret. Starting with modelling individual models and ending with hierarchical forecasting approaches. At the end of each section, the results are concluded and the best approach according to the tuning approach is presented. In Chapter 5, these models are tested on the test set to assess final performance and decide which model to use for each aggregation level.

4.2 Forecasting design

Following the research model as described in Figure 4, see Chapter 2, the transportation volumes are forecasted using a data driven approach. A process model for data mining projects is CRISP-DM. It consists of six iterative phases from business understanding to deployment, see Figure 10 (Schröer et al., 2021). Business and data understanding is discussed in Chapter 2. The next step is to collect and pre-process the data such that it can be used to create the forecasting models. These can then be evaluated by their performance. The accuracy is the main performance measure because the more accurate the forecast the better decisions can be made. Accuracy is measured by mainly using the MAPE since this gives a good managerial point of view to the forecast performance and it can be used to compare different data sets, like between the different aggregation levels. Also RMSE is used since it is on the same scale as the data and gives a higher weight to high errors (Hyndman and Koehler, 2006). This is beneficial because a large error leads to a big difference in fleet capacity which we want to avoid. As discusses in Chapter 2, an improvement of 1.5% of MAPE results in requiring 1 truck and 2.3 trailers less. So to compare the performance of different models on the same aggregation level RMSE is used, and to compare between different data sets MAPE is used. Also, since MAPE gives an easier interpretation we will mostly talk about that.

Next to accuracy, the models can be compared on the basis of speed and comprehensibility (Singh et al., 2016). The running time of the model is not of the most importance since the model outcome is not used for a quick decision on daily basis. The model is run once per week to generate a new forecast, it is therefore no problem if it takes a couple of minutes before giving an outcome. However, this is not expected to be an issue since the data sets do not contain a large amount of values, shortening the required time to train a model. Comprehensibility is of more importance since Farm Trans and CAPE Groep are interested in gaining knowledge of the models and methods used, such that the models can be used after the research is finished.

Multiple models are tested in order to gain insight into what performs the best. The models hyper parameters are tuned such that its performance on the tuning set is maximized. For choosing which models to use we consider the criteria mentioned earlier. Besides those some methods like random forest and neural networks tend to require more data to work properly than methods like ARIMA and exponential smoothing due to the nature of the models. We test ARIMA and exponential smoothing since these are traditional and tend to work well in most cases. Furthermore, a random forest is fitted to test whether a more sophisticated machine learning approach can be beneficial for the accuracy. A neural

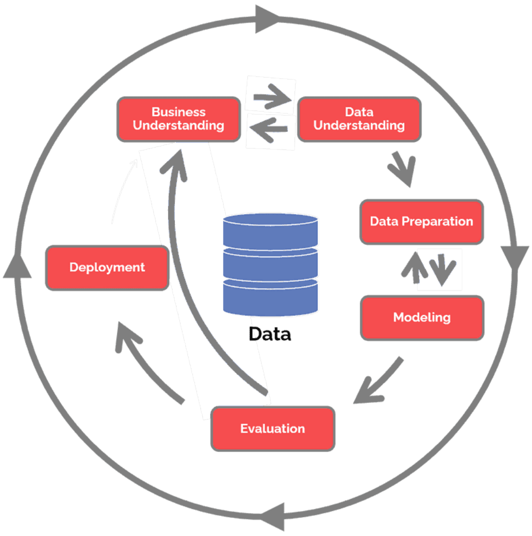


Figure 10: CRISP-DM (Schröer et al., 2021)

network approach, specifically LSTM, is also reviewed.

4.2.1 Data preparation

Data from the beginning of 2018 until May 2022 is available for usage. It is important to mention that periods (or weeks) at the beginning or end of the year may contain more or less than 28 days (the periods are defined such that each period contains 4 weeks). When this is the case, the total truckload is adjusted such that there is no difference in the data between the different years. Therefore, when generating the forecasts we need to re-normalize to account for the right number of days in the period. This is done by extra- or interpolating the values. Before we can apply forecasting methods to the data, we need to prepare the data such that it is in the right form. From the data warehouse of Farm Trans we can obtain the data. From this we applied some filters in order to obtain the correct information. The data is filtered by "IDVestiging", which must represents Fresh and Frozen. We also only consider information from 2018 onwards. Outliers are removed by only considering a truckload of less then 14 meters per order line. The data is split into a training, tuning and test set. Data from the year 2018 and 2019 is used as training set. Then data from the beginning of 2020 until period 6 of 2021 is used as tuning set. This is chosen such that 13 periods (one year) remains for the final test set. This is needed because the forecasting horizon is 13 periods. We can represent the data in the following form, where Y is the response and X the explanatory variable.

$$Y = \begin{bmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_T \end{bmatrix} X = \begin{bmatrix} y_{t-1} & y_{t-2} & \cdots & y_0 \\ y_t & y_{t-1} & \cdots & y_0 \\ \vdots & \vdots & \ddots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_0 \end{bmatrix} \tag{8}$$

This represents the univariate approach, assuming the past demand can explain demand in the future. When we want to include more variables, explanatory variables, we can add them to X so that they are included in the analysis. As described earlier the effect of covid-19, GDP and the budget is assessed. Explanation about them can be found back in 2.1.2.

4.2.2 Multiple-period ahead forecasting

When training for example an auto-ARIMA model the hyper parameters are chosen such that the training error (residual), in this case AIC, is minimized. Forecast errors are different from residuals in two ways. First, residuals are calculated on the training set while errors are calculated on the test set. Second, residuals are based on one-step forecasts while errors can involve multi-step forecasts (Hyndman and Athanasopoulos, 2018). Therefore, when we want to forecast multiple periods ahead we also need to train the model for this purpose. This is not possible for all models since not all models allow for separate Y and X definitions. There are multiple ways to cope with this. We can fit a new model for each time period that we want to forecast ahead, when forecasting 13 periods ahead, 13 models are generated. This is called a direct multi step forecast strategy and this is preferable since then the models are trained for the right forecasting procedure. Another option is to use a recursive multi step forecast. We then create a forecast for the prior time step which is then used as an input for making a prediction on the following time step.

4.2.3 Stepwise approach

We start at the highest aggregation level, forecasting global period demand. To get a first glimpse of performance we use the auto-ARIMA model to automatically select the model parameters. There is a possibility of overfitting since the parameters are selected based on the training set, not necessarily resulting in a good performance on the tuning set. Therefore, we manually perform hyper parameter tuning to get models which perform well on data they have not yet seen before. This is first done by a single training and testing set and later with cross validation. Cross validation is the standard approach used for tuning but the first time a single set is used to show the importance of cross validation. How the cross validation is performed is explained in Section 4.3.1, where we first apply it. After tuning the models the best individual models are selected. For this best model the explanatory variables, as stated in Section 2.1.2, are added to the models to assess the performance. Performance may increase because these variables have a impact on the demand and may therefore also have predictive power. When we know the best ARIMA model, the random forest and exponential smoothing models are assessed. Here we only perform cross validation for tuning the models and then also add the explanatory variables to the best individual models. When we know the best overall forecasting model we move on to the next aggregation level and perform the same steps, using cross validation to tune the models. Next to that also hierachical forecasting methods become available for use. This is explained in Section 4.2.4. In Figure 11 the overall approach is shown.

4.2.4 Hierarchy approach

As described earlier in Section 1.4.3, and Table 2, the focus lies on generating a forecast at four levels. We start at the highest aggregation level, predicting the demand for fresh and frozen (global) per period, and work our way down towards forecasting on regional level per week. As discovered during the literature study, hierarchical forecasting can be used in many ways. Our approach is as follows: First individual models are created, thereby tuning the models and exploring the use of explanatory variables. Then, if applicable, a bottom up or top down approach is used to forecast. For each aggregation level this has a different meaning. For example we can forecast global weekly by using a top down approach with the forecast of global period. When applying such a method we only move one aggregation level. So for example from global period to global weekly and region period, but not to region weekly. We do this because that level is not directly related and in switching levels generally accuracy is lost.

Second, the forecasts are combined using a simple average. Andrawis et al. (2011) showed this generally provides more accurate forecasts than individual models. For example, the individual forecast of global period and global weekly are combined to generate a forecast for both. The weeks are summed, like in the bottom up approach, and then the average is taken to gain the forecast for the global period. In the same way the forecast for global week is gained by using the top down approach on the global period to get forecasts per week. To get the forecasts the best individual models are used according to the cross validation approach in the tuning set. We refer to this as the combination approach. Lastly, the forecast of other levels may be incorporated as explanatory variables in the models as used by Zhang (2003) and Daum et al. (2021). We essentially search for a relation between the earlier generated forecast of a higher level and the demand of a lower aggregation. This is referred to as the hybrid approach. This is a bit different from the approach that Zhang (2003) uses. Since we are including other aggregation levels, and in Zhang (2003) they are forecasting for the same aggregation level. This also indicates the novelty on forecasting methods included in this research. On the right side of Figure 11 the way aggregation can be used top down is shown. On the left the general forecasting approach is shown.

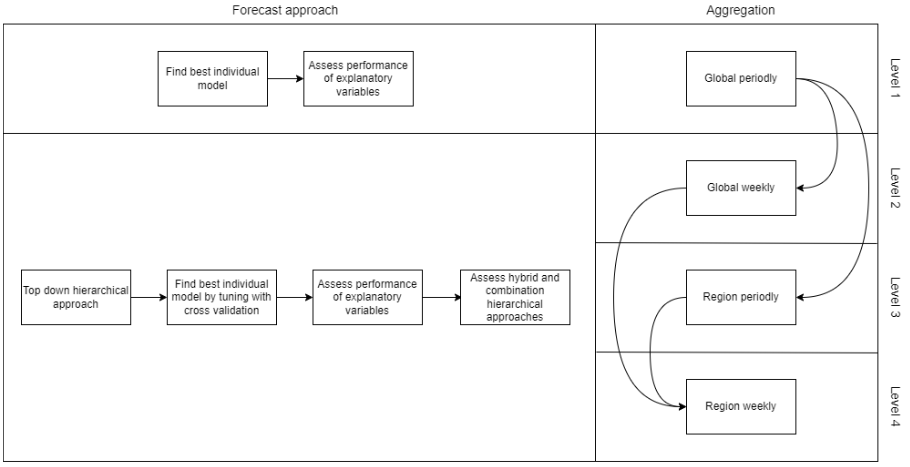

Forecast approach
Aggregation
Find best individual model
Assess performance of explanatory variables
Top down hierarchical approach
Find best individual model by tuning with cross validation
Assess performance of explanatory variables
Assess hybrid and combination hierarchical approaches
Global periodly
Global weekly
Region periodly
Region weekly
Level 1
Level 2
Level 3
Level 4

Figure 11: General forecasting approach

4.3 Global periodly forecasting

First, we create a model for the global demand per period. The three different models are tested and we start with using auto-ARIMA to get a first glimpse of forecasting performance. The working is discussed in Section 3.3.1. Then we manually tune the models to optimize the forecasting performance.

4.3.1 ARIMA

To get a first indication, we make a forecast for the year 2021 by using all previous information as training data. The best set of parameters is chosen by auto-ARIMA which results in the lowest AIC regarding the training data, thereby assessing performance of the model when forecasting 1 period ahead. This is done by making use of the pmdarima package in python (Smith et al., 2017). The result is an ARIMA(0,1,0) model and in Figure 12b the forecast is shown when predicting 1 ahead. The y-axis label is removed due to confidentiality. The historical demand and the training prediction, the blue and yellow line respectively, are shown. This data used to train the model and choose the parameters. Then from period 39, the start of 2021, the forecast together with the real demand is shown, the orange and grey line. On these last two the accuracy measures are calculated.

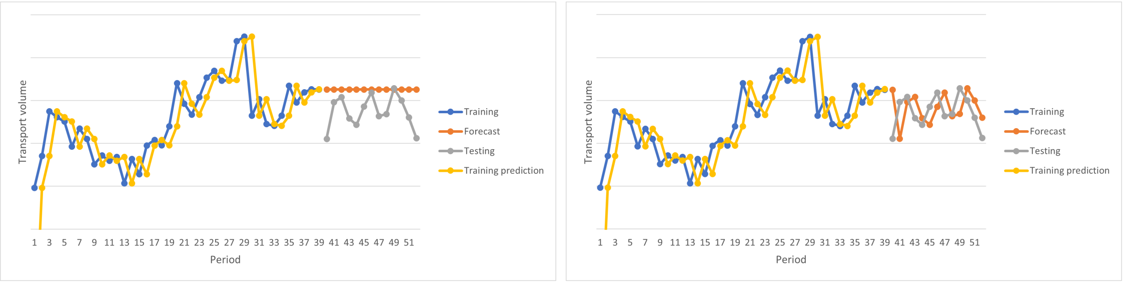
Since the parameters p and q, which stand for the order of auto regression and moving average, are both 0, the model can be seen as a random walk (Hyndman and Athanasopoulos, 2018). A random walk is defined in Equation 9. The forecasted value y_t depends on an intercept B_o and a fraction B_1 of the demand of the previous time step X_{t-1}. This means that according to the best ARIMA model, the time series is quite unpredictable.

$$y_t = B_0 + B_1 X_{t-1} + e_t \tag{9}$$

We can use the same model for forecasting multiple periods ahead. Since auto-ARIMA does not allow for an option to use the multi step forecast strategy, we opt for using recursive multi step forecasting. ARIMA(0,1,0) comes close to the naïve method, the forecast is updated by the same value each time step. The result is shown in Figure 12a and provides a MAPE of 9.6%, calculated according to the equations explained in 3.6. From this we can conclude that no relationship is found between the lagged values, errors and the demand in the next period by using the auto-ARIMA method.

This is however the model that performs best on the training data. It may be the case that a different configuration of the hyper parameters will lead to better results on the tuning data set. Therefore multiple configurations of the ARIMA model are trained and evaluated. In the Appendix the results are shown. Based on the tuning set and looking at the RMSE and MAPE, the best model is an ARIMA(0,0,0) model for forecasting 13 periods ahead with a MAPE of 5.6%. This is a good result in terms of accuracy but the model does not indicate the use of auto regression and moving averages. This is not desirable when looking at support for implementation at the organisation. Also since we test the model only once, the result is not credible. It may very well be possible that if we move on one period we get a different result. This will be assessed when performing cross validation in the next section. Therefore cross validation is a better approach on determining performance of the models.

Cross validation To get a better understanding of the model performance, we need to get closer to the real situation. In practise, each time we obtain new data, a new forecast is



(a) 13 periods ahead

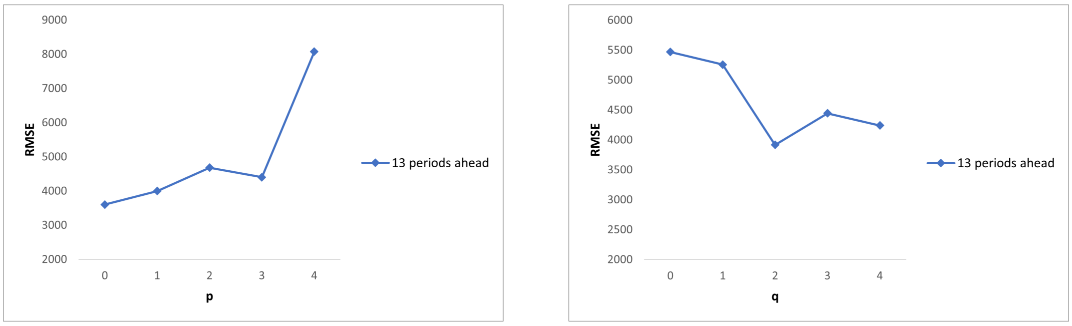
(b) 1 period ahead

Figure 12: Historic demand together with ARIMA(0,1,0) forecast of global truckload per period

generated. This can be replicated by using cross validation. The advantage is that we can train and fit the model multiple times to get a better understanding of the performance. This is important because it may be the case that a model performs well on one part of the data but not on another part, this is likely the case for the ARIMA(0,0,0) model. We make use of the training and tuning data sets. The goal is to forecast one year (13 periods) ahead, so this is the forecasting horizon. We start in 2020 and forecast one year ahead and calculate the performance measures. Then we add the new period, fit a new model and generate the forecast one year ahead. We continue in this fashion until no more data is available in the tuning set, which results in 6 iterations. Each iteration provides the MAPE and RMSE as accuracy measure, calculated according to the formulas explained in Section 3.6. We take the average of the performance measures over the iterations to determine the overall performance of the model. This means that some periods are incorporated more often than others. Meaning that if a period has a large one off in real demand this can impact the accuracy of the model. These kind of problems are left out of scope for this research.

We test different ARIMA configurations and compare the performance on the tuning set. The best performing model is selected. In Chapter 5 the model is tested on the test data set, using information which the model has not yet seen before to assess performance. Tuning ARIMA means altering the order and seasonal order. We tested with the order (p, d, q) ranging from $(0, 0, 0)$ to $(1, 1, 2)$ and seasonal order (P, D, Q) ranging from $(0, 0, 0)$ to $(1, 1, 1)$ using a grid search approach. From it is concluded that d=0 always outperform other values for d, the same yields for the seasonal order $(0, 0, 0)$. Further altering the seasonal order is not possible due to the available data and we showed that we only need to difference once to make the data stationary in Section 2.1.1. Therefore, we further tested the p and q values. An overview of the results is summarised in Figure 13. Here we see that the best average performance is achieved with a value of 0 for p and a value of q of 2 or 3. The best performing individual model however is an ARIMA(0,0,4) model with a MAPE of 8.2%. Therefore the tuning process only gives an indication of the range of values which fit the model well. From there we can search in the neighborhood of the good performing parameters to find a, local, best model. In case of the ARIMA(0,0,4) model the neighborhood consists of (1,0,4),(0,0,3) and (0,0,5). These models did not result in a better performance.

