Forecasting Demand within the Chemical Industry

A comparison of machine learning and traditional forecasting models

Master's Thesis

submitted for the degree of Master of Science in

Industrial Engineering and Management

August 2024

R. J. Gerhardus

VIVOCHEM
PART OF BÜFA

Supervisor Vivochem:

P. Slaghekke

UNIVERSITY OF TWENTE.

Supervisors University of Twente:
Dr. M. C. van der Heijden
Dr. L. Xie

Forecasting Demand within the Chemical Industry

Date
16-08-2024

Student
R.J. (Ruben) Gerhardus
Industrial Engineering & Management (IEM)
Production & Logistics Management (PLM)

Supervisor Vivochem
P. (Patrick) Slaghekke
Plant Manager, Vivochem B.V.

Vivochem B.V.
Darwin 5, 7609 RL Almelo
www.vivochem.com

Supervisors University of Twente
First supervisor:
Dr. M.C. (Matthieu) van der Heijden
Associate Professor, Industrial Engineering & Business Information Systems

Second supervisor:
Dr. L. (Lin) Xie
Assistant Professor, Industrial Engineering & Business Information Systems

University of Twente
Faculty of Behavioural, Management and Social Sciences
Drienerlolaan 5, 7522 NB Enschede
www.utwente.nl

Acknowledgement

Dear reader, I hope you will enjoy reading my thesis, '*Forecasting Demand within the Chemical Industry*,' which I submitted for completion of the master's degree in Industrial Engineering & Management. This marks the end of my academic education at the University of Twente. The research effort was conducted for Vivochem B.V. under the supervision of Patrick Slaghekke.

First, I would like to thank Vivochem for the opportunity to do this research. I also express my gratitude to Patrick for his guidance and encouragement, from which I have learned much. I extend this appreciation to my colleagues, who made my time at Vivochem very enjoyable. I look forward to continuing to work with you. Secondly, I thank Matthieu and Lin for being my supervisors. Your valuable feedback and recommendations enabled me to challenge myself, for which I am thankful.

Furthermore, I would like to express the immense enjoyment I experienced studying for the past seven years with my two friends, Sten and Twan. These years will always remain special to me. Additionally, I need to mention the support from my mother, family, and friends, for which I am forever thankful. Lastly, I thank my father, who passed away, not knowing I would continue studying for my master's; you will always inspire me.

Deo gratias,

Ruben Gerhardus

Enschede, 16th of August 2024

Management Summary

This research on forecasting was conducted for Vivochem, a chemical distribution company located in Almelo, Netherlands. Demand patterns within the chemical industry are subject to irregularities, making forecasting difficult. Despite this difficulty, it is necessary to predict future demand to ensure customers can be served with the appropriate inventory. When forecasts are unreliable, inventory levels increase to maintain the customer service level. The rise in inventory results in increased inventory costs. Therefore, the objective of our research is:

'To research and develop a forecasting model suitable for sporadic chemical demand patterns that lowers inventory costs while maintaining a high service rate.'

By analysing the current situation, we determined that the current forecasting method is non-systematic. A forecast with the current process is generated only after a pre-determined inventory level is reached, and this forecast is the average demand for the past three months. This forecast is then subject to judgemental interpretation, making it not fully quantitative. This method must be revised to forecast demand as it does not incorporate peaks and intermittent demand periods. We use shipment data of base items to represent historical demand. Forecasting is done per base item, representing the chemical to which products belong. We categorised these base items into lumpy, intermittent, erratic, and smooth demand.

Our literature research found that statistical models are less fitted for this demand but generally perform better than more complex models. In more recent years, there has been a change in the status quo; *Machine Learning* (ML) models have started to outperform traditional forecasting models. Literature indicates that the *K Nearest Neighbour* (KNN), *Light Gradient Boosting Machine* (LightGBM), and *Neural Network* (NN) models are well-performing methods which could be applicable. In addition, we included the statistical models Holt and ARIMA as they are commonly used methods and offer a valuable comparison. We include a moving average of the demand over the past three months to represent the current forecasting model, referring to it as the benchmark.