The cross validation approach gives a different result than before. Where we first concluded an ARIMA(0,0,0) model now an ARIMA(0,0,4) model is the best. Increasing performance from 10.0% to 8.2% based on the cross validation results. The best ARIMA(0,0,4) model indicates that 4 moving average terms are used. When predicting 13 periods ahead this results the forecast moving towards the average (Hyndman and Athanasopoulos, 2018). In fact, the first 5 periods a straight line is predicted. Indicating that further into the future, predictions become less accurate such that a more robust model performs better. Demand fluctuates a lot between the periods and therefore an average gives a good performance.



(a) Performance of p

(b) Performance of q

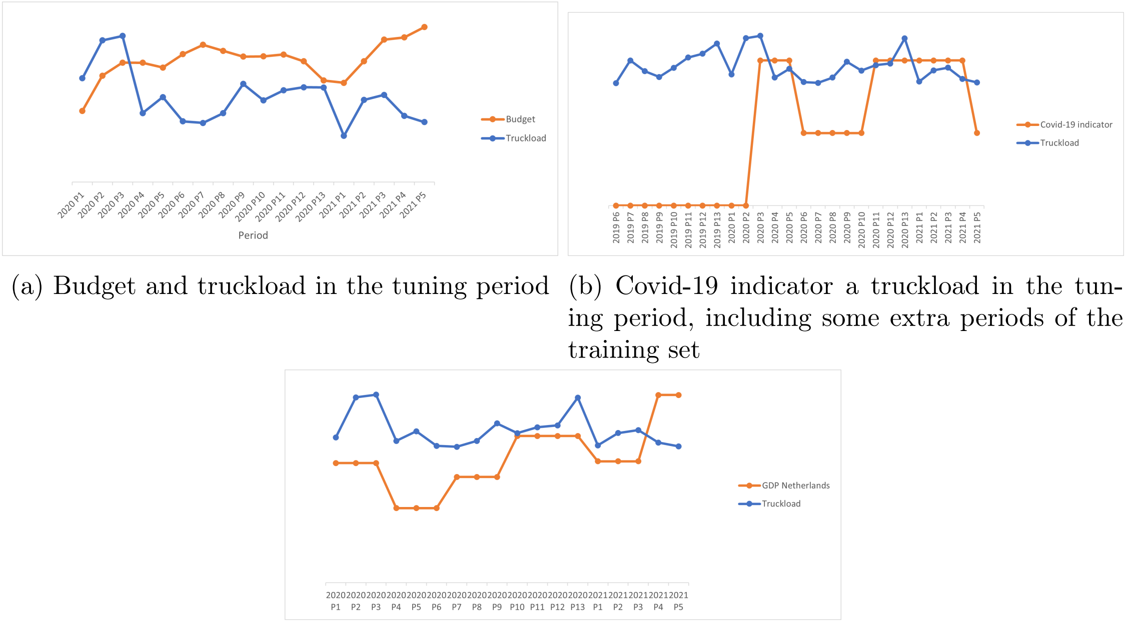
Figure 13: Average performance of p and q in ARIMA

Explanatory variables To increase the accuracy of the forecasting model the defined explanatory variables may provide extra explanatory power. These explanatory variables are added to X as defined in Section 4.2.1. We perform cross validation and use the ARIMA(0,0,4) model since it performed the best without any explanatory variables. When predicting 13 periods ahead we see a small improvement in terms of accuracy. The lowest RMSE is achieved when using the budget as explanatory variable resulting in a MAPE of 7.9%. This is interesting because when comparing the budget directly to the revenue in the same cross validation period, the MAPE is 13.6%. When comparing the budget to the truckload there are however some indications that when truckload increases, budget increases. But this is not trivial for the whole period and is shown in Figure 14a.

Including the covid-19 indicator as explanatory variable did not lead to an increase in forecasting accuracy. When we plot the covid-19 indicator with demand over the tuning data we see that it is not evident that when there is a lockdown effective, the demand is always lower than when there is no lockdown. This is due to the fact that before 2020 covid-19 was not present and there was an upward trend. This means that the model is not able to learn that a lower demand is due to covid-19, since lower demand is also associated with no lockdown. This can be seen in Figure 14b where the covid-19 indicator is plotted against the truckload over the tuning data and 8 period before to show why the model is not able to learn from it. Therefore we assess different training and tuning period is Section 5.5. Lastly the GDP of the Netherlands is plotted against the truckload in Figure 14c. Here we can see in some cases that when GDP decreases or increases the truckload does the same. Although for example in the last period the GDP increases but the truckload decreases so it is also true the other way around.

4.3.2 Exponential smoothing methods

Another method that is tested are exponential smoothing methods. As discussed in Section 3.3.1, these models are pure time series based. Since ARIMA showed simple methods may



(a) Budget and truckload in the tuning period

(b) Covid-19 indicator a truckload in the tuning period, including some extra periods of the training set

(c) GDP of the Netherlands and truckload in the tuning period

Figure 14: Explanatory variables plotted against the truckload in the tuning period

produce good results, exponential smoothing methods may produce accurate results. Nine different combinations of models are tested including no, additive or multiplicative trend and seasonality. The model parameters are estimated by maximizing the log-likelihood, using the statsmodels package in python (Seabold and Perktold, 2010). The results are also computed using cross validation and are shown in the appendix.

The best performing model, using no trend and no seasonality results in a MAPE of 14.7%, which is worse then ARIMA. A high smoothing parameter of 0.8 is used for the last cross-validation iteration. This smoothing parameter may change each time new data becomes available and the model is trained again. This smoothing parameter is calculated such that it minimizes the training error, therefore we may optimize and change it each time the model is trained. A high smoothing level indicates a reactive model, something we also saw as result of the auto-ARIMA model (which looked like the naïve method). Each time new data becomes available the level is updated with 0.8 times the new demand which then becomes the new forecast. Also since the best model involves no trend and no seasonality we can conclude that a more simple method yields the best result. Overall the exponential smoothing methods do not result in higher accuracy of the forecast.

4.3.3 Random Forest

It is proven that a single decision tree results in less accurate results than combining multiple trees (James et al., 2021). A random forest combines multiple trees, resulting in a more accurate model. Another advantage of a random forest is that we can specify the response and independent variables separately, where for ARIMA this is not the case. Therefore we can easier create models to predict multiple periods ahead by altering the independent variables. In this case we create a model for each period that we want to predict ahead, therefore using the direct multi-step approach. This suggests that forecasting with random forest has a high probability to be more accurate than ARIMA. Model estimation is done through the use of the scikit machine learning package in Python (Pedregosa et al., 2011).

We make use of the cross validation approach. Since we need to specify our predictor variables ourselves we choose to use 13 lagged values of the response variable as independent variables. This represents one year of data, such that seasonality can be taken into account. Using more lagged variables is not beneficial due to the amount of data which is available. Also from the ARIMA model it was concluded that using more lagged values was not beneficial in terms of accuracy. A downside of using random forest is that the first 13 observations cannot be used, there are no 13 lagged values for these observations. As hyper-parameters the standard settings are used at first which are: (mtry=all, maxdepth=none, samplessplit=2, samplesleaf=1, ntrees=100).

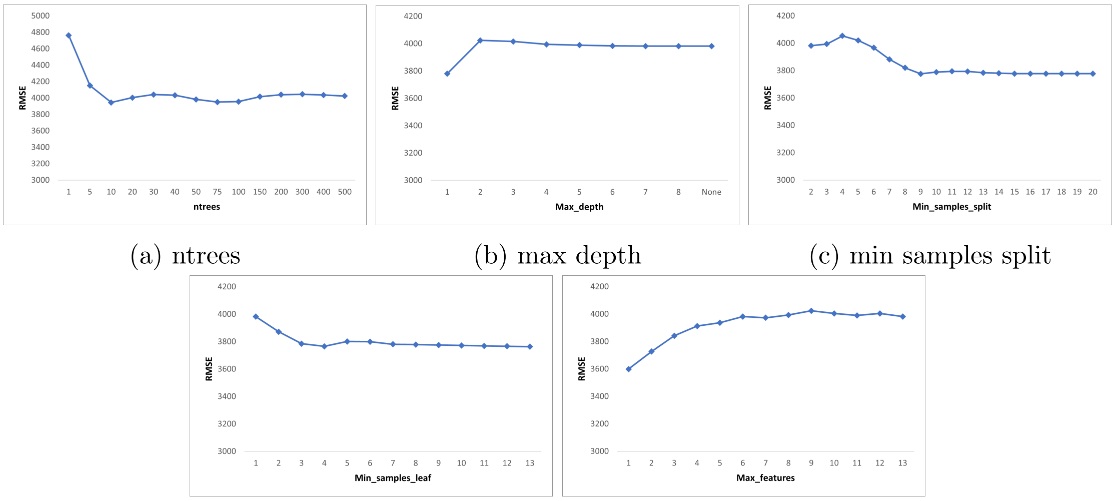
When we take a look at a separate iteration of cross validation approach, the training data and prediction of that training data do not seem to differ much. This indicates a good model. However when we look at the tune data there is a bigger difference between the real and predicted value. This can indicate that the model is overfitting, meaning that the training data is modeled too well such that it negatively impacts the performance on the tune data. To solve this problem hyper parameter tuning is performed.

Tuning the random forest Tuning means setting the hyper parameters such that the performance on the tuning data set is improved. Tuning is done for predicting 13 periods ahead. We can change the following settings in our random forest: mtry, max depth, samples split, samples leaf and ntrees.

The maximum number of features (mtry) determines the number of prediction variables considered at each split. Thereby forcing each individual tree to be different which improves the overall accuracy of the random forest. Max depth represents the depth of each tree. The deeper the tree, the more splits are made and therefore more information is captured in the tree. The minimum sample split determines the minimum number of samples that is required at a split at an internal node. When we increase this parameter, each tree in the forest becomes more constrained as it has to consider more samples at each node thus less splits can be made. The minimum sample per leaf is similar, but for the final nodes (leafs). Finally, the number of trees determines how many decision trees are generated. The more trees are generated, the longer the time required to fit the random forest. Usually the higher the number of trees, the better the data can be learned. Therefore there is a trade off between performance, time and overfitting (Probst et al., 2019). To determine the "sweatspot" of the hyper-parameters we conducted multiple tests, each time altering one of the hyper-parameters in a certain range while keeping the others at their default values also using the cross validation approach. The result is summarised in Figure 15.

When we look at the number of trees we see that the RMSE decreases as the number of trees increases. This makes sense since when we only construct one tree, we model a single decision tree and therefore do not make use of the properties of a random forest. When using more than 20 trees we do not see an improvement in accuracy, while we increase the running time. The optimal value of number of trees is therefore be around 20. We say around 20 because the other hyper-parameters may also have an influence on the optimal number of trees to use. The lower the number of features considered, the better the performance. This indicates that a more simple model works better. The maximum depth of 1 gives the lowest RMSE also indicating a non complex tree. This may indicate that there is one feature which has a high importance, this is investigated later on. The minimum samples required at a split flattens out after using 10 samples and at the leaf around 7 samples.

Searching for combinations of the above mentioned hyper parameters we find a best



(a) ntrees (b) max depth (c) min samples split

(d) min samples leaf (e) max features

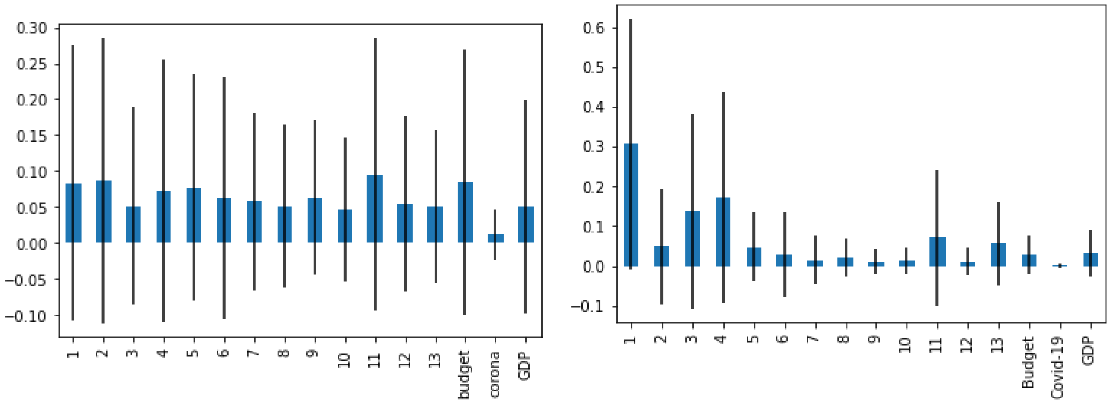
Figure 15: Tuning random forest hyper parameters

model with (mtry=1, maxdepth=None, samplessplit = 2, samplesleaf=1, ntrees=50) with a MAPE of 9.6%. Searching in the neighborhood of this does however not lead to any improvement. A mtry of one is rather curious, since at each split a random feature is chosen to split upon. When repeated a lot of times this results in forecasting close to averages. This is something we also saw at the ARIMA model, which tended towards averages when forecasting a long time period ahead. This indicates that more robust models tend to work well in this environment. When mtry is 1, each feature has almost the same feature importance. This is combined with no max depth and a no restriction on the sample split and sample leaf parameter resulting in large trees with low difference between the decisions in the final node due to a mtry of 1.

Explanatory variables By adding explanatory variables the accuracy of the model does not change that much. The highest accuracy is achieved by using the GDP resulting in a MAPE of 9.6% meaning no improvement is found here. The same yields for including the other variables. Since multiple features are given as input, not all features may be of the same importance. Some may have a great influence on the outcome while others do not. We can assess feature importance by computing the mean and standard deviation of accumulation of the impurity decrease within each tree. Feature importance is shown in Figure 16, where in (a) mtry is set to 1. The blue bars are the feature importance of the forest, along with their inter trees variability represented by the error bars. When mtry is set to 1 the feature importance is almost the same as we would expect. Therefore, we also included the feature importance when all variables are assessed at each split. This indicates that the first lagged variables is of much higher importance than the others. But, using this does not result in more accurate forecast which is due to overfitting as discussed earlier.

We see a high variability between the different trees, represented by the black bars in the graph, which means that we cannot say with certainty that a feature has a high importance on predicting the transportation volumes. This means that the predictive performance of the features is not very high, something we also noticed when testing the ARIMA models. Also we see that the covid-19 feature does not have any importance at all. The budget has the highest importance of the explanatory variables. This was already discovered by the fact that the model performance increased when using budget data.

Figure 16: Feature importance



(a) mtry = 1

(b) mtry = auto

4.3.4 LSTM

Deep learning methods are capable of identifying structure and pattern of data such as non-linearity and complexity in time series forecasting. In particular, LSTM has been used in time-series prediction (Siami-Namini et al., 2018). LSTM has been tested with its default configuration from the keras package in python (Chollet et al., 2015). This however yielded the same result as we have seen earlier: copying a large amount of the demand of the last period and therefore coming close to a naïve method. Due to time reasons it is therefore decided not to further investigate the LSTM model.

4.3.5 Conclusion forecast global period

We have tuned ARIMA, exponential smoothing and random forest models to forecast the transportation volumes of Farm Trans by evaluating them with cross validation on the tuning set. In Table 5 the best models of each method are summarised. We also included the result of a naive forecasting method, which performed worse than all models. This shows why it is important to train these individual models. Performance is, at best, increased by 7% which represents 4.6 trucks less need to be acquired on the long term.

Overall we have seen that when predicting further in the future a more robust model achieves a higher accuracy. In this case a robust model represents a more simple model where the forecast moves towards the average historic demand. The ARIMA(0,0,4) model with budget achieves the highest performance when predicting 13 periods ahead. The order of (0,0,4) verifies that forecasting further into the future is more difficult and therefore an average forecast is more accurate. Because d in the order (p, d, q) is set to 0, the long term forecast will go to the mean of the data. Also since the demand of Farm Trans can fluctuate a lot between the periods, and no clear relationship is found in these fluctuations, forecasting an average gives the best performance.

As said the demand fluctuated in the past. Therefore, the models have a hard time in finding patterns and relationships between the past and future. A large upward trend was present in 2019 and in 2020 there was a big decrease due to the closing of catering industry. This could explain why no relationship are found. Since march 2020 the demand has not come back to the level of before covid-19. However including the covid-19 indicator as explanatory variable did not lead to an increase in forecasting accuracy. This can be seen in Figure 14b. In Section 5.5 it is investigated if using data from that point in time has a positive effect on accuracy. Including the budget did have a positive effect on the ARIMA model since accuracy improved by 0.3%. We have seen that the budget sometimes follows the same patterns as the truckload but this is not always the case, for example due to price fluctuations. Therefore accuracy increase is not significant.

Table 5: Results of the best models

Model	RMSE 13	MAPE 13	RMSE 1	MAPE 1
Naive method	5,142	14.9	4,168	10.0
ARIMA(0,0,4)	2,946	8.2	2,809	7.9
ARIMA(0,0,4) variables (b)*	2,904	7.9	2,983	9.0
Exponential(N,N)	4,732	14.7	2,711	7.4
RF	3,598	9.6	3,184	8.3
RF variables (g)*	3,592	9.6	3,296	8.7

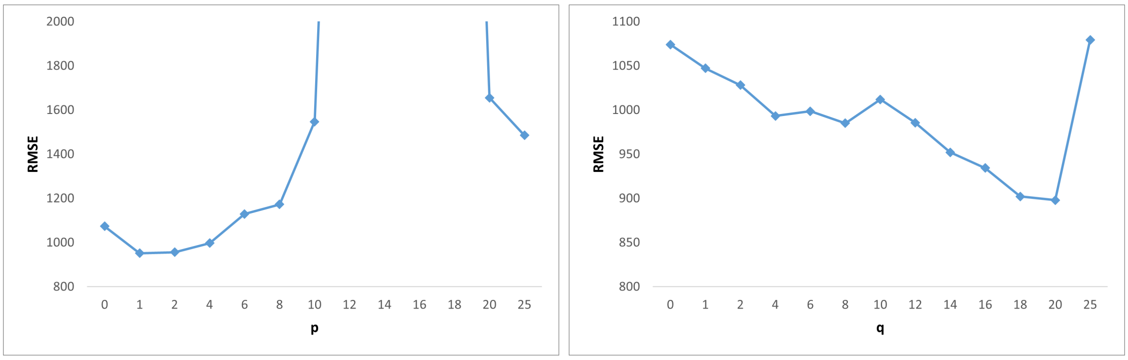
* c= corona data, g = GDP, b = budget

4.4 Global weekly forecasting

As stated in Figure 11, the next aggregation level is the global demand per week. The focus is to forecast 12 weeks, which is equal to 3 periods, ahead. However, we also forecast 53 weeks ahead such that we can compare the performance with the models of global period which forecast 1 year ahead. Also, at this aggregation level we start making use of the hierarchical properties of the data. We start with a top down approach of the forecast, dividing the forecast per period over the weeks and assess the performance. For this we make use of an average percentage of demand that has occurred in a week in a period in the first 2 years, the training data. We use the best model found for forecasting 13 periods ahead, which is the ARIMA(0,0,4) model. Again the cross validation approach is used to evaluate accuracy. The result is a MAPE of 10.3%. This score is close to the performance of the global period forecasts which indicates a good result for such an easy method.

4.4.1 Training and tuning

At forecasting the global period demand, we have seen that tuning the models result in better accuracy. Because of that the individual models of this aggregation are also trained and tuned. Results of the previous tuning cannot be used since the time frame is different, which may indicate different demand patterns. Since we now predict the weeks, we have more input variables and therefore the range of values to test is bigger. The tuning is therefore done by altering the parameters one at the time. The result is summarised in figure 17. A low value of p provides a good result, as does a value of q of around 20. The best individual model is of order (0,0,19) resulting in a MAPE of 8.4%. An improvement of 1.9% compared to the top down approach. This model can be said to be comparable to the best ARIMA(0,0,4) model used to forecast global period since 4 periods is close to 19 weeks.



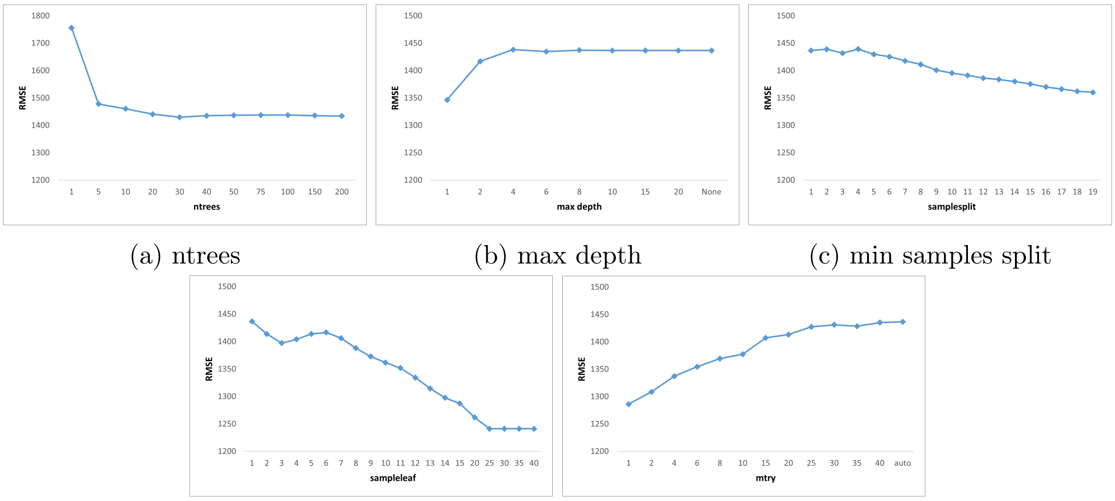
(a) Tuning p

(b) tuning q

Figure 17: ARIMA p and q tuning with standard model order (0,1,1) and seasonal (0,0,0) on weekly global demand

Regarding exponential smoothing, all different models are tested. The best exponential smoothing model is again a model with no trend and no seasonality, just like in the forecasting model for predicting the demand per period, resulting in a MAPE of 13.4%. The smoothing parameter is 0.6 and therefore a bit lower than in the global period model, indicating the model is less reactive and more information about the past is used.

The results of tuning the random forest is summarised in Figure 18. When we look at the feature importance for predicting one week ahead we see that the first lagged variables has a very high importance. This means that the demand of last period has the most impact on the next week, indicating a reactive model. This is something that we see returning also in the ARIMA models. The other lagged variables do not have a high importance, therefore also indicating seasonality is not present.



(a) ntrees

(b) max depth

(c) min samples split

(d) min samples leaf

(e) max features

Figure 18: Tuning random forest hyper parameters weekly forecast

The results are the same as when tuning the random forest on the period data. An optimal number of trees is around 20 trees, the lower the max depth the better, a high number of samples is required at the split and leaf and a low value of max features benefits performance. The best individual random forest has (mtry=1, maxdepth=2, samplessplit=6, samplesleaf=3, ntrees=100) resulting in a MAPE of 12.6%. This model was found by performing a local search on the best model found until no better model was found in the neighborhood. An mtry of one indicates that one feature is considered at each split at random, therefore the decision tree cannot give priority to a feature that has high predictive power. This results in a prediction going towards averaging. However the random forest is a more sophisticated method, the performance does not increase.

4.4.2 Including explanatory variables

When including the explanatory variables into the models, no improvement is found for the ARIMA (0,0,19) model. The best performance is achieved by including the budget as explanatory variable. We would have expected that including budget would result in a better accuracy since this was the case for global period. Because budget information is available on period level, not on weekly level, every 4 weeks the budget is updated. Combined with the fact that the demand fluctuates week to week, it indicates that no clear relationship is found between both. Covid-19 and budget did not lead to an improvement so it is no surprise that the same yields for the lower aggregation level. As for the random forest, small improvement is gained when including the explanatory variables. The accuracy measures are shown in Section 4.4.5.

4.4.3 Hybrid model

Using the forecast from the higher aggregation levels as explanatory variables in the lower levels is referred to as a hybrid model approach. We make a forecast per period with the best model, which is the ARIMA(0,0,4) model. For each week the forecast for the corresponding period is added as explanatory variable and the model is fitted. This approach can be used by the ARIMA and the random forest models. Unfortunately, exponential smoothing methods are not able to incorporate this approach. Using the best individual models, ARIMA gives a MAPE of 9.0% where the random forest performs at an mean average error percentage of 12.6%. For both this does not mean a improvement. Although for the ARIMA model the coefficient of the model parameter is 0.32 with a low p-value. Meaning that there is a high chance that the variable has a impact on the output. The forecasts do change, but the overall accuracy is not improved. However the model with hierarchy is more reliable since the parameters have lower p-values. This result is not the same as we have seen in the literature, where Zhang (2003) showed an increase in accuracy when feeding output of an ARIMA model into a machine learning model. We applied this in combination with hierarchy instead of on the same aggregation level which shows no improvement in this case.

4.4.4 Forecasting 12 weeks ahead

The goal for the weekly level is to forecast 12 weeks ahead such that capacity can be altered on the short to medium term. Therefore the forecast is also evaluated for this time frame. Since the forecasting horizon is different, the models are tuned with their new forecasting horizon.

Tuning the ARIMA model results in optimal model of order (1,0,20). Which is close to the model for predicting 53 weeks ahead. For the random forest tuning leads to a different result compared to the tuning done for 53 weeks ahead. The best model is found with a configuration of (mtry=1, maxdepth=None, samplessplit=2, samplesleaf=1, ntrees=200). Again a mtry of 1 results in the best performance, indicating random variables are selected at each split. In combination with no depth this results in an average prediction. An example of one of the decision trees used can be found in the appendix.

4.4.5 Conclusion forecast global weekly

Overall we see the same result as for the global period forecast, the ARIMA model gives the best result. 19 lagged weeks are used compared to 4 periods to predict global period transport volumes which is comparable. After 19 weeks an average of the demand is taken as the forecast. This more robust method gives better performance since the demand fluctuates a lot between the weeks. It also shows that no clear relation ship can be found by the models since the best performing models tend towards forecasting averages. This can be explained due to the fact that covid-19 had a big impact on the demand, making it less predictable. Including the hybrid hierarchical forecast approach did not result in an increase in performance. This can be the case because a comparable ARIMA method is used for both the global period and weekly time series. No new information is provided to the model. Also it should be noted that the top down approach achieves a high performance for the simplicity it has. Tuning the individual forecasting models lowers MAPE by 1.9%, an improvement of 18%, which can be seen as the benefit of explicitly modelling forecasts for each aggregation level. Also, the approach is again much better than the naive method.

Table 6: Results of the best models for global week

Model	RMSE 53	MAPE 53	RMSE 12	MAPE 12
Naive method	1,996	24.9	1,910	21.9
top down global period	985	10.3	1,200	12.9
ARIMA(0,0,19)	800	8.4	942	10.7
ARIMA(0,0,19) with hierarchy	854	9.0	962	10.7
ARIMA(0,0,19) with budget	1,061	11.4	1,037	11.3
ARIMA(1,0,20)	919	9.9	925	10.1
Exponential(N,N)	1,134	13.4	1,040	12.1
RF	1,070	12.6	965	11.1
RF with budget, corona, GDP	1,068	12.4	954	10.9
RF with hierarchy	1,071	12.6	959	11.0
RF best 12 week	1,069	12.5	967	11.2

4.5 Region period forecasting

We continue with the same approach for generating the forecast per region. First, we use the earlier generated forecast, global period and divide it over the regions, top down, by looking at the average percentage of demand in each region in the training data set. A MAPE of 8.0%, 51.1% and 32.6% is achieved for the Benelux, Germany and UK respectively. This achieves a good result for the Benelux which comes closes in accuracy compared to global period forecast. This can be explained by the fact that the Benelux is responsible for 80% of the demand, therefore having demand patterns which are very much alike that from the global demand. Germany and the UK however do not follow these patterns which is why the accuracy is lower.

Next, the tuning approach is used to determine the best configuration for the individual models. The results are shown in the appendix since these are comparable to the tuning results per period when tuning for the global demand. The best ARIMA model is an ARIMA(0,0,4) for the Benelux. This is the same model as for the global period and is also explained by the fact that around 80% of the global demand originates from the Benelux. For Germany the best is ARIMA(1,0,1) and for UK ARIMA(3,0,2). In these models also the auto regressive term is included, indicating that demand from 1 and 3 lagged periods is used in generating the forecast. The accuracy for Germany and the UK is much worse than for the Benelux. The predictions become much less accurate when we aggregate the regions. This confirms the fact that the more we aggregate the demand, the less predictable it becomes. Making use of hierarchy could therefore be beneficial for these regions. But, since the demand patterns are different for the UK and Germany it could also be that the forecast does not provide much predictive power for these regions, this is investigated later on.

Regarding the tuning of the random forest we see the same patterns as when tuning the global period model, especially for the Benelux time series. This mean an optimal number of trees of around 30, a low maximum depth, high minimum number of samples per split and leaf and a low number of features to consider at each split. Roughly the same yields for Germany and the UK but the exact results can be found in the appendix. The best individual models are (mtry=1, maxdepth=None, samplessplit=2, samplesleaf=1, ntrees=30) for the Benelux, (mtry=auto, maxdepth=None, samplessplit=2, samplesleaf=6, ntrees=100) for Germany and (mtry=auto, maxdepth=None, samplessplit=8, samplesleaf=1, ntrees=50) for the UK. The Benelux model is the same as for the global period model only a smaller number of trees is used. We already reflected on the model parameters impact on the forecast in Section 4.3.3.

The exponential smoothing methods are simple and fast and therefore all combinations are run. The model with no trend and no seasonality performs best for the Benelux with a smoothing parameter of 0.5. This model is less reactive than the model for predicting the global demand, meaning historical information is more important. For Germany and the UK an additive and multiplicative seasonality provide better results. This indicates that there is more seasonality present than in the Benelux. Looking at the demand there are indeed larger fluctuations between the different periods then compared to the Benelux. When decomposing the time series of Germany there is more demand in the summer periods compared to the other time series. Furthermore, to the best ARIMA and random forest models the explanatory variables are added and the performance determined. This also includes the hybrid hierarchy approach.

4.5.1 Conclusion forecast region

The tuning results are summarised in Table 7. The best performing method is an ARIMA model where for the Benelux we use the individual model and for Germany and the UK we use explanatory variables. The best individual model configuration varies for each region which is why this is not indicated in the table. Taking a weighted average gives a representation of what model performs best overall. We choose weighted average since demand in the Benelux is much higher than in the other regions. Using a weighted average and looking at the MAPE, because we compare different data sets, we conclude that ARIMA with explanatory variables has the highest accuracy. In the implementation phase it is considered whether one generic model is used or that a different model for each time series is used. Interestingly, for the Benelux, including hierarchy as hybrid model at ARIMA does not give an increase in accuracy while it does when we include it in the random forest. Since the forecast from global period is also made by an ARIMA, the forecast is really comparable, especially for the Benelux. We are therefore providing an explanatory variable which is very much alike the forecast. Searching a linear relation with that proves to be ineffective. The random forest makes use of decision trees, looking at the feature importance the hierarchy variable has the second highest importance, after the first lagged value. Zhang (2003) also fed the outcome of an ARIMA model into a machine learning approach yielding better results, this is also what we see here. For Germany and the UK including hierarchy does not lead to an increase in accuracy which is due to the fact that the demand patterns differ a lot from the global forecast.

4.6 Region weekly forecasting

Finally, forecasting models are developed for the region per week. Regarding the hierarchy we now have 2 inputs, one from the region forecast per period and one from the global forecast per week. For the Benelux and UK the global forecast per week provides better results, for Germany the forecast per period and region. But the models are not better than the individual models for this aggregation level.

When tuning the ARIMA models, it is concluded that a lower p and q value is beneficial for all regions. The graphs can be found back in the appendix. The results do differ a bit per region, which results in a different best model for each region. An ARIMA(2,0,1), (2,0,0) and (10,0,2) are the best individual models according to the cross validation on the tuning set. These models differ from the earlier individual models, more auto regression terms are used.

Table 7: Tuning results of the best models for each region per period predicting 13 periods ahead

	Benelux		Germany		UK		Average
Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	MAPE
Naive method	3,365	13.5	785	39.3	1,610	29.8	17.8
Top down global period	2,220	8.0	938	51.1	1,895	32.6	15.0
ARIMA	1,737	6.6	642	33.3	1,401	24.8	11.3
ARIMA with variables	1,879	7.0	491	23.7	1,266	22.8	10.5
ARIMA with hierarchy	1,921	7.1	664	34.6	1,753	30.7	12.5
EXPON	3,031	12.4	785	43.8	1,243	23.0	16.5
RF	2,579	10.2	717	39.3	1,444	25.2	14.7
RF with variables	2,684	10.7	717	39.3	1,434	25.2	15.1
RF with hierarchy	2,383	9.5	694	36.2	1,494	25.8	13.9

When tuning the random forest method again the same patterns arise as before. The best individual models according to the tuning are (mtry=4, maxdepth=6, samplessplit=2, samplesleaf=1, ntrees=50) for the Benelux. This differs from the Benelux period model that more features are considered at each split which makes sense since a lot more features are available (53 weeks instead of 13 periods). For Germany the best configuarion is (mtry=auto, maxdepth=None, samplessplit=45, samplesleaf=1, ntrees=50). But the performance of the different configurations do not differ much of eachother. (mtry=auto, maxdepth=None, samplessplit=2, samplesleaf=24, ntrees=50) for the UK. The best exponential smoothing methods are with no trend and no seasonality for the Benelux and additive seasonality for Germany and the UK. Indicating that there is more seasonality present those countries.

For the best models the explanatory variables are added but the impact is negligible, like at the models for the region period forecast. Adding hierarchical forecast of the global period forecast improves accuracy.