From the *Enterprise Resource Planning* (ERP) system, we collected 12 years of data on 29% of base items (constituting 139 base items) responsible for almost 80% of inventory costs, with a weekly time bucket. The ARIMA and Holt models were trained solely on this historical demand data. ML models were trained on this data and additional variables, such as feature engineered variables from the historical data set and exogenous data from external sources. Model parameters were optimised using various methods to minimise forecasting errors. The forecasting models were trained on 80% of the data and tested on the remaining 20%. We evaluated the forecasting performance of the models with the following measurements: bias, *Mean Squared Error* (MSE), *Symmetric Mean Absolute Percentage Error* (sMAPE) and *Mean Absolute Error* (MAE). We evaluated inventory performance with a measurement that is based on the service rate, inventory cost, and order variance called *Root Means Squared* (RMS). To estimate the inventory performance, we created a *continuous Order Up To* (OUT) model to simulate the inventory, where the safety stock (ss) and order up to a point (S) are based on the forecasts. The performances of the forecasting models were compared to the benchmark models' performances.

We decided to consider a forecasting model to improve forecasting when two measurements (either MSE, sMAPE or MAE) indicated lower error values than those of the benchmark model. This was necessary as there was a degree of inconsistency in the number of improved base items according to the different measurements. The main observation from our results is the relatively good performance of ML models compared to the statistical models. ML models improved significantly more base items than the statistical models at an improvement threshold of 5% and 10%, with the best ML model (NN) improving more than twice as many base items as the best statistical model (ARIMA). This indicates that ML models are a better fit for sporadic demand patterns. Additionally, we observed certain models improving more base items from specific demand types than other models; all ML models (especially the LightGBM model) performed well for lumpy base items. Holt improved a high number of smooth base items while the NN improved more (base items) at the higher threshold. NN and LightGBM both performed well for the intermittent and erratic classes.

As stated, we trained the ML models on three different data sets. The second data set (with exogenous variables) improved a small number of base items (not improved without this data) for the KNN and NN models (15 and 20, respectively). The exogenous variables, gas price index, oil price index, and *Production Price Index* (PPI), seemingly improved the forecasting of these models. However, we could not find an explanation of why these variables improved for these specific chemicals from a chemical, demand type, or application perspective. We observed the third data set (with endogenous data) improving not more than five base items for each model.

Next, we researched the performance of the models when extending the forecast horizon to six and 26 weeks. For longer horizons, we observed that the ML models have a significant increase (almost doubling for all models) in the number of improved base items at higher thresholds of improvement. This indicates that these are significantly more accurate than the benchmark model at longer horizons. We observed a less significant increase at the higher threshold for the statistical models, especially for the Holt model. This indicated that all models maintain their relative performance at greater horizons. Regarding data sets, models (and especially the KNN model) trained on the second data set improved more base items at extended horizons. The third data decreased in the number for the KNN and NN models, only increasing to eight for the LightGBM. Considering that the third data set still improved only a few base items, it is of little added value.

Regarding inventory, according to the RMS measurement, almost all ML models improved 40 more base items compared to the statistical models. This indicates that the forecasting improvement of ML models can translate into a potential improvement in inventory performance. We hypothesized a weak correlation between RMS and bias, as the ML models all had a strong performance according to both measurements, while both were lower for the statistical models.

Considering the model's performances and minimising the number of new models, we evaluated five possible solutions: NN, NN with LightGBM for lumpy demand, NN trained on both data sets one and two (selectively applied based on inventory performance), NN with Holt for smooth demand, and lightGBM. These five solutions are chosen due to their relatively high performance for the corresponding demand types and overall performance. We evaluated the inventory reduction based on the inventory simulation with the OUT policy. The solution with only the NN model trained on historical data had the most significant reduction in inventory while maintaining an acceptable service rate. This solution reduced an average of 22,3% of inventory (in weight units) for 65 base items and improved the service rate (fill rate) by an average of four percentage points for 19 base items.

We recommend implementing the proposed solution according to the implementation plan. This will reduce inventory, free up the associated financial costs, improve customer service, and enable purchasing to make better decisions. The improved performance of the new forecasting model will enable Vivochem to more cost-efficiently align inventory to customer demand.