4.6.1 Conclusion forecast weekly region

in Table 8 the tuning results are displayed using the cross validation method. Overal the best method is the ARIMA model using variables. However, it differs per region. Including the global period forecast with the hybrid model approach does not lead to an increase in performance.

4.7 Hierarchical forecasting

Now that we constructed all the individual models, the hierarchical forecasting approaches can be used. First the top down and bottom up approaches were used. The top down results are already incorporated into the conclusions of the previous sections. For all models the best individual ARIMA models are used for the computation and the error measures are based on forecasting one year ahead. We choose the ARIMA methods because these are proven to be effective for almost every aggregation level and we can then compare it to those results effectively. The bottom up results are summarised in Table 9. On the left side of the table we see the aggregation level of the forecasts which will be summed up to the aggregation level specified in each column. For example when summing the forecast of region week to forecast the Benelux per period we get a MAPE of 7.0%.

Table 8: Tuning results of the best models for each region per week predicting 53 weeks ahead

	Benelux		Germany		UK		Average
Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	MAPE
Naive method	1,469	24.9	238	50.0	458	36.7	28.6
top down region period	857	12.9	190	39.7	463	31.0	16.9
top down global weekly	781	11.5	271	60.4	481	32.6	18.5
ARIMA (best)	630	9.3	175	36.4	345	22.2	13.3
ARIMA with variables	672	10.1	137	25.4	326	22.6	13.0
ARIMA with hierarchy	631	9.4	165	32.0	412	27	13.6
EXPON	840	13.4	176	34.5	290	19.5	16.0
RF	724	11.3	184	38.6	346	22.3	15.1
RF with variables	723	11.3	184	38.6	346	22.3	15.1
RF with hierarchy	725	11.3	192	40	363	23	15.3

Table 9: Bottom up cross validation results

	Region period				
	Benelux	Germany	UK	Global week	Global Period
Region week	7.0	36.6	23.5	10.1	
Global week					7.2
Region period					9.2

Remarkable is the global week to global period forecast. The best individual model for global period achieved a MAPE of 7.9% where here we achieve a performance of 7.2%. This is unexpected because the accuracy of the global week model is 8.4%. However it can be argued that the weekly forecast captures a different time frame which is better predictable. In Chapter 5 this is further investigated by assessing the test set. When adding the forecasts of each week together we get better results which proves the effectiveness of hierarchical forecasting methods, even simple ones like this bottom up approach. Also the forecast of Benelux per period achieves a low error measure, but this is not better than the individual model. Also the other bottom up forecast do not achieve a better result than the individual models. It is assumed that because of the different demand patterns of the UK and Germany the hierarchical approaches do not work correctly for those regions. It could be the case that using different methods for dividing the forecast over the region does lead to better results but these are not investigated.

Furthermore we apply the combination approach as suggested by Andrawis et al. (2011). This method is introduced in Section 3.2 and 4.2.4. The results of the application to our forecasts can be found in Table 10 and 11. The simple average approach is used since this is a robust and simple way to use combinations of forecasts. When combining for example global period and week we first sum up the global week forecast to get two forecast for global period, then we take the average. It is also done the other way around, we get a top down forecast from global period for global week and then take the average to get the final forecast for global week.

Accuracy is improved only for the Benelux per period by a small margin, improving from a MAPE of 6.6% to 6.5%. Therefore we do not see the same results as Andrawis et al. (2011). Here also using a different combination method may produce better forecasts but

Table 10: Combining individual forecast models cross validation results period results

	Period			
Combination	Global	Benelux	Germany	UK
Global period+week	7.2			
Region period+week		6.5	34.5	23.3

Table 11: Combining individual forecast models cross validation results week results

	Week			
Combination	Global	Benelux	Germany	UK
Global period+week	9.8			
Region period+week		10.4	37.5	24.3

this is out of scope for this research.

4.8 Conclusion

In this chapter we discussed the tuning of the different forecasting models and decided which model and method gave the best accuracy over the tuning set using cross validation. This is summarised in Table 12. For every aggregation level, we have seen that all models perform better than a naive forecast. This is important to note since naive forecast are very easy to generate and since we achieve a much higher accuracy we can say that it is worth it to put time and effort in generating these forecasting models. For example for the global period the forecast is improved by 7% compared to the naive method. For the other aggregations it is only more. Overall the best method to use is the ARIMA model, where sometimes we include the explanatory variables. For almost all regions this provides the highest accuracy. We see that the best models tend to go towards average forecast predictions in the long term. This is also something that is explained by Hyndman and Athanasopoulos (2018). This creates robustness in the model which has a positive result since the demand can fluctuate a lot between the periods. If the fluctuations cannot be predicted, forecasting in between these fluctuations gives the highest accuracy. It may be ineffective since it is of importance to grasp these fluctuations for monitoring and controlling of the business activities. The models have shown that this is hard to do since there are no clear seasonal patterns and also the best random forest methods tend towards robuster forecasts. This can be the case since covid-19 has had a big impact on the demand and this period is present in the tuning set. However using the covid-19 indicator as explanatory variable did not always result in better performances than the robust models, it did in some cases in combination with the budget. This is also due to the training set chosen. In the current training period only a few observations include an impact of covid-19 on the demand. Therefore in Section 5.5 it is tested if whether using only data with covid-19 impact has a positive effect on the accuracy. We do not see an improvement inaccuracy when incorporating the GDP. This is different from what we have noticed in literature.

What we can conclude from the hierarchical approaches is that forecasting only the top level and dividing it along the way gives reasonable accurate results. Therefore this may be seen as a low effort manner to generate good forecasting results. It does have some problems when the forecast is divided over the regions since the demand patterns differ between them. In Chapter 5, we elaborate on which methods are the best for each aggregation level using the test data set. We also concluded that when using the bottom

Aggregation	Best tuning model/ method	MAPE tuning
Global period	ARIMA (0,0,4) with budget	7.9
Global period	Bottom up global week	7.2
Global week	ARIMA (0,0,19)	8.4
Benelux period	ARIMA (0,0,4)	6.6
Benelux period	Combination region period + week	6.5
Germany period	ARIMA (1,0,1) with budget and covid-19	23.7
UK period	ARIMA (3,0,2) with budget and covid-19	22.8
Benelux week	ARIMA (2,0,1)	9.3
Germany week	ARIMA (2,0,0) with budget and covid-19	25.4
UK week	Exponential (additive seasonality)	19.5

Table 12: Overview best models per aggregation according to cross validation on the tuning data set

up approach from global week to global period, better results where achieved than with the individual model. Therefore, this approach is also shown in Table 12.

5 Results

In this chapter we assess the accuracy of the models on the test data set. It is important to note is that the models that require tuning, like ARIMA and random forest, are not yet tested on data that the model has not yet seen before. Therefore we use an extra, third, test set to assess the performance of the best models that followed from the cross validation approach. Since the forecasting horizon is 13 periods, the test set contains one year of data ranging until the fifth period of 2022. The models are retrained with all data before the start of the test set to get the best estimation of the model parameters and the best configuration of hyper parameters resulting from the tuning are used. This is done because we want to include the most recent information of demand in the models since this is the most representative for the current situation. We test the best individual ARIMA, exponential smoothing and random forest models together with the inclusion of explanatory variables. Also the hierarchical forecasting methods are tested on the test set.

Ideally we want to use cross-validation here to get the best understanding of the performance, like in the tuning approach. But due to the limited data which is available, this approach is not applied on the test data set. We therefore apply a single train and test set approach here.

Performance of the budget versus the revenue in the test data period gives an MAPE of 4.6%. This is fairly accurate but, as said before, the budget is also used as guideline for the business. This means that when not on target to meet the budget certain actions are taken to increase sales. Also the budget is not directly related to the truckload since margins may differ throughout time due to for example fuel costs.

5.1 Global period test

Starting with the highest aggregation level, we found the best performing individual model for the tuning set was the ARIMA(0,0,4) model. The hierarchical bottom up method using global week was the method to achieve the lowest accuracy in the tuning set. All models are now also tested on the final test data. The result is shown in Table 13.
In general the performance improves. This can be explained by the fact that the structural break in the data due to covid-19 is further away and therefore has less influence on the demand. This information has therefore less impact on the forecast. Patterns are more stable in the recent years which increases model predictability, also due to the robustness. This also yields for the other forecasting aggregations. The random forest achieves the best accuracy in this test case and therefore the results differ from the cross validation approach where bottom up forecasting performed the best. It does however not vary much, only 0.2%. Since the bottom up model performed much better in the cross validation tuning approach, we still opt for that solution for predicting global period transport volumes. Also the ARIMA model achieves a good performance in the test set. When we include the budget in the ARIMA model we see a major increase in MAPE, which is bad. The budget in this period is a lot higher compared to the transportation volumes which causes a big difference in performance. This can be explained by the increase in fuel prices, as mentioned before. Therefore the budget will be higher but this does not say anything about the volumes that need to be transported. The budget is made by the employees and we can therefore also interpret the budget as a human influence on the forecast. Because the budget is used as explanatory variable these factors need to be taken into account when producing the forecast. Covid-19 could also have had an impact on the demand but this is not the reason for the budget to increase, rather for the volume to decrease. While the volumes fluctuate we see a steady increase in budget.

Table 13: Results of the best models for forecasting periodly, tested on 2022

Model	RMSE 13	MAPE 13
Bottom up global week	2,368	6.8
Bottom up region period	2,741	7.9
ARIMA(0,0,4)	2,365	6.7
ARIMA(0,0,4) with budget	8,967	28.2
Exponential(N,N)	2,588	7.3
RF	2,318	6.6
RF with GDP	2,390	6.6

In Figure 19 the performance of the ARIMA(0,0,4) model is shown on the test set. The forecast is made at the end of the fifth period in 2021 for 13 periods ahead. Because of using 4 moving average terms we get a straight line forecast after 4 periods of prediction, indicating a robust model gives a good accuracy and that it is difficult to predict more than 4 periods ahead. According to Farm Trans, this is not beneficial to gain support for the forecast within the organisation. An increase throughout the periods is more desirable because of the wish to let the business grow. But according to the research no upward trend is found in the last years. Furthermore, this forecast shows a "too simple" model which does not lead to increase trust in the model. However since it is shown that this model creates the most accurate forecast we still opt for this solution. In the implementation phase of this project the opportunity is given to Farm Trans to choose which ARIMA method is used to create the forecasts. Such that we are not bound to this one solution. This is more elaborated in Chapter 6.

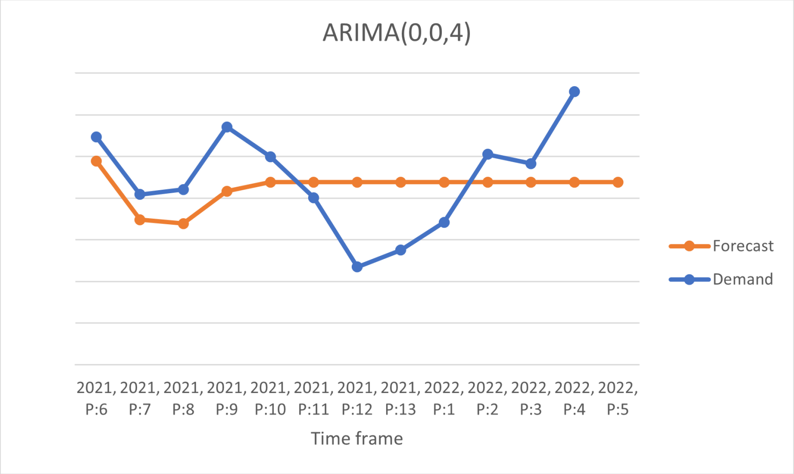


Figure 19: ARIMA(0,0,4) forecasting the test set

5.2 Global weekly test

The best performing model following from the cross validation is the individual ARIMA(0,0,19) model. We test this model together with the other models and methods with the best configurations following from tuning the parameters. The performance is shown in Table 14. We conclude that the ARIMA(0,0,19) model performs the best when predicting one year ahead and the random forest with budget, covid and GDP as features performs the best when predicting 12 weeks ahead. This is a different result compared to the tuning results where in both cases ARIMA performed better. We know from the previous results that the ARIMA tends to a robust, average, forecast when predicting long periods ahead. We can therefore conclude that the random forest is better in predicting shorter term ahead. On the long term the difference between the models is small. Furthermore using the increase in budget due to fuel prices results in a worse performance when including this variable, this was also the case in the forecast of global period. The hierarchical methods top down, bottom up and combination also do not result in a lower error measure. The combination approach comes the closest to the best performing model. This ARIMA model performs good in both cases, which is why ARIMA is favorable. The performance increase of the random forest for 12 weeks ahead is significant but a decrease of 1% of MAPE does not have that much of an impact on the decision making of Farm Trans. This is elaborated in Chapter 2 and 6.

Table 14: Results of the best models to forecast weekly, tested on 2022

Model	RMSE 53	MAPE 53	RMSE 12	MAPE 12
Top down global period	1059	10.7	787	8.4
Bottom up region weekly	811	8.5	828	9.5
Global period+week	652	7.7	606	6.8
ARIMA(0,0,19)	643	7.6	504	5.5
ARIMA(1,0,20)	757	8.3	589	7.7
ARIMA(0,0,19) with budget	2,112	26.8	1,301	16.8
ARIMA(0,0,19) with hierarchy	707	8.2	813	9.7
Exponential(N,N)	1,200	13.9	1,203	14.8
RF	671	7.7	454	4.6
RF with budget, corona, gdp	730	8.2	440	4.5
RF with hierarchy	683	7.9	462	4.7

5.3 Region period test

The next aggregation level is region per period. Again the overal performance improves compared to the cross-validation results. From the tuning set the best model was the ARIMA model where for Germany and the UK we used the budget and covid-19 indicator as explanatory variables. Also for the Benelux the combination approach where we combined the region period and week performed well. When retraining the models with all data and making the forecast for the test set the results differ a bit. The results are presented in Table 15, the configurations used are not shown because these differ per region but they are the best following the tuning approach. For the Benelux using a hybrid model approach where we include the forecast of global period from the ARIMA(0,0,4) model, as explanatory variable results in the highest accuracy. The combination approach also achieves a good accuracy but not as good as from the tuning set. Germany and the UK follow different demand patterns and since the global forecast can be compared to the demand of the Benelux it has a negative impact on the other regions. ARIMA with variables performs the best for Germany, which is the same as in the tuning approach. For the UK the hierarhical combination approach works the best. Here we combine the Region period and week, where we use bottom up from the week to get a period forecast, by taking the average. We also see that this approach on average gives the best result for this aggregation level. The second best model is the ARIMA model with hierarchy as explanatory variable.

Table 15: Results of the best models forecasting each region 13 periods ahead

	Benelux		Germany		UK		Average
Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	MAPE
Top down global period	1,968	7.4	972	51.6	1,941	34.1	14.7
Bottom up region week	1,277	5.6	789	23.0	771	12.5	8.0
Region period+week	1,245	5.2	718	20.6	725	12.0	7.4
ARIMA	1,354	5.7	819	23.9	877	14.3	8.4
ARIMA with variables	2078	8.3	449	13.5	834	14.9	10.0
ARIMA with hierarchy	1,162	4.6	745	21.5	956	16.4	7.6
EXPON	1200	5.0	641	17.6	772	13.3	8.1
RF	1908	8.4	1,072	32.9	692	13.1	13.4
RF with variables	1809	8.0	805	24.8	686	13.3	12.4
RF with hierarchy	1,985	8.8	1,063	32.8	804	15.5	11.7

5.4 Region weekly test

The final aggregation level is the region per week. From the tuning set the best models where ARIMA for the Benelux and Germany and exponential smoothing for the UK. After retraining the models and making a forecast one year ahead for the test set we get the results as presented in Table 16 and 17. The configurations used are not included here since these differ for each region. Again the best performing models are ARIMA for the Benelux and Germany, where for the latter we again include the covid-19 indicator. However for the UK the best model is the random forest including the forecast of global period as explanatory variable. The best model for the UK earlier was the exponential smoothing model which now performs as one of the worst. The difference can be found in the middle of the test set where demand is much lower than the prediction. Here the seasonal estimate is not accurate, indicating that not each year the seasonal effects are the same. This also was the period where there was again lockdown so this is also a reason why the forecast may be off. Therefore we do not recommend using exponential smoothing since it is less robust than for example ARIMA.

Table 16: Results of the best models forecasting each region 53 weeks ahead

	Benelux		Germany		UK		Average
Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	MAPE
Top down region period	778	11.2	1610	33.2	398	26.2	15.0
Top down global weekly	781	11.5	271	60.4	481	32.6	18.5
Region period+week	937	12.4	143	26.9	345	20.9	14.7
ARIMA	414	6.5	207	23.3	227	13.9	8.9
ARIMA with variables	611	9.9	156	17.6	252	18.8	12.2
ARIMA with hierarchy	417	6.7	229	26.6	256	16.4	9.7
EXPON	494	7.5	222	24.7	308	25.8	12.9
RF	583	9.9	322	39.5	178	12.4	15.1
RF with variables	521	8.8	188	21.5	179	12.6	12.8
RF with hierarchy	565	9.7	324	39.7	174	12.1	12.6

Table 17: Results of the best models forecasting each region 12 weeks ahead

	Benelux		Germany		UK		Average
Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	MAPE
Top down region period	825	12.5	163	33.6	284	22.0	15.5
Top down global weekly	1,559	8.3	274	61.0	374	28.1	15.5
Region period+week	1,566	21.7	147	17.7	247	14.3	20.3
ARIMA	410	5.8	208	24.9	243	14.2	8.5
ARIMA with variables	422	6.6	141	13.7	207	15.3	8.7
ARIMA with hierarchy	346	5.3	225	27.5	296	18.8	9.0
EXPON	417	6.1	135	15.5	209	15.6	9.2
RF	685	12.0	288	35.0	204	13.0	16.7
RF with variables	585	10.2	141	15.6	204	13.0	13.6
RF with hierarchy	616	10.9	287	35.0	209	13.2	13.2

5.5 Altering the training periods

Covid-19 had an impact on the demand of Farm Trans, from March 2020 the upwards trend disappeared and we can argue whether the data before 2020 can be used for modelling or not. Therefore, it is tested whether it is beneficial to use information from the point where covid-19 was present. Because there is less data available, we only make use of a training and test set. The test set contains information from 2021 period 6 until 2022 period 5, this has not changed such that we can compare performance. Therefore, the data from 2020 period 4 until 2021 period 6 are used as training data. Regarding the random forest, the models cannot be trained in the same way as before since we need at least 26 observations to train all the 13 models (one for each period to forecast). Therefore the recursive multi step approach is used, training one model and create a forecast one step ahead. This forecast is then used as input for the following time step in the series. The random forest is then retrained with the new time series, and again a one step ahead forecast is made. This continues in this way until we have reached the goal of predicting one year ahead.

Since there is less training data available, the models can be estimated less accurately which may have a negative impact on the performance. Especially for the random forest there is less information available which also changes the impact of the configuration of the hyper parameters. For example a high minimum number of samples spit is not feasible since not enough observations are available. But to compare the performance we still use the same configurations. The results are summarised in Table 18.

Table 18: Results of best models, using data from the point where covid-19 impacted the demand

	Global		Benelux		Germany		UK	
Model	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
ARIMA (period)	2,114	5.8	1,681	7.5	1,094	34.1	799	14.9
ARIMA (weekly)	797	8.6	480	8.0	286	34.8	188	14.5
RF (period)	2,630	7.4	1,158	4.8	859	25.8	616	11.3
RF (weekly)	706	7.7	399	6.0	222	25.3	215	13.9

The results vary across the different forecasting levels. The global forecast per period gives an MAPE of 5.8% which is better than the 6.7% (found back in Table 13) achieved by the model using all available information. However for the different regions it does not result in better performance. Since this is a one time forecast it is also a less accurate result than for example the result of the training set. Because we want the final advice to be comprehensive we make some trade-offs. It is not desirable to have a different approach and method for each aggregation level since this is not worth the time it take to implement. It is therefore recommended to use the original approach.

5.6 Conclusion

In Table 19 an overview is shown of all the aggregation levels with their best models according to the test set when predicting one year ahead. For the weeks we want to predict 12 weeks ahead but we have seen that the best weekly models when forecasting one year ahead also perform well, or are the best, on predicting 12 weeks ahead. Overall we see that the performance of the testing set is better than the tuning set. This is explained by less influence of covid-19 and a more stable demand in this period. The models are quite robust since the ARIMA models tend towards averages when predicting long term ahead (Hyndman and Athanasopoulos, 2018). Therefore, these models perform well on the test set since demand is more stable here. Because of this it needs to be taken into account that when future demand is expected to increase, these models might not perform the best. This is why budgets can be taken into account as explanatory variables. These can incorporate expected increase or decrease in the demand and thereby alter the forecast. Overall we see that the ARIMA models perform the best, let alone the UK. For the UK the random forest model performs better. Also, the UK period is the only aggregation level where one of the hierarchical methods, i.e., top down, bottom up or combination, achieves a better accuracy. Using an ARIMA model for the UK would result in a MAPE of 14.3% and 13.9% for period and week respectively which is an increase of 2% compared to the best models. We can argue if it is worth it to implement the random forest models for Farm Trans since they are only used for the UK. Since for all other aggregation levels the ARIMA models perform, the best the decision is made to implement the ARIMA model as prototype for Farm Trans in a generic way. This will be elaborated in Chapter 6.

Incorporating hierarchy into the forecast has a positive impact on forecasting the Benelux and UK per period but further than that no performance improvement is gained. For the UK we see that the combination approach gives the best results and for the Benelux the hybrid approach. Also for the weekly UK the hybrid model gives the best result. Although it may be said that the top down hierarchical methods come close to the optimized individual models. For example for the global weekly forecast this method reaches a MAPE of 10.7% where the best individual model scores 7.6%. This difference of 3% may not have the biggest impact on the business decision of Farm Trans since this results in one or two trucks more or less required. We strive towards an MAPE which matches the MAPE of the budget of Farm Trans, which was 10.5%.

Aggregation	Best testing	MAPE testing
Global period	ARIMA (0,0,4)	6.7
Global week	ARIMA (0,0,19)	7.6
Benelux period	ARIMA (0,0,4) with hierarchy	4.6
Germany period	ARIMA (1,0,1) and covid-19	13.5
UK period	Random forest	13.1
UK period	Region period+week	12.0
Benelux week	ARIMA (2,0,1)	6.5
Germany week	ARIMA (2,0,0) and covid-19	17.6
UK week	Random forest with hierarchy	12.1

Table 19: Overview of the best models per aggregation level according to the test set

6 Implementation

Now we know what model works the best for each aggregation level, it is important to implement it at Farm Trans. This is discussed in this chapter. As indicated by Couillard (1993), forecasting is an essential piece of information which is needed for the fleet planning process. Therefore the model is implemented within the control tower application, supplied by CAPE Groep to Farm Trans. Implementation is two fold: on the one hand a prototype of the best forecasting method is created for Farm Trans, this is discussed first together with the benefits. On the other hand the model is set up in a generic fashion such that it can be used later on, for example for other parts of the business, like bulk transport, or other clients of CAPE Groep.

6.1 Prototype Farm Trans

At Farm Trans a prototype is implemented within the current application, the control tower. A new forecasting module is created and added to the application. By making use of the already developed software the project is quickly integrated within the organisation.

In Figure 20, an example is given of the graphical representation of the output of the prototype. Not only the forecast, but also the historical demand of the last year is shown to give information about how the business is developing. Furthermore, the number of trucks used is shown. This is included to support the decision makers by including an estimation of the required capacity. The forecast is transformed into the number of trucks required by using the average volume transported per truck in the past year, in this case of the year 2021. This is done for all aggregation levels and for truck as well as for the trailers that are used for transportation. In this way for all aggregation levels an estimation can be made, being that the global period is the most important. This is the highest level and therefore represents the whole fresh and frozen department. Relating it to the forecasting accuracy, as indicated in Section 2.2.1 and 2.4, Improving the accuracy by 1.5% results in that 1 truck and 2.3 trailer less are required on the global period level. This is a quick transformation of the forecast which is sufficient since this is also the current practise regarding the transformation of the budgets to capacity. More research can be done on how to more accurate transform the forecast into required capacity.

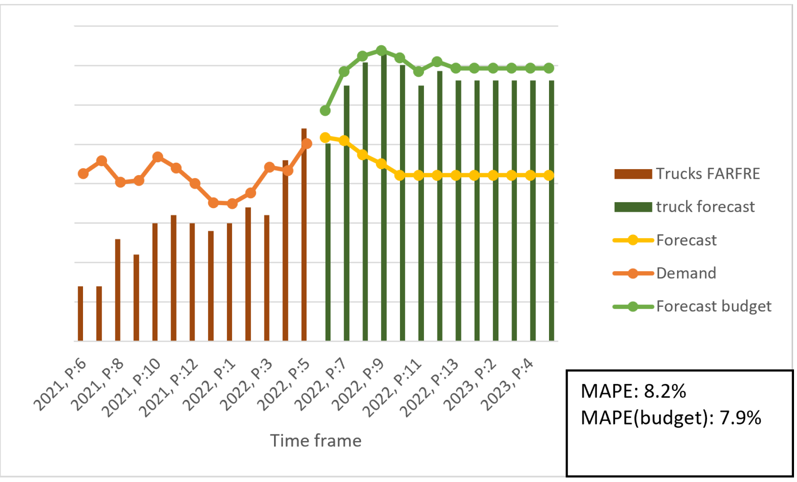


Figure 20: Final forecast ARIMA(0,0,4)

The forecast is incorporated into the weekly reports. From here the information is spread through the organisation through the weekly meetings. It is secured in the organisation and it can be used during the fleet capacity decision process. The current decision making process needs to change to incorporate the forecasts. Instead of using the budget to determine required capacity, the forecast of transportation volume is used. This requires a different working method. The approximation of the capacity gets more accurate because transportation volume better represents the demand of the business than the budget. The budget is vulnerable for increase in fuel prices for example which impacts the fleet capacity decision process. On the basis of historical data from the business, the average transported volume per truck, region and time frame can be estimated. When this is used together with the expected transport volumes, the required capacity can be determined. This leads to a better substantiated estimation of capacity which in turn result in less short term adjustments, which are undesirable. Of course this is impacted by the decision of how much demand Farm Trans wants to fulfill themselves, this is a decision left to the management and currently left out of consideration.

Next to that the new forecast also helps in making timely decisions. As indicated in Table 2, knowing 12 weeks up front supports the tactical decisions and 1 year up front supports strategical decision making. The forecast of transportation volumes is further divided into the regions, where the budget is only globally available. Therefore, the forecasts adds another dimension to the decision making which further substantiates decisions because estimations can be made more accurately when looking at the regions separately. The forecast itself is less accurate the more we aggregate from the global period level, but since this aggregation was first non-existing a better estimation can be made for each aggregation level separately.

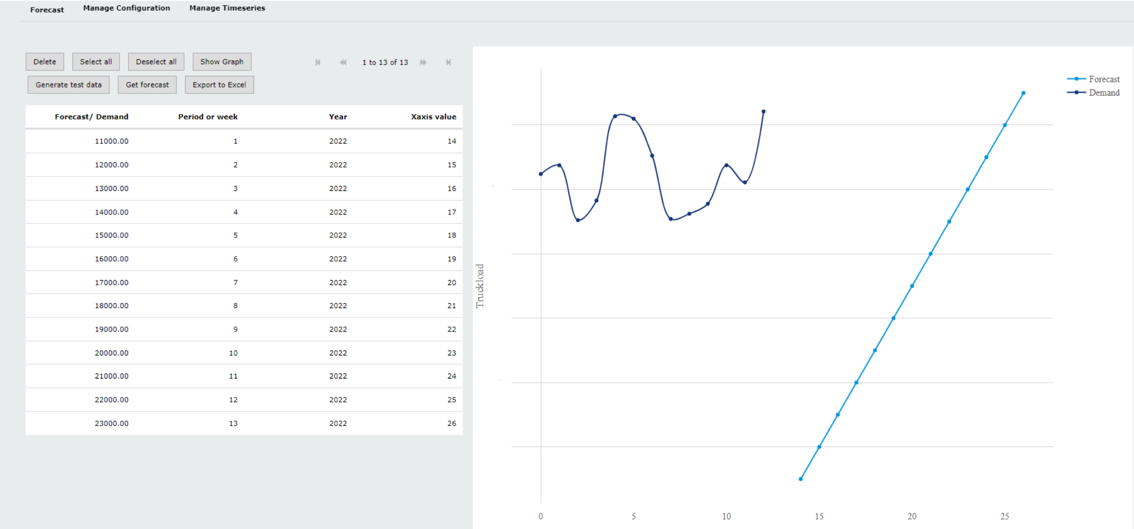
This approach of forecasting and implementation further benefits Farm Trans in the way that no human knowledge about customer behavior, new customers or expected revenue is needed. The budgets are based on that information, this is lost when employees leave Farm Trans. It is important to establish this knowledge into the organisation which this forecasting model contributes towards. Next to that it also creates consciousness of the employees who normally do not think about future demand. It can start the discussion about what can be done if demand is high or low. But also about whether the forecast is accurate or not, what is the expectation of sales and does this match our data driven forecast? This can result in small things like sending drivers on vacation in quite periods.

6.2 Generic model

The model is set up in a generic way such that a time series together with explanatory variables can be provided and a forecast is generated with that information.

6.2.1 How does it work?

The prototype consists of three pages. In Figure 21 the main page of the prototype is shown. The forecast shown here is an example of the forecast, not the real forecast. From the main page, a time series and configuration can be selected for which a forecast is to be generated. To start the process the user needs to click on the button "Get forecast". A pop up appears asking for what aggregation level a forecast needs to be generated. The user can select from a list of levels and click the button "Create forecast" to send the information to the API. The working of the API is discussed in Section 6.2.2. A standard configuration is used, which can be managed in the tab "Manage configuration". The API returns the forecast which is then shown on this page in the form of a graph and a table. The table can be exported to Microsoft Excel to save the information, this needs to be done manually by the user. The buttons: Select all, Deselect all and Delete can be used to select and delete the forecasting information from the table. The button Show Graph gives the graph as pop out and the button "Generate test data" can be used to get an example of the forecast as shown in Figure 21.



Forecast/ Demand	Period or week	Year	Xaxis value
11000.00	1	2022	14
12000.00	2	2022	15
13000.00	3	2022	16
14000.00	4	2022	17
15000.00	5	2022	18
16000.00	6	2022	19
17000.00	7	2022	20
18000.00	8	2022	21
19000.00	9	2022	22
20000.00	10	2022	23
21000.00	11	2022	24
22000.00	12	2022	25
23000.00	13	2022	26

Figure 21: Main page of the forecasting prototype implemented at Farm Trans

The prototype requires input information, namely the historical information about the data that we want to forecast. This can be arranged by an Excel upload in the tab "Manage timeseries". This page consists of two tabs, one for the period and week time frames. An Excel file is provided which need to be in a format consisting of the columns: Year, Period, Global truckload, Benelux truckload, Germany truckload, UK truckload and the explanatory variables which are included, in this case covid-19 and budget. After uploading the information, it is displayed such that it can be checked if the right information is uploaded. When the model is going to be used for different time series this needs to be adjusted.

Lastly we need information about the configuration that is used, this can be managed in the tab "Manage configuration". Here for each aggregation level the standard model to use is already configured, based on the conclusions of this research. The configuration consists of information about model parameters (order, seasonal order and intercept), accuracy iterations, number of periods to predict ahead and the trainperiod. How these are used is explained in Section 6.2.2. The flow of activities to perform in the prototype to get a forecast is also shown in Figure 22. All this information is also specified in a document with instructions and a recording of a demonstration of the prototype is available.

6.2.2 Application Programming Interface (API)

To create the prototype, the ARIMA model is made accessible by making use of an API. To this API the information about the timeseries, configuration and explanatory variables is sent. This is input from the user and needs to be supplied in the prototype. The API returns the forecast for a certain amount of periods ahead together with the accuracy in the past by using cross validation. The choice for an API is made because we want to be able to access the model from any online environment, mainly the application of Farm Trans where the forecasting module is implemented. In this way we incorporate the model in the control tower of Farm Trans and it also becomes possible to implement at other clients.

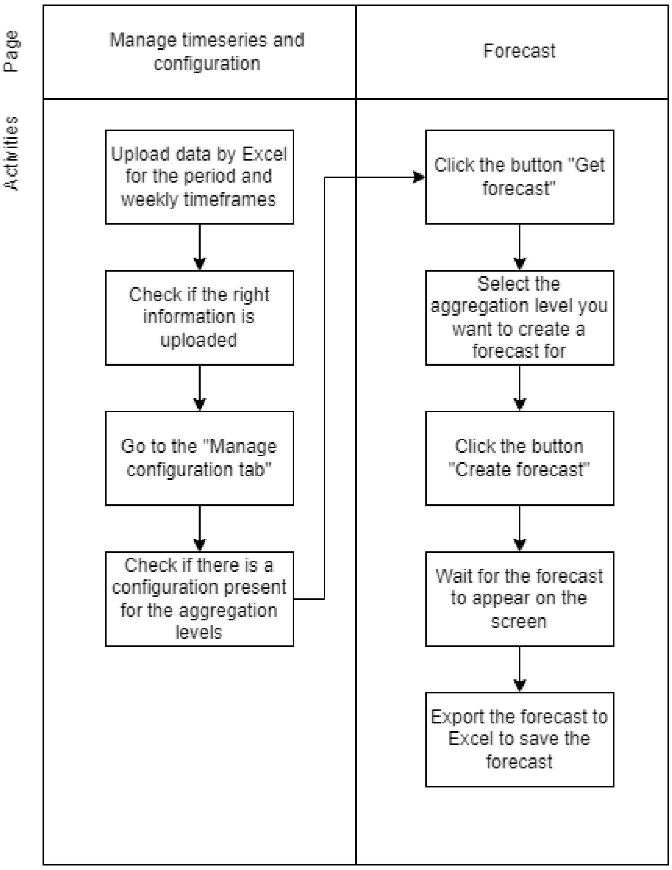

Page
Activities
Manage timeseries and configuration
Forecast
Upload data by Excel for the period and weekly timeframes
Check if the right information is uploaded
Go to the "Manage configuration tab"
Check if there is a configuration present for the aggregation levels
Click the button "Get forecast"
Select the aggregation level you want to create a forecast for
Click the button "Create forecast"
Wait for the forecast to appear on the screen
Export the forecast to Excel to save the forecast

Figure 22: Activities needed to perform in the prototype to get forecast

The API takes some time to start up if not used for a time longer than one hour. This is to save the costs of keeping it running since it is only used once per week. Therefore, when it is first used it can take some time before a response of the model is handled.