To improve upon this research effort, we have several recommendations:

- Researching if a systematic inventory policy (currently non-existent) can be developed as it relates to how forecasting performance is translated to inventory.
- Determining if there is an economic benefit by extending the forecast horizon, as we observed our NN model to maintain its performance at greater horizons.
- Researching the performance of aggregated forecasts or hybrid models since we only researched pure models during this effort.
- Researching the possibility of adding more data, such as daily data and demand from the company group, to the models to improve accuracy.
- Developing a forecasting model to forecast product-specific demand, allowing Vivochem to predict which packaging should be used during the dispensing process.
- Improving the reliability of the requested delivery date so that it can replace the shipment date, thereby improving the data quality.

List of Abbreviations

ADR	*Accord européen relatif au transport international des marchandises Dangereuses par Route*
AI	*Artificial Intelligence*
ERP	*Enterprise Resource Planning*
GBM	*Gradient boosting machine*
GDP	*Gross Domestic Product*
IBC	*Intermediate bulk container*
KNN	*k-Nearest Neighbours*
KPI	*Key Performance Indicator*
MAE	*Mean absolute error*
ML	*Machine learning*
MPSM	*Managerial Problem-solving method*
MSE	*Mean squared error*
NN	*Neural Network*
PPI	*Production Price Index*
RMS	*Root mean square*
SKU	*Stock Keeping Unit*
sMAPE	*Symmetric Mean Absolute Percentage Error*
SS	*Safety Stock*

Content

1. Introduction

In the first chapter, we introduce Vivochem B.V. (referred to as Vivochem from now on) and define this research effort's core problem and the corresponding research approach.

1.1 Company

Vivochem is a chemical distribution company based in Almelo, The Netherlands. Located at the 'XL Business Park' close to the logistical centre: Port of Twente. Customers are supplied with basic chemicals from a product range containing more than 450 chemicals. These chemicals are not produced by Vivochem but are supplied by other distribution companies or directly from the producers. Chemicals arrive mostly in bulk (via tanker trucks) and, for a lesser part, as general cargo (via regular lorries). The chemicals that arrive in bulk form are dispensed in smaller quantities.

In addition to distribution, Vivochem also offers supplementary services: drumming, warehousing, and exporting. With drumming, customers can enjoy preferred packaging that is in accordance with their desired quality standards and quantitative needs. Packages range from small cans of twenty litres to IBC (Intermediate Bulk Container) of 1000 litres. Examples of packages can be seen in Figure 1. Vivochem offers warehousing to customers who do not have adequate storage conditions (due to regulatory or capacity restrictions). Vivochem's modern warehouse can store chemicals according to ADR[1] regulations, ensuring safety and quality. To customers desiring international transportation of their chemical goods, Vivochem offers its export service. Thereby unburdening customers with all administrative and legal workloads involved with exporting chemicals.

Vivochem provides value to customers by enabling them to purchase smaller quantities of chemicals, which is impossible at the producers. This is made possible by the numerous services provided by Vivochem. Customers who use these services are found in the following industries (but not limited to): agriculture, coating and adhesive, cleaning, technical, and food. Chemicals are shipped to customers throughout the Benelux and exported further throughout Europe.

Since 2021, Vivochem has been part of the BÜFA group, which, in addition to chemical distribution, is active in the cleaning (BÜFA Cleaning) and composites (BÜFA Composites) markets.

Figure 1: 200 litre steel pallet drums, IBC of 1000 litre, and cans of 20 litre (Vivochem, 2024).

[1] Abbreviation for: *"Europese overeenkomst inzake het internationale vervoer van gevaarlijke goederen over de weg"* which is the Dutch reference to the French: *"Accord européen relatif au transport international des marchandises Dangereuses par Route."*

1.2 Problem Description

In today's globalised market, companies must perform business operations as efficiently and effectively as possible to stay competitive. For distribution companies like Vivochem, its ability to serve and supply its customers entirely depends on its inventory. The contents of its inventory are replenished according to the expected demand. Currently, Vivochem employs no systematic method to support its inventory management. The parameters of its inventory management are based on the expertise and experience of its employees in the sales and purchasing departments. While practical, it will, however, result in an inefficient inventory as human intuitively based forecasting performs worse than systematic data-driven techniques (Makridakis et al., 1993). The inefficiencies resulting from the current forecasting process negatively affect Vivochem's ability to serve its customers.

Customers could experience longer lead times when backorders occur due to stockouts resulting from an inaccurate demand expectation. This leads to a worse customer experience and could push customers to seek out competitors. Service rates are kept high by increasing order quantities. The inventory cost will increase when orders (to suppliers) increase more than demand (due to more cycle inventory), resulting in less profit. Operational capacity must increase to account for the larger quantities, increasing costs as well. The lack of an appropriate forecasting model represents the research gap. The demand patterns within the chemical distribution industry are known for their sporadic characteristics, having relatively high irregularities such as periods of intermittence and bursts in demand. The core problem can be summarised into one statement: *The current lack of knowledge on quantitative forecasting for sporadic demand, together with the unavailability of an applicable methodology, increases various costs to maintain a high service rate.*

For an enterprise like Vivochem, developing such a forecasting system is essential to stay competitive. (Axsäter, 2015). The objective of this research is thus: *'To research and develop a forecasting model suitable for sporadic chemical demand patterns that lowers inventory costs while maintaining a high service rate.'*

1.3 Research Structure

The main research question corresponding to the objective is: *How can the sporadic chemical demand be forecasted while maintaining an acceptable service rate and reducing inventory?* The research question is supported by several supplementary research questions (referred to as RQs). Answering these questions will provide intermediate results, which, when put together, will answer the main research question and achieve the objective.

- RQ1: What is the current state of forecasting and inventory?
  - RQ1.1: What data are available that can be considered relevant to the expected demand?
  - RQ1.2: What is the quality of available historical data?
  - RQ1.3: What is the quantity of available historical data?
  - RQ1.4: Which criteria can be used to select SKUs with the highest potential for cost savings?
  - RQ1.5: How can the SKUs be classified?
  - RQ1.6: How are forecasts currently produced and used?
  - RQ1.7: What are the requirements for a forecasting system?
- RQ2: Which forecasting methods and measurements are available in the literature?
  - RQ2.1: Which methods are available in the literature?
  - RQ2.2: How can the forecasting accuracy be measured?
  - RQ2.3: How can the inventory performance be measured?
- RQ3: How can the performance of the forecasting methods be evaluated?
  - RQ3.1: How is data prepared and sourced?
  - RQ3.2: Which methods are to be developed into models?
  - RQ3.3: Which parameters need to be set for the methods?
  - RQ3.4: How can we assess model performance?
  - RQ3.5: How can we validate and verify the models?
- RQ4: How are the models designed?
  - RQ4.1: How should the forecasting models be configured?
  - RQ4.2: How should the inventory model be designed?
- RQ5: What is the performance of the models?

  - RQ5.1: Which models perform best according to the forecasting performance?
  - RQ5.2: Which models perform best according to forecasting performance for the different types of demand?
  - RQ5.3: Does more data improve forecasting for ML models?
  - RQ5.4: How does a longer forecast horizon affect forecasting performances?
  - RQ5.5: Do model performances differ according to inventory performance compared to forecasting performance?
  - RQ5.6: Which model(s) should be implemented based on our results?
- RQ6: How can the most appropriate method(s) be implemented?

The research phases will be structured according to the *Managerial Problem-Solving Method* (MPSM) from (Heerkens et al., 2017). However, the phases after implementation are left out and not considered part of this research. Phase 1, *Defining the Problem,* has already been conducted, and phase 2, *Formulating the Approach*, is done in this chapter. This method allows for creative problem-solving while remaining a systematic approach.

1.3.1 Phase 3 Analysing the Problem.

Knowing which data is available and its quality is essential to evaluate the applicability of methodologies during the literature research. For this purpose, we have created RQ1.1 (*What data are available that can be considered relevant to the expected demand?)*. The question is supported by RQ1.2 *(What is the quality of available data?)* and RQ1.3 *(What is the quantity of available data?).* Data quality and quantity are two critical parameters that must be answered to develop a workable solution. These questions will be answered by investigating the data within the *enterprise resource planning* (ERP) system. RQ1.4 (*How can SKUs be classified?)* requires appropriate solutions per class of SKU as they might not all show the same type of sporadic demand. For this purpose, we have created RQ1.5 (*Which criteria can be used to select SKUs with the highest potential for cost savings?*). This allows the research to focus on the area where it will have the most impact (in reducing cost savings).