The input of the model is a time series together with a configuration of the ARIMA model. The output is a forecast and accuracy based on cross validation. A standard configuration for each aggregation level is used based on this research but these can be altered by the user if needed in the "Manage configuration" page. The model parameters indicate how the ARIMA model behaves and these are explained in Section 3.3.1.1. The number of accuracy iterations is input from the user, standard set to 10, which refers to the number of cross validation iterations that are performed to assess accuracy. This differs a bit from the approach we used throughout the research since now we do not need to divide the data in different data sets. We can perform the accuracy iterations from a point in time to get an idea about the accuracy of the model. This is standard set to the observation in the middle of the time series. If not enough observations are available, the number of iterations is decreased. The trainperiod represents the point in time from where the accuracy iterations are done. Also new configurations can be made, which are then automatically selected as the configuration to use when generating a new forecast.

Because of this, the model can also be implemented for the other business units of Farm Trans. This is out of scope of this research but may very likely be done in the future. Then the accuracy of the different configurations need to be assessed to generate reliable forecasts since the demand patterns are different.

One of the goals of CAPE Groep is to be a strategic partner of their clients. By showing to be innovative and thinking ahead value is created. This research contributes to that because it is a new piece of knowledge that can be applied at clients, showing CAPE is an innovative business partner. Due to the generality and a prototype in the software the forecasting module can be easily incorporated. Therefore easily creating value for their customers. Knowledge is transferred by this research report and a specific document about the working of the model.

6.3 Conclusion

The benefits for both Farm Trans and CAPE Groep are stated in this chapter, thereby answering the two research questions: (1) How can farm trans utilize the forecast and what will be the benefit? and (2) How should decision making processes be altered such that the forecast is incorporated? To conclude, Farm Trans utilizes the forecast in their weekly reports. The forecast is transformed into approximations for fleet capacity per region and time period based on the average volume transported per truck and trailer in 2021. This starts the discussion based on this information. The process should be altered because instead of budgets, transport volumes are used. Therefore an estimation is made by using information about average volumes transported per region and timeframe. Which is the total volume transported divided by the number of truck and trailers used in the certain time period. Here we only look at the demand served by Farm Trans themselves and not the outsourced demand. CAPE Groep benefits from the generality of the model, such that it can be implemented at other clients and this strengthens their position as strategic partner for their clients.

7 Conclusion, discussion and recommendations

This chapter provides the conclusion, discussion and recommendations of the research that is conducted at the analyse and control department of Farm Trans. The main research question was: *How can Farm Trans predict the transportation volumes within the conditioned business unit with the available information?* This question was raised to solve the core problem: *There is no knowledge available within Farm Trans on how to use available data to gain insight in future demand.* Therefore, mainly quantitative research was conducted and by formulating and answering sub questions structure was given to the research.

7.1 Conclusion

We developed a forecasting model to determine the transportation volumes of Farm Trans. The model uses historical information about the transportation volumes and explanatory variables as input. The main usage of the forecast is to support the fleet capacity decision process. We are interested in forecasting multiple aggregation levels which are based on the decision they support. This is per region (Global, Benelux, Germany and United Kingdom) and timeframe (period and week) resulting in 8 aggregation levels. Short term (12 weeks up front) we can alter capacity by chartering and long term (one year) we can purchase new capacity. The most important question was: what forecasting method works the best for Farm Trans at what aggregation level? The better the forecasting accuracy, the better the decision making process can be supported. We calculated that improving the forecast error by 1.5% results in requiring 1 truck and 2.3 trailers less. The decision process does not boil down to one truck more or less but having an error of 10% result in the possibility of having 10 trucks which are not used or short. This has a big impact on the business and, therefore, it is important to generate forecasts with a good accuracy. The goal was set to meet the accuracy of the budget currently used by Farm Trans, which had a MAPE of 10.5% in the last 4 years.

Multiple models and methods are tested and evaluated on their accuracy. We tested ARIMA, exponential smoothing and random forest models extensively in combination with explanatory variables and a hierarchical forecasting structure to determine the best approach for Farm Trans. The data is split up in three data sets for training, tuning and testing the models. First the models are trained and tuned to search for the best configuration of hyper parameters according to the performance on the tuning set using cross validation.
The results of tuning are shown in Table 20. Overall the ARIMA models achieve the highest accuracy on the tuning set. In some cases explanatory variables like the budget and the covid-19 indicator are used to increase accuracy. The global period achieves the best accuracy with the bottom up approach, using the forecast of global week by the ARIMA(0,0,19) model. The main difference is that 19 lagged variables are used instead of 16 week by the global period ARIMA(0,0,4) model, 1 period equals 4 weeks. In combination with the different time frame this produces more accurate results for the short term ahead, since after 19 weeks the forecasts is the average demand. Also for the Benelux period a hierarchical method performs better, but only by a small margin of 0.1%. This is however the combination approach where the forecast from the period and week are combined. The hierarchical methods improve the accuracy of the two aggregations with highest demand, global and Benelux period.

Aggregation	Best tuning model/ method	MAPE tuning
Global period	ARIMA (0,0,4) with budget	7.9
Global period	Bottom up global week	7.2
Global week	ARIMA (0,0,19)	8.4
Benelux period	ARIMA (0,0,4)	6.6
Benelux period	Combination region period + week	6.5
Germany period	ARIMA (1,0,1) with budget and covid-19	23.7
UK period	ARIMA (3,0,2) with budget and covid-19	22.8
Benelux week	ARIMA (2,0,1)	9.3
Germany week	ARIMA (2,0,0) with budget and covid-19	25.4
UK week	Exponential (additive seasonality)	19.5

Table 20: Overview best models per aggregation according to cross validation on the tuning data set

For the other aggregation levels the best tuning performance is achieved with individual models where often explanatory variables are included. Mainly the budget has proven to be of good predictive value since 4 out of 8 models benefit from it. The budget prone to price fluctuations which negatively impacts the predictive power. Including budget has a positive effect on the global period but not on the week. This can be explained by the fact that the budget is created per period. Therefore, the weekly fluctuations are not incorporated into the budget and thus both increases and decreases in demand explained by the same budget. Only for Germany week the budget is included but this performance is mainly due to the covid-19 indicator and not the budget. Because we see a clear relationship between covid-19 and the demand in Germany. For the other regions this was not noticed which is why it is not included in all models. The GDP is never beneficial for accuracy of the models, thereby concluding that Farm Trans her business is not much affected by an increase or decrease in GDP. GDP has increased, with some hick ups, over the years where the demand has been more stable.

The optimal configurations are also tested on the test set, this is important because this is information the model has not yet seen before. The results determine the models that are going to be used for implementation and are shown in Table 21. This are the models that are recommended to use for each aggregation level. However, since only the UK benefits from a model different from the ARIMA it is chosen to not implement the random forest model. The benefit is small, 2%, and using one model for all aggregation levels is more comprehensible. The results differ a bit from the tuning results. The performance of the testing set is better than the tuning set which is explained by less influence of covid-19 and a more stable demand in this period. The models are quite robust when predicting multiple periods into the future. Combining this with a more stable demand in the test set results in a good forecasting performance. Therefore, the forecasting performance needs to be taken into account in the future. When the demand patterns change, which could be the case due to changing market situations, maybe other forecasting methods will achieve a better performance. Including the budget as explanatory variable is not as beneficial as in the tuning set. This is because the budget increased while the demand did not, therefore losing its predictive value.

Aggregation	Best testing	MAPE testing
Global period	ARIMA (0,0,4)	6.7
Global week	ARIMA (0,0,19)	7.6
Benelux period	ARIMA (0,0,4) with hierarchy	4.6
Germany period	ARIMA (1,0,1) with covid-19	13.5
UK period	Random forest	13.1
UK period	Combination region period+week	12.0
Benelux week	ARIMA (2,0,1)	6.5
Germany week	ARIMA (2,0,0) with covid-19	17.6
UK week	Random forest with hierarchy	12.1

Table 21: Overview of the best models per aggregation level according to the test set

7.1.1 Implementation

Implementation of this research was twofold. A prototype of the best forecasting model is created for Farm Trans and also this model is set up in a generic way, meaning that it can also be used by other business units of Farm Trans. One of the goals of CAPE Groep is to be a strategic partner of their clients. This research contributes to that end since the model can be used for any timeseries to be forecast and therefore it can be implemented at other clients. The gained forecasting knowledge is usable by the use of this research and the created prototype together with instructions. The forecast for Farm Trans is incorporated into their weekly reports. Together with historical information average volume transported per truck and trailer, an estimation of the required capacity is made by Farm Trans. This is input for the capacity decision process, where first there were no estimations, now there is a data driven forecast and estimation on which the discussion is based. Therefore the capacity decision process changes. Together with explicit knowledge of the employees and the forecast a better decision can be made. Farm trans indicated that this can be a difficult process and that the forecast helps in creating more insight in the future demand.

7.2 Discussion

One of the difficulties of this research is the selection of what models and methods to test for this particular situation. A lot of combinations can be made and therefore choices must be made. We have tested ARIMA, exponential smoothing and random forest models in combination with the usage of explanatory variables and hierarchical forecasting methods. From the literature we have seen that ARIMA models often result in a good performance and since the simple version of ARIMA deemed effective also exponential smoothing models where tested. To incorporate the effect of a more sophisticated model the random forest was tested. There are still many more models which could be tested which maybe result in a better accuracy. We however achieved results accurate to 6.7% for the global period, the most important aggregation level, which is comparable to the performance of the budget. Also, the chance to increase that accuracy even more is small. We for example quickly assessed LSTM, a promising method according to the literature, which showed that it behaves the same as the other models. The models tend towards robustness for predicting long time periods ahead. Furthermore, due to the available time for the research we have limited the research to the named models and methods. Furthermore, the research incorporates novelty of hierarchical forecasting methods. Where (Zhang, 2003) forecasted a timeseries by ARIMA and fed it into a ML algorithm to forecast the same level, we made use of hierarchy. We forecasted a higher level aggregation time series and fed it into the model of an other aggregation level to assess the performance. This was proven to be effective in some cases. We mainly saw the higher aggregation levels, like the Benelux period, to benefit from this approach.

Another point of discussion is the impact of covid-19 on the data. We have seen that volumes decreased at the moment of the lockdowns. This is in the middle of the data set that was available and therefore this had an impact on the results. Using only data before or after that moment resulting in having to less observations to setup training, tuning and testing data. It was assessed what the impact would be if only the observations with covid-19 impact are used. This resulted in an improvement for one aggregation level, global period. Therefore it was decided to use all available data because of comprehensibility. It may in a later point in time be possible that deleting the oldest observations results in a higher accuracy since these observations do not represent the current situation anymore. This is a point for future research if deemed necessary by Farm Trans or CAPE Groep.

Regarding the implementation of the prototype the wish was to have it working automatic as much as possible. This was not possible in the time frame of this research. Setting up the required data structure took more time than intended and was therefore not possible. However, implementation shows the practical relevance of this research. By implementing the forecasting model as prototype in the control tower, Farm Trans can easily make use of the model and it is ensured that the research is really used by Farm Trans.

7.3 Recommendations and further research

The main objective of the research was to develop a model to forecast transportation volumes. This model is developed in Python, and implemented in the control tower application of Farm Trans. To ensure future usage of the model we propose some recommendations as well as possible future research possibilities.

7.3.1 Forecast usage

The forecasts need to be made each week and period for their respective aggregation levels. This needs to be done manually by Farm Trans since data needs to be uploaded to the model. Farm Trans can choose themselves whether to include the explanatory variables, like the budget, in their forecast. Currently, this is not recommended since the budget increased in the test period where demand did not. Therefore, including the budget led to worse forecasting performance. If the budget is including this needs to be taken into account. Another option is using a budget without fuel prices included to create a predictor which is more stable. Furthermore, it is important to keep track of the forecasting performance. This is incorporated into the model by showing the accuracy based on cross validation using half of the data supplied to the model. Because we have worked with covid-19 impacted data, it may be the case that a different model performs better in the future. Therefore this accuracy number needs to be taken into account when generating the forecasts.

7.3.2 Fleet capacity planning

A subject for further research is fleet capacity planning. Now that a forecast of the transportation volumes is available this is used to estimate required capacity by taking into account historical averages, i.e., how much volume a truck can transport in a week or period. This can be further researched by the fleet size and composition vehicle routing problem (FSCVRP) for example. Here the goal is to decide on the fleet size and composition of the vehicles needed to complete all orders. There should be a connection made between the orders and the transportation volume in this case such that the forecast can be translated to required capacity. Another option which is less time consuming would be to also account for the turnover rate of a truck, instead of only the average volumes transported per time period. This will result in more accurate representation of required capacity. Due to time reasons this is not taken into account in this research.

7.3.3 Implementation

As discussed before, implementation can be further extended by connecting the data warehouse of Farm Trans to the forecasting model. Currently the input data needs to be uploaded and the forecast exported manually. By connecting the database, the required information can be retrieved automatically at the end of each week or period and the forecast send back. This automation step was not possible in the current time window of the research due to the data infrastructure that needed to be set up for this. Furthermore, the forecasting model can also be implemented at the other business units of Farm Trans, for example at bulk transport. This requires re-assessing the model configuration. This can already be done in the current prototype since it returns the model accuracy based on the data. Also new data may be required since other explanatory variables may provide a high predictive power. The main steps taken in this research can be used as guidelines to implement the model at the other business units.

References

Andrawis, R. R., Atiya, A. F. and El-Shishiny, H. (2011), 'Combination of long term and short term forecasts, with application to tourism demand forecasting', *International Journal of Forecasting* **27**(3), 870–886.

Andreoni, A. and Postorino, M. N. (2006), 'A multivariate arima model to forecast air transport demand', *Association for European Transport and contributors* **1**(1).

Archer, B. H. (1980), 'Forecasting demand: Quantitative and intuitive techniques', *International Journal of Tourism Management* **1**(1), 5–12.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N. and Petropoulos, F. (2017), 'Forecasting with temporal hierarchies', *European Journal of Operational Research* **262**(1), 60–74.

Bauwens, L., Koop, G., Korobilis, D. and Rombouts, J. V. (2015), 'The contribution of structural break models to forecasting macroeconomic series', *Journal of Applied Econometrics* **30**(4), 596–620.

Baykasoğlu, A., Subulan, K., Taşan, A. S. and Dudaklı, N. (2019), 'A review of fleet planning problems in single and multimodal transportation systems', *Transportmetrica A: Transport Science* **15**(2), 631–697.

Brockwell, P. J. and Davis, R. A. (2016), *Introduction to Time Series and Forecasting*, 3rd edition edn, Springer.

Brown, R. G. (1959), *Statistical forecasting for inventory control*, McGraw-Hill.

cbs (2022*a*), 'Centraal economisch plan 2022', https://www.cpb.nl/forecasts.

cbs (2022*b*), 'Horeca; omzetontwikkeling, index 2015=100', https://www.cbs.nl/nl-nl/cijfers/detail/83862NED.

cbs (2022*c*), 'Opbouw binnenlands product (bbp); nationale rekeningen', https://opendata.cbs.nl/#/CBS/nl/dataset/84087NED/table.

Chollet, F. et al. (2015), 'Keras'. accessed 11-08-2022.
URL: *https://github.com/fchollet/keras*

Christopher, M. (2011), *Logistics and Supply Chain Management*, 4th edition edn, Pearson Education Limited.

Couillard, J. (1993), 'A decision support system for vehicle fleet planning', *Decision support systems* **9**(2), 149–159.

Croxton, K., Lambert, D., García-Dastugue, S. and Rogers, D. (2002), 'The demand management process', *The International Journal of Logistics Management* **13**(1), 51–66.

Daum, D., Prak, D. and Minner, S. (2021), 'A macro-micro framework for truck rate forecasting', *Unpublished working paper* .

Engle, R. F. and Brown, S. J. (1986), 'Model selection for forecasting', *Applied Mathematics and Computation* **20**(3), 313–327.

Fildes, R., Goodwin, P., Lawrence, M. and Nikolopoulos, K. (2009), 'Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning', *International Journal of Forecasting* **25**(1), 3–23.

Heerkens, H. and van Winden, A. (2012), *Geen probleem: Een aanpak voor bedrijfskundige vragen en mysteries*, van Winden Communicatie.

Holt, C. C. (1957), 'Forecasting seasonals and trends by exponentially weighted averages', *International Journal of Forecasting* **20**, 5–10.

Hutter, F., Kotthoff, L. and Vanschoren, J. (2018), *Automated Mahine Learning; Methods, Systems, Challenges*, 1st edition edn, Springer.

Hyndman, R. J. and Athanasopoulos, G. (2018), *Forecasting: Principles and Practice*, 2nd edition edn, OTexts.

Hyndman, R. J. and Khandakar, Y. (2008), 'Automatic time series forecasting: The forecast package for r', *Journal of Statistical Software* **27**(3), 1–22.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. and Grose, S. (2002), 'A state space framework for automatic forecasting using exponential smoothing methods', *International Journal of Forecasting* **18**, 439–454.