The current situation needs to be defined before any literature research can be conducted. RQ1.6 (*How are forecasts currently produced and used?)* will qualitatively be answered by mapping the forecasting processes. A quantitative state of the art should be answered by estimating the performance of the current state of forecasting. After describing the current state of the art, the requirements for a new system should be defined, thus answering RQ1.7 *(What are the requirements for a forecasting system?).* Answering these research questions should enable us to conduct the literature research in a specific manner, only selecting appropriate literature by which a solution can be developed. RQ1 (*What is the current state of forecasting and inventory?)* has thus been answered.

1.3.2 Phase 4 Formulating Solutions

By researching the literature, a solution will be developed in accordance with the latest scientific developments. For this purpose, we have created RQ2 *(Which forecasting methods and measurements are available in the literature?).* Methods related to forecasting SKU demand will be researched, and selection will be based on their expected applicability to the context of this research. With this answering: RQ2.1 (*Which methods are available in the literature?).* Appropriate measurements for forecasting performance are needed for this purpose: we have created RQ2.2 *(How can the forecasting accuracy be measured?)* and RQ2.3 *(How can the inventory performance be measured?).* RQ2.2 is used to assess the accuracy of the methods. RQ2.3 estimates the effect forecasting will have on the inventory processes. This is necessary as forecasting may be accurate, but its purpose enable inventory management to reduce costs while maintaining an acceptable service rate.

1.3.3 Phase 5 Choosing a solution.

The knowledge of the state of the art (provided by the answers to the sub-questions from RQ1) enables the selection of applicable methods from the relevant literature (provided by the answer from RQ2). The selected methods can be quantified with the performance measures (provided by the answer from RQ2.2 and RQ2.3). We combine the information to answer RQ3 (*How can the performances of the forecasting methods be evaluated?)*. The methodology includes the data preparation and cleaning: RQ3.1 (*How is data prepared and sourced?)*. We need to translate the methods from the literature research into functional models; thus, we need to define which models are to be developed: RQ3.2: (*Which methods are to be developed into models?*). We need to define the parameters to be set: RQ3.3

(*Which parameters need to be set for models?).* For these models, we need to determine how to measure their performance: RQ3.4: (*How can we assess model performance?*). The models and measurements should be validated and verified to ensure replicability and reproducibility: RQ3.5 (*How can we validate and verify the models?).*

After determining how we produce verified results from validated models, we still need to configure our models with RQ4: (*How are the models designed?).* We need to set the many parameters for the forecasting methods: *RQ4.1 (How should the forecasting models be configured?).* To estimate the models' inventory performance, we must design the inventory model: RQ4.2 (*How should the inventory model be designed?*).

After developing the models and inventory system, RQ5 *(What is the performance of the models?)* should be answered. The forecasting performance of the models should be evaluated: RQ5.1 (*Which models perform best according to the forecasting performance?)* as well as the specific performances for the different demand types: RQ5.2 *(Which models perform best according to forecasting performance for the different types of demand?*). For the ML models, we want to analyse how different data sources affect their performance, specifically RQ5.3 (*Does more data improve forecasting for ML models?*).After establishing the forecasting performances, we want to see how these performances are affected by increasing the forecast horizon: RQ5.4 (*How does a longer forecast horizon affect forecasting performances?).* Answering this will provide insight into the robustness and limitations of the models. We will analyse the results from an inventory perspective: RQ5.5 (*Does the forecasting performances of models differ compared to inventory performance?*). Combining all results allows us to answer RQ5.6: (*Which model(s) should be implemented based on our results?*).

1.3.4 Phase 6 Implementation

The implementation will not be part of this research, but a implementation plan will be formulated on how to successfully achieve it. The last remaining RQ6 (*How can the most appropriate model(s) be implemented?)* will be answered based on the results of phase 5.

After all research phases have been conducted and all research questions answered, the main research question (*How can the sporadic chemical demand be forecasted while maintaining an acceptable service rate and reducing inventory?)* is answered.