Hyndman, R. and Koehler, A. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**(4), 679–688.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021), *An Introduction to Statistical Learning*, 2nd edition edn, Springer.

Kwiatkowski, D., Phillips, P. C., Schmidt, P. and Shin, Y. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?', *Journal of Econometrics* **54**(1), 159–178.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Petropoulos, F. (2022), 'Forecasting: theory and practice', *International Journal of Forecasting* **1**(1).

Probst, P., Wright, M. N. and Boulesteix, A.-L. (2019), 'Hyperparameters and tuning strategies for random forest', *WIREs Data Mining and Knowledge Discovery* **9**(3), e1301.

Profillidis, V. A. and Botzoris, G. N. (2019), *Modelling of transport demand: Analyzing, Calculating and Forecasting Transport Demand*, Elsevier.

PWC (2017), 'Shifting patterns: the future of the logistics industry', https://www.pwc.com/gx/en/industries/transportation-logistics/publications/the-future-of-the-logistics-industry.html. accessed on 12-04-2022.

Rijksoverheid (2022), 'Coronavirus tijdlijn', https://www.rijksoverheid.nl/onderwerpen/coronavirus-tijdlijn.

Rohaan, D., Topan, E. and Groothuis-Oudshoorn, C. (2022), ‘Using supervised machine learning for b2b sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider’, *Expert Systems with Applications* **188**, 115925.

Rudakov, K., Strizhov, V. and Kashirin, D. (2017), ‘A multivariate arima model to forecast air transport demand’, *Autom Remote Control* **1**(78), 75–87.

Schröer, C., Kruse, F. and Gómez, J. M. (2021), ‘A systematic literature review on applying crisp-dm process model’, *Procedia Computer Science* **181**(1), 526–534.

Seabold, S. and Perktold, J. (2010), ‘Statsmodels: Econometric and statistical modeling with python’.

Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018), A comparison of arima and lstm in forecasting time series, *in* ‘2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)’, pp. 1394–1401.

Singh, A., Thakur, N. and Sharma, A. (2016), ‘A review of supervised machine learning algorithms’, **1**(1), 1310–1315.

Smith, T. G. et al. (2017), ‘pmdarima: Arima estimators for Python’. accessed 11-08-2022.
URL: *http://www.alkaline-ml.com/pmdarima*

van Wee, B. and Annema, J. A. (2012), *The Transport System and Transport Policy - An Introduction*, 1st edition edn, Edward Elgar Publishing Ltd.

Winters, P. R. (1960), ‘Forecasting sales by exponentially weighted moving averages’, *Management Science* **6**, 324–342.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H. and Deng, S.-H. (2019), ‘Hyperparameter optimization for machine learning models based on bayesian optimizationb’, *Journal of Electronic Science and Technology* **17**(1), 26–40.

Zhang, G. (2003), ‘Time series forecasting using a hybrid arima and neural network model’, *Neurocomputing* **50**(1), 159–175.

Zotteri, G. and Kalchschmidt, M. (2007), ‘Forecasting practices: Empirical evidence and a framework for research’, *International Journal of Production Economics* **108**, 84–99.

Zotteri, G., Kalchschmidt, M. and Caniato, F. (2005), ‘The impact of aggregation level on forecasting performance’, *International Journal of Production Economics* **93-94**, 479–491. Proceedings of the Twelfth International Symposium on Inventories.

8 Appendix

8.1 Decision tree global weekly forecast one week ahead

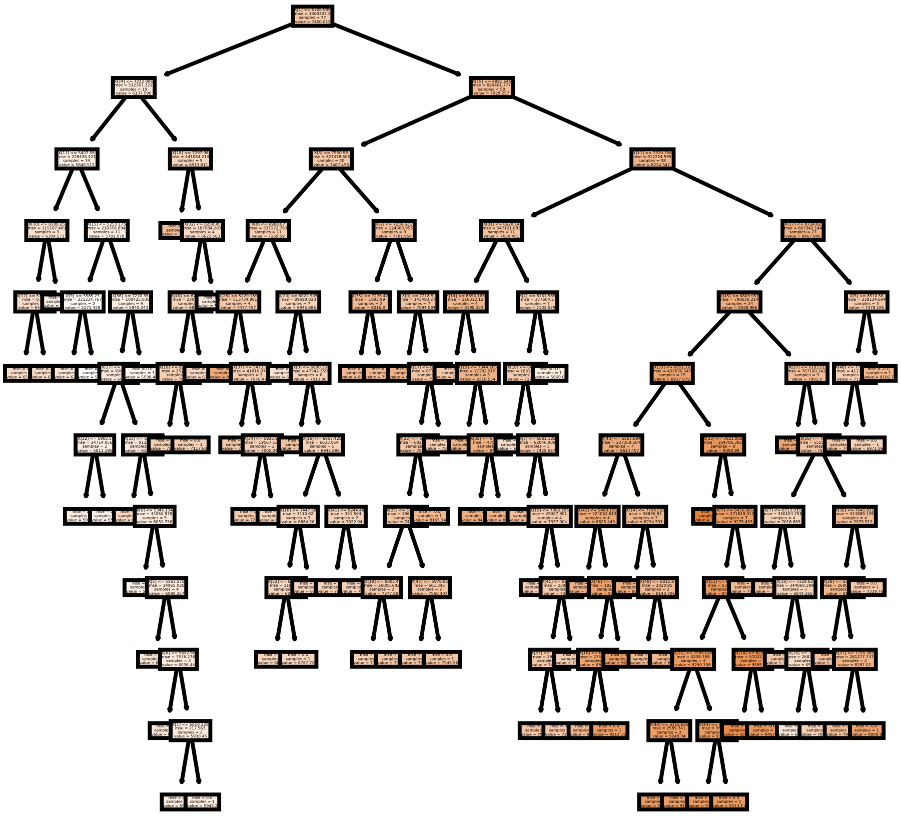
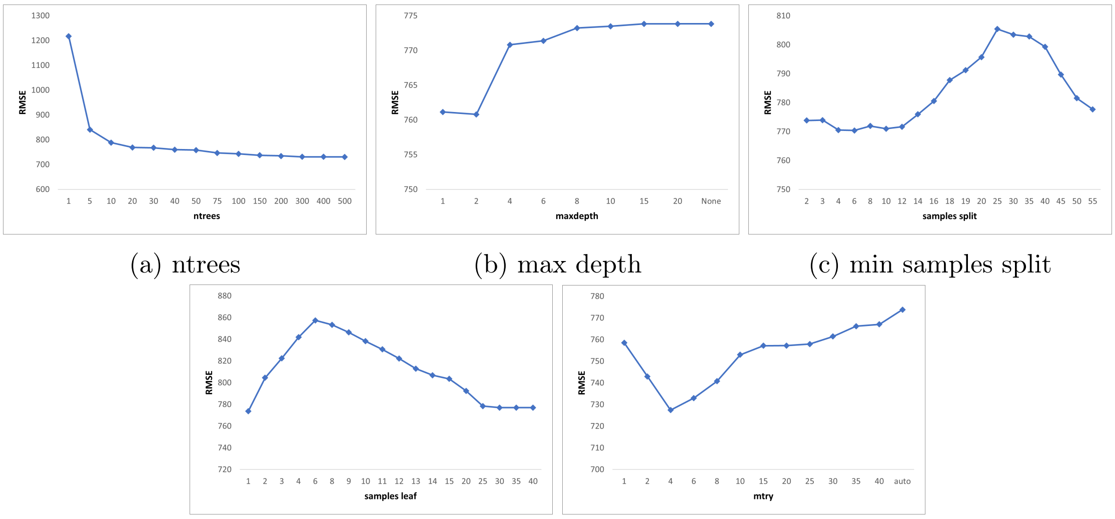


Figure 23: Decision tree 1 of many, forecasting global weekly

8.2 Tuning results weekly region



(a) ntrees

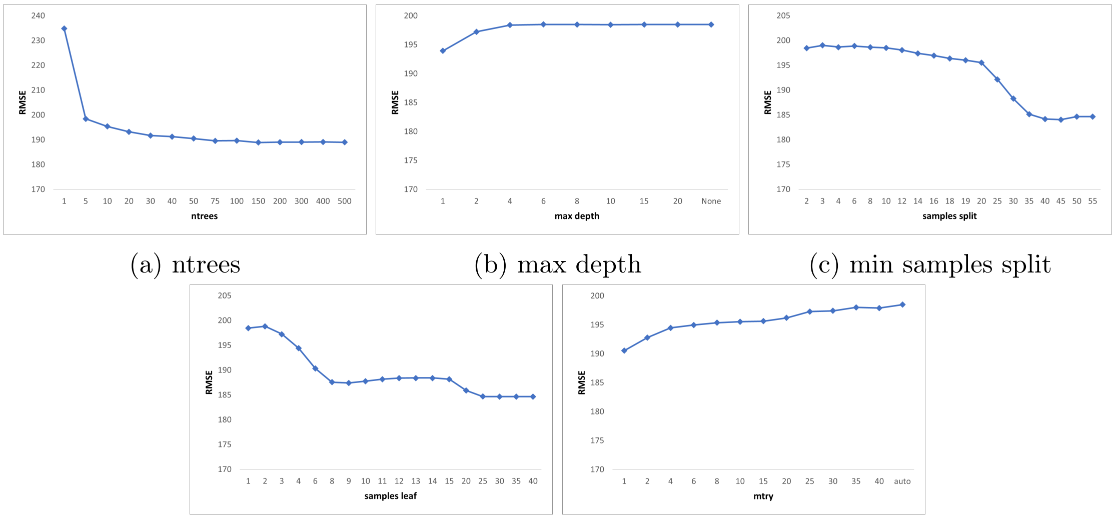
(b) max depth

(c) min samples split

(d) min samples leaf

(e) max features

Figure 24: Tuning random forest hyper parameters weekly forecast Benelux



(a) ntrees

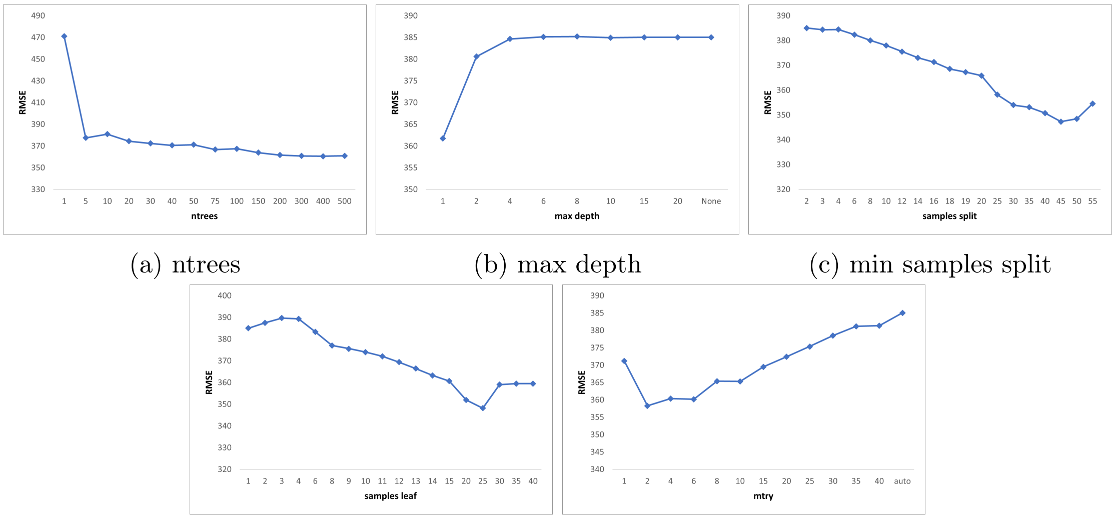
(b) max depth

(c) min samples split

(d) min samples leaf

(e) max features

Figure 25: Tuning random forest hyper parameters weekly forecast Germany



(a) ntrees

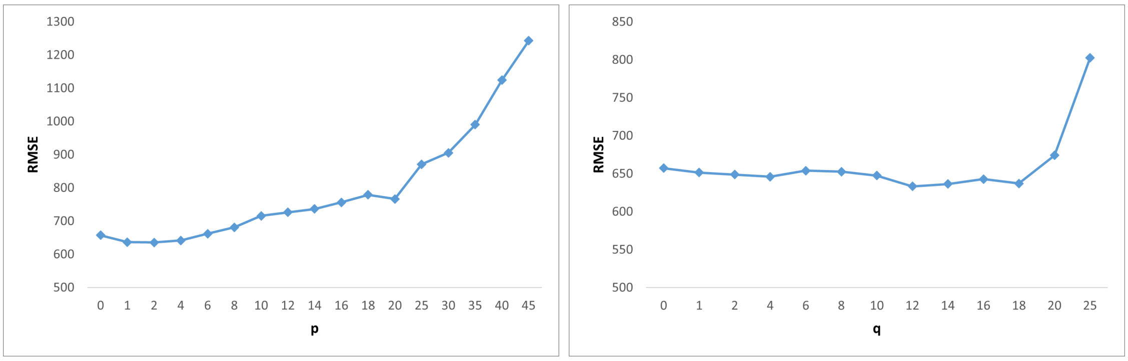
(b) max depth

(c) min samples split

(d) min samples leaf

(e) max features

Figure 26: Tuning random forest hyper parameters weekly forecast UK



(a) p

(b) q

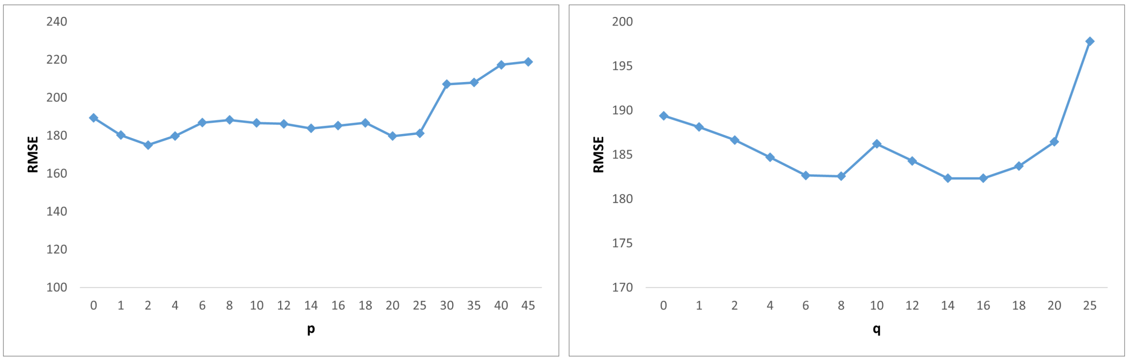
Figure 27: Tuning ARIMA hyper parameters weekly forecast Benelux

8.3 Read me forecasting API

8.3.1 Forecasting timeseries with ARIMA API

Created to generate a forecast based on input of a timeseries.

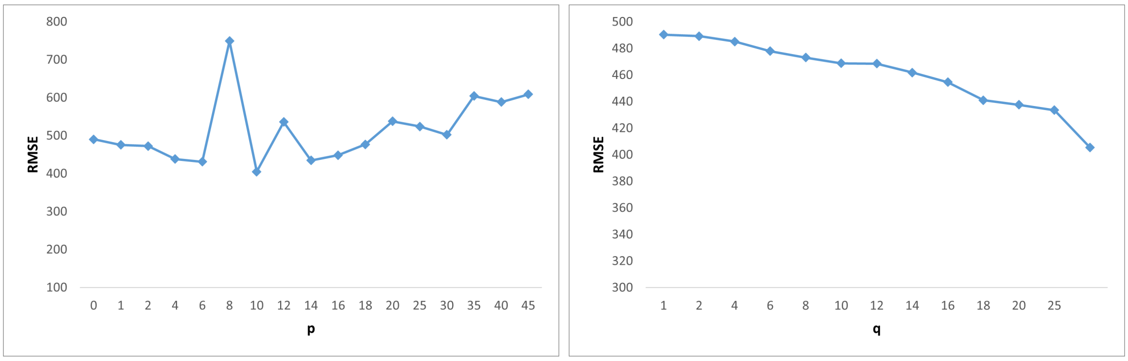
8.3.1.1 Description This API is created for a graduation project about forecasting. It set up in a generic way such that it may be used for different organisations or purposes. Based on the input given, the model is trained with all the data and returns a forecast some periods ahead from that point onwards. What input is required is specified below. The code also runs a cross validation on the data, using "AcuraccyIteration" iterations from the point "TrainPeriods". This essentially fits the model multiple times and calculates the error corresponding to the corresponding real data and gives back an accuracy measure. In the config it is specified what ARIMA(p,d,q) model is used "Order" and "SeasonalOrder", how many periods u want to use for training, how many periods you want to forecast ahead and the number of accuracy iterations. For more information on how ARIMA works, the following website can be consulted: https://otexts.com/fpp2/arima.html Explanatory variables are optional, if they are not included no information needs to be sent to the API.