1.4 Scope and limitations

The research is subject to limitations to ensure that it can be conducted within the limited time and achieve a realistic objective. Firstly, the research is limited to researching literature on methodologies with similar data, meaning with comparable quantity and quality. If a method cannot be operationalised due to dissimilar data, it will not be considered. Regarding exogenous (meaning not directly demand-related) data sources, since little is known about their correlation with demand, inclusion will done in consultation with the purchasing and sales departments. Secondly, only quantitative techniques will be considered. Adding a quantitative data-based system is assumed to improve forecasts, thus leaving qualitative techniques out of scope. Additionally, only pure models will be researched as hybrid models require much more development time and would extend the range of possible models extensively. A requirement for including a method within this research is that previous research must have shown potential for said method within a similar context. Thirdly, the forecasting will only be done on a selection of chemical products. This selection will be based on inventory costs and data availability; chemical products responsible for high inventory costs will be selected but only if the product has sufficient historical data. Furthermore, we will research alternative forecasting conditions compared to the current standard but will not optimise models for different conditions. This would require more research effort, and alternative conditions would be merely used to estimate the robustness of the models. Next, advice will be given on implementing the developed model(s), but its implementation will not be considered part of this research. Lastly, since economic and strategic considerations drive order quantities and lot sizes, optimising these will not be considered part of this research.

1.5 Deliverables

This research effort will yield primarily a quantitative solution based on relevant scientific literature, which can be used to forecast SKU demand within the chemical distribution industry. The product of this

research will most probably be a prototype of a single or multiple models. Additionally, an estimation of the solution's effect on inventory. Lastly, an advisory section (with recommendations) will be part of the final report to support the successful implementation of a solution.

The problem has been defined in the section: *Problem Description.* An appropriate approach has been formulated in the section: *Research Structure.* This chapter concludes the first and second research phases *(Defining the Problem* and *Formulating the Approach)*.

2. Context Analysis

This chapter presents and analyses all relevant information about the forecasting system (or lack thereof). Its purpose is to answer RQ1 (*What is the current state of forecasting and inventory?).* In section: *2.1 Data Availability*, we discuss the data from a quantity and quality perspective. In section: *2.2 Demand Patterns*, we discuss the types of demand patterns found for the products of Vivochem and how these can be classified. In the last section: *2.3 Forecasting*, we discuss the current forecasting process.

2.1 Data Availability

Within this section, RQ1.1 (*What data are available that can be considered relevant to the expected demand?*), RQ1.2 (*What is the quality of available historical data?)* and *RQ1.3 (What is the quantity of available historical data?*) will be answered. Firstly, the information found in the data will be explained. Secondly, the period of the data will be defined. Lastly, quality will be estimated by investigating the degree of missing data, as well as accuracy and sensitivity to misrepresentation. Data will be selected based on this quality assessment.

2.1.1 Historic Data

Sales data is a time series data set related to the customer's orders. An order contains the quantity of a chemical, the order date, the requested delivery date, the shipping date and the delivery date. The order date is simply the date customers send in an order. The requested delivery date is the date a customer requests a product. The promised delivery date is the expected delivery date, which is subject to change. The shipping date is the date an order physically leaves Vivochem.

Sales orders for Vivochem can be sourced from the current ERP system for almost three years. An additional ten years can be sourced from the old ERP system. Regarding the quality of ERP data, the shipping and order dates are quite reliable since these dates are used within the order intake and expedition processes and are thus updated accordingly. Delivery date could be less accurate if chemicals are subject to a (more) complex supply chain with more risk of delay. The requested delivery would be the best date to use. However, this date has only been systematically tracked for one year. Additionally, when there is no requested delivery, it is estimated and is not updated automatically after further contact with a customer. Due to the less reliable nature of the requested delivery date and delivery date, these will not be used. Order dates could be the most reliable data. However, customers can (and do) occasionally order months in advance; thus, this date can be skewed. The shipment date is continually automatically updated to the actual shipping date, is available for the complete data set, and is thus chosen to represent historical demand for the time series. This data is available for all SKUs over 12 years.

2.1.2 Other Data Sources

According to the sales and purchasing departments, other factors are (potentially) related to the demand for Vivochem’s chemical products. It is unknown if this relation is causal or correlational and how strong said relation is. The factors are gas prices, oil prices, economic growth/decline (national, Chinese, USA, and EU), and ambient temperature. Gas and oil prices influence both the production (including transportation) of chemicals and application processes and influence demand. Macroeconomic growth/decline indicates the need for production in a region and thus for the demand. Ambient temperature affects chemical processes and applications based on the chemicals' properties and thus could be related to their demand. This data will be sourced from official governmental sources.

2.2 Demand Patterns

This section will answer RQ1.4 (*Which criteria can be used to select SKUs with the highest potential for cost savings?*) and RQ1.5 (*How can SKUs be classified?*).

2.2.1 Selection of SKUs

A base item number refers to a chemical; different SKUs can belong to the same base item number. The total number of SKUs is around 1400, corresponding to 600 base items. Products (SKUs) differ in concentration and packaging but can belong to the same chemical (base item). Orders to suppliers are

based on base item numbers. Therefore, we decided to forecast the level of base items. It is more representative of the demand for chemicals as well. We start aggregating by base item number, starting with all base items. The average inventory was calculated for the base items with the corresponding cost value. This showed that 29% (of base items) are responsible for 80% of inventory cost. Thus, we focus on this impactful group of base items. Due to the massive quantities and high-cost price of these chemicals, the potential for improvement is highest for these base items. Since forecasting depends on historical data, only base items on stock in the past three years are considered. This is done to ensure sufficient sales data exists to allow for forecasting. Filtering for sales in the past three years results in 97% of base items (from the previous 29%) remaining, constituting 139 base items.

2.2.2 Classification of SKUs

All SKUs show very sporadic behaviour. Numerous demand peaks and intermediate periods of no (or less) demand exist. Even though this is already a type of demand, we can classify demand even more precisely. According to (Williams, 1984) there are four types of sporadic demand, which can be seen in Figure 2. This two-dimensional classification system allows for better classification than other one-dimensional (classification) systems. It has been used for stock control and forecasting purposes and thus applies to our research. Demand variance is calculated by taking the square root of the standard deviation divided by the average demand. Demand interval is the average time between demand orders.

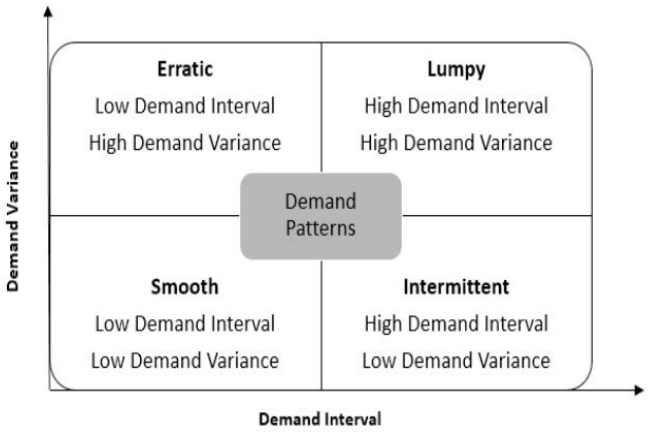

Figure 2: demand classifications for sporadic demand (Williams, 1984).

The average variance and interval were calculated after calculating the values for each base item. The average values (across all base items) were used as the limit values by which to decide if a base item had a 'high' or 'low' variance or interval, which was also done by (Adur Kannan et al., 2020). The average variance value was 7.7, and the average interval was 2.5 weeks. Table 1 shows the classification results. An experimental design should consider the distribution of these classes or evaluate forecasting per class.

Table 1: Classification of SKU demand.

	Items with lumpy demand	Items with intermittent demand	Items with smooth demand	Items with erratic demand
No.	35	12	84	8

2.3 Forecasting

Within this section, several topics related to the current state of forecasting will be discussed, and RQ1.6 *(How are forecasts currently produced and used?)* and RQ1.7 (*What are the requirements for a forecasting system?)* will be answered.

2.3.1 Process of Forecasting

Forecasting is currently a judgemental forecast from the purchasing department. The time horizon of these forecasts is three weeks, as the maximum replenishment lead time is two weeks. Judgemental forecasts are only reactively generated when inventory falls to a (predetermined) order level. This has apparent downsides