(a) p

(b) q

Figure 28: Tuning ARIMA hyper parameters weekly forecast Germany



(a) p

(b) q

Figure 29: Tuning ARIMA hyper parameters weekly forecast UK

8.3.1.2 Getting Started Python v3.8 with packages: Pandas 1.1.3 Json 0.9.5 Numpy 1.22.2 Pmdarima 1.8.5 Math 3.8 (standard python)

8.3.1.3 Executing program The main function is Main(JsonTimeseries, Config, ExplanVariables = 0) The function is called via an API. When run locally through the following url: http://127.0.0.1:5000/Arima?Config=&Timeseries=&ExplanatoryVariables=

1. Create input data is formatted in the following manner: Below in the appendix a request example is shown. Timeseries = “Data” : float, “Year” : int, “PeriodOrWeek” : int Config= "Order": "p": "int", "d": "int", "q": "int" , "SeasonalOrder": "P": "int", "D": "int", "Q": "int", "PeriodsPerYear": "int" , "Intercept": “Boolean” "Trainperiod": int, "PredictAhead": int, "AccuracyIterations": int

Optional: ExplanatoryVariables= “Data” : float, “Year” : int, “PeriodOrWeek” : int

2. Output is formatted in the following manner: Output = ‘Forecast’ : ‘data’: float, ‘Year’ : int, ‘PeriodOrWeek’ : int, ‘Accuracy’: ‘MAE’: float, ‘MAPE’: float, ‘MSE’: float, ‘RMSE’: float

8.3.2 Request example

http://127.0.0.1:5000/Arima?Config=%22Order%22:%20%22p%22:1,%20%22d%22:1,%22q%22:1,%22Season
[19823.999999999964, 23531.89999999997, 28735.959999999963, 28060.01999999997, 27542.199999999986, 24637.92999999996, 26728.809999999958, 25507.339999999975, 22564.10999999995, 23580.999999999967, 22979.59999999995, 23423.499999999938, 20344.31724137928, 23168.879999999957, 21417.659999999978,

24742.969999999965, 25398.949999999964, 24763.279999999966, 26989.90999999998, 32006.859999999997, 29620.899999999976, 28341.360999999968, 30378.780999999966, 32659.299999999974, 33468.85999999998, 32293.04129032254, 32388.495999999966, 36907.85099999996, 37436.27999999994, 28225.733999999982, 30147.25999999998, 27248.579999999973, 27052.30999999996, 28224.45059999997, 31714.899999999965, 29764.459999999963, 30950.61999999996, 31317.349999999962, 31277.866666666625, 25522.223999999973, 29808.859999999946, 30397.46999999996, 27913.831999999966, 27158.150099999973, 29294.37999999998, 30927.720999999958, 28173.849999999962, 28420.969999999954, 31423.639999999967, 29993.15999999996, 28015.95999999996, 25613.6533333333, 25502.739999999965, 26842.07999999997, 30102.99999999996, 29658.99999999997, 33109.85999999996], "Year": [2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2022, 2022, 2022, 2022, 2022], "PeriodOrWeek": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5]&ExplanatoryVariables="Data": [1473660.1429503844, 1642269.2571428847, 1646842.56216571, 1650007.1488107098, 1584573.1808782087, 1827425.3240907104, 1838026.9889782092, 1803006.9264282102, 1762055.6218407096, 1740050.3386957105, 1627341.2655422164, 1564224.6421589763, 1470960.7144648938, 1017545.1791028298, 1205403.023335969, 1529680.2907790432, 1804858.2000708324, 1824654.8376745952, 1666846.8164229065, 1728749.2736187696, 1576731.345617525, 1643482.4187848214, 1779152.9319127297, 1703273.9033896732, 1621288.9971017342, 1497488.2075726546, 1509735.739271177, 1761493.2719124302, 1855020.0859983307, 1854298.6875345993, 1819447.2962169996, 1915579.8677359994, 1982061.3226145005, 1939514.1155767997, 1897573.147692698, 1899307.8369904999, 1912651.8213557997, 1864622.1731399999, 1726931.1561005446, 1710218.6653512232, 1864582.544130449, 2019236.2366378694, 2035092.584457545, 2109453.7615546533, 2272283.227424595, 2419850.885982056, 2373380.791958266, 2354498.7120367414, 2332033.9873386966, 2223501.923862819, 2305362.4080670807, 2154931.4560021027, 2342599.7188164, 2562865.334456711, 2694845.930871951, 2876104.130249147, 3041444.338931827], "Year": [2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2022, 2022, 2022, 2022, 2022], "PeriodOrWeek": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5]

8.4 Global period forecast results

8.4.1 ARIMA models on training and test set

Table 22: Results of ARIMA models, 13 period ahead forecast. training data = 2018-2020, test = 2021

ARIMA model (p,d,q)	AIC	MSE	RMSE	MAPE	MAE
(0,0,0)	765	3,926,522	1,982	5.6	1,635
(0,0,1)	746	4,378,574	2,093	5.8	1,662
(0,1,0)	712	26,702,559	5,167	16.9	4,721
(0,1,1)	715	16,149,092	4,019	12.8	3,563
(1,0,0)	732	4,512,747	2,124	5.7	1,617
(1,0,1)	733	4,489,064	2,119	5.6	1,577
(1,1,0)	714	23,743,509	4,873	15.9	4,431
(1,1,1)	716	26,923,417	5,189	17.0	4,739

8.4.2 ARIMA models 13 period ahead forecast with cross validation

Table 23: Results of ARIMA models, 13 period ahead forecast. cross-validation

ARIMA model (p,d,q)	MSE	RMSE	MAPE	MAE
(0,0,0)	16,533,636	3,880	10.0	3,166
(0,0,1)	13,974,131	3,584	9.5	2,991
(0,0,2)	13,352,854	3,532	9.6	2,991
(0,0,3)	12,435,204	3,424	9.2	2,855
(1,0,0)	12,335,414	3,457	9.3	2,813
(1,0,1)	19,249,872	4,086	12.3	3,646
(1,0,2)	18,403,668	3,933	11.4	3,410
(1,0,3)	23,357,481	4,530	13.7	4,059
(2,0,0)	14,870,536	3,776	10.4	3,115
(2,0,1)	34,257,622	5,409	16.6	4,966
(2,0,2)	20,156,522	4,190	12.6	3,740
(2,0,3)	31,547,622	5,355	15.1	4,592
(3,0,0)	23,445,707	4,252	13.2	3,853
(3,0,1)	27,738,075	4,934	14.8	4,430
(3,0,2)	17,813,656	3,957	11.9	3,471
(3,0,3)	23,179,288	4,472	13.0	3,865
(4,0,0)	316,997,412	11,980	30.4	8,959
(4,0,1)	116,719,143	8,279	22.0	6,449
(4,0,2)	17,264,650	3,971	11.8	3,470
(4,0,3)	150,327,068	8,077	21.4	6,292
(0,0,4)	8,975,663	2,946	8.2	2,516
(0,0,5)	8,994,049	2,965	8.2	2,579
(0,0,6)	14,664,233	3,700	10.2	2,937
(1,0,4)	21,996,794	4,557	13.3	3,934
(2,0,4)	19,627,484	4,331	12.7	3,755
(3,0,4)	24,436,064	4,765	14.5	4,287
(4,0,4)	22,424,334	4,608	13.7	4,008

8.4.3 Best ARIMA(0,0,4) model with explanatory variables

Table 24: Results of ARIMA models, 13 period ahead forecast. cross-validation with explanatory variables

ARIMA model (p,d,q)	Added explanatory variable	MSE	RMSE	MAPE	MAE
(0,0,4)	0	8,975,663	2,946	8.2	2,516
(0,0,4)	budget	8,602,655	2,904	7.9	2,369
(0,0,4)	corona	262,610,504	12,852	39.9	11,745
(0,0,4)	gdp	14,495,778	3,759	10.1	3,126
(0,0,4)	bc	12,454,861	3,368	9.6	2,856
(0,0,4)	bg	25,104,812	4,578	12.0	3,743
(0,0,4)	cg	14,429,469	3,601	9.8	2,998
(0,0,4)	bcg	26,221,235	4,584	12.2	3,762

8.4.4 Exponential smoothing results cross validation

Table 25: Results of exponential smoothing models 13 periods ahead using cross validation

Model(trend,seasonal)*	MSE	RMSE	MAPE	MAE
A,A	106,615,789	9,040	28.7	8,344
A,M	96,203,459	8,483	26.6	7,700
A,N	54,496,456	6,600	21.2	6,127
M,A	204,289,456	12,247	38.4	11,154
M,M	117,605,885	9,254	29.0	8,425
M,N	356,860,513	15,512	46.6	13,633
N,A	34,894,818	5,490	16.8	4,822
N,M	44,147,588	6,178	18.7	5,373
N,N	26,666,167	4,732	14.7	4,245

* A = additive, M = multiplicative, N = none

Table 26: Results of exponential smoothing models 1 period ahead using cross validation

Model(trend,seasonal)*	MSE	RMSE	MAPE	MAE
A,A	11,321,932	3,352	9.0	2,605
A,M	11,884,128	3,433	9.2	2,670
A,N	7,634,343	2,747	7.4	2,106
M,A	11,434,194	3,368	9.1	2,626
M,M	12,005,214	3,450	9.2	2,678
M,N	7,613,687	2,742	7.4	2,098
N,A	9,022,447	2,995	7.8	2,243
N,M	11,552,862	3,389	9.2	2,651
N,N	7,411,604	2,711	7.4	2,129

* A = additive, M = multiplicative, N = none

8.4.5 Random forest tuning results cross validation (not all results are shown)

Table 27: Results of RF models, 13 period ahead forecast. cross-validation

RF model *	MSE	RMSE	MAPE	MAE
(auto,None,2,1,1)	23,173,217	4,763	12.8	3,718
(auto,None,2,1,5)	17,606,275	4,152	11.4	3,306
(auto,None,2,1,10)	15,665,959	3,945	10.9	3,182
(auto,None,2,1,20)	16,201,437	4,004	11.5	3,361
(auto,None,2,1,30)	16,496,113	4,041	11.7	3,398
(auto,None,2,1,40)	16,488,200	4,034	11.7	3,394
(auto,None,2,1,50)	16,072,167	3,982	11.5	3,349
(auto,1,2,1,50)	14,467,519	3,781	10.6	3,103
(auto,2,2,1,50)	16,377,072	4,024	11.6	3,381
(auto,3,2,1,50)	16,357,368	4,016	11.6	3,365
(auto,4,2,1,50)	16,177,302	3,995	11.5	3,355
(auto,5,2,1,50)	16,122,837	3,989	11.5	3,354
(auto,None,2,1,50)	16,072,167	3,982	11.5	3,349
(auto,None,2,1,50)	16,072,167	3,982	11.5	3,349
(auto,None,3,1,50)	16,186,996	3,994	11.5	3,360
(auto,None,4,1,50)	16,652,192	4,054	11.7	3,407
(auto,None,5,1,50)	16,307,205	4,020	11.5	3,366
(auto,None,2,1,50)	16,072,167	3,982	11.5	3,349
(auto,None,2,2,50)	15,114,021	3,872	11.0	3,217
(auto,None,2,3,50)	14,425,520	3,785	10.5	3,064
(auto,None,2,4,50)	14,283,071	3,765	10.3	3,028
(auto,None,2,5,50)	14,545,738	3,801	10.2	3,012
(1,None,2,1,50)	13,149,168	3,599	9.7	2,836
(2,None,2,1,50)	14,176,662	3,727	10.4	3,051
(3,None,2,1,50)	15,048,732	3,842	10.9	3,188
(4,None,2,1,50)	15,557,998	3,913	11.1	3,237
(5,None,2,1,50)	15,718,344	3,937	11.2	3,270

(mtry, maxdepth, samplessplit, samplesleaf, ntrees)

8.4.6 Random forest tuning results cross validation

Table 28: Results of RF models, 13 period ahead forecast. cross-validation

RF model *	MSE	RMSE	MAPE	MAE
(1,None,2,1,50) budget	13,314,427	3,616	9.5	2,803
(1,None,2,1,50) corona	13,286,796	3,610	9.5	2,791
(1,None,2,1,50) gdp	13,171,197	3,592	9.4	2,781
(1,None,2,1,50) budget,corona	13,850,067	3,695	10.6	3,084
(1,None,2,1,50) budget,gdp	14,298,875	3,759	10.8	3,137
(1,None,2,1,50) corona,gdp	14,127,272	3,730	10.6	3,091
(1,None,2,1,50) budget,corona,gdp	15,190,290	3,873	11.0	3,196

(mtry, maxdepth, samplessplit, samplesleaf, ntrees)

8.5 Global weekly results

8.5.1 ARIMA models 53 period ahead forecast with cross validation

Table 29: Results of tuning q in ARIMA, 53 weeks ahead, cross validation

ARIMA model(p,d,q)	MSE	RMSE	MAPE	MAE
(0,0,0)	1,217,846	1,074	10.8	859
(0,0,1)	1,156,953	1,047	10.6	838
(0,0,2)	1,111,175	1,028	10.4	821
(0,0,4)	1,035,530	993	10.1	795
(0,0,6)	1,046,028	999	10.3	805
(0,0,8)	1,012,620	985	10.3	801
(0,0,10)	1,060,452	1,012	10.6	826
(0,0,12)	1,002,768	986	10.2	794
(0,0,14)	930,708	952	10.0	769
(0,0,16)	888,719	934	9.7	739
(0,0,18)	825,086	902	9.5	712
(0,0,19)	805,344	893	9.5	708
(0,0,20)	821,378	898	9.6	707
(0,0,21)	822,121	900	9.8	718
(0,0,25)	1224805	1079	12.7	893

Table 30: Results of tuning p in ARIMA, 53 weeks ahead, cross validation

ARIMA model(p,d,q)	MSE	RMSE	MAPE	MAE
(0,0,0)	1,217,846	1,074	10.8	859
(1,0,0)	923,983	952	10.1	772
(2,0,0)	931,301	956	10.2	768
(4,0,0)	1,028,710	998	11.0	809
(6,0,0)	1,390,576	1,130	13.2	948
(8,0,0)	1,509,410	1,173	13.9	993
(10,0,0)	4,840,525	1,549	17.7	1,273
(12,0,0)	54,245,125	5,466	55.3	4,114
(14,0,0)	14,104,046	2,743	28.7	2,109
(16,0,0)	75,453,476	6,877	69.4	5,164
(18,0,0)	53,641,409	5,849	58.9	4,386
(20,0,0)	3,285,972	1,655	19.0	1,374
(25,0,0)	2,277,034	1,486	17.1	1,244

8.6 Exponential smoothing weekly results cross validation

Table 31: Results of exponential smoothing models, 53 weeks ahead, cross validation

Model(trend,seasonal)*	MSE	RMSE	MAPE	MAE
A,A	15,245,795	3,273	41.2	2,966
A,M	8,458,373	2,476	30.9	2,222
A,N	7,672,616	2,541	32.1	2,315
M,A	21,127,673	4,232	52.7	3,804
M,M	20,711,747	3,906	48.9	3,537
M,N	11,825,969	2,991	37.4	2,704
N,A	3,042,454	1,639	20.0	1,415
N,M	3,230,645	1,694	20.3	1,440
N,N	2,282,059	1,410	17.4	1,234

* A = additive, M = multiplicative, N = none

8.7 Random forest weekly tuning results cross validation (not all results are shown)

Table 32: Results of RF models, 53 weeks ahead forecast. cross-validation

RF model *	MSE	RMSE	MAPE	MAE
(auto,None,2,1,1)	3,145,274	1,756	20.2	1,439
(auto,None,2,1,5)	2,251,150	1,478	17.6	1,240
(auto,None,2,1,10)	2,207,172	1,461	17.6	1,237
(auto,None,2,1,20)	2,150,801	1,441	17.4	1,223
(auto,None,2,1,30)	2,116,674	1,430	17.3	1,213
(auto,None,2,1,40)	2,135,883	1,435	17.4	1,221
(auto,None,2,1,50)	2,144,200	1,437	17.4	1,222
(auto,1,2,1,50)	1,836,329	1,347	16.5	1,160
(auto,2,2,1,50)	2,071,517	1,417	17.2	1,208
(auto,4,2,1,50)	2,148,578	1,438	17.4	1,225
(auto,6,2,1,50)	2,137,166	1,435	17.4	1,219
(auto,None,2,1,50)	2,144,200	1,437	17.4	1,222
(auto,None,2,1,50)	2,144,200	1,437	17.4	1,222
(auto,None,3,1,50)	2,153,162	1,439	17.4	1,223
(auto,None,4,1,50)	2,131,034	1,432	17.3	1,218
(auto,None,5,1,50)	2,152,963	1,439	17.5	1,226
(auto,None,2,1,50)	2,144,200	1,437	17.4	1,222
(auto,None,2,2,50)	2,074,681	1,414	17.1	1,202
(auto,None,2,3,50)	2,015,948	1,397	17.0	1,189
(auto,None,2,4,50)	2,025,889	1,404	17.1	1,197
(auto,None,2,5,50)	2,048,038	1,414	17.3	1,210
(1,None,2,1,50)	1,696,032	1,286	15.6	1,095
(2,None,2,1,50)	1,765,958	1,309	15.9	1,110
(4,None,2,1,50)	1,859,589	1,337	16.2	1,136
(6,None,2,1,50)	1,908,364	1,355	16.4	1,152
(8,None,2,1,50)	1,953,122	1,370	16.6	1,164
(10,None,2,1,50)	1,972,062	1,378	16.7	1,169
(15,None,2,1,50)	2,059,874	1,407	17.1	1,196

(mtry, maxdepth, samplessplit, samplesleaf, ntrees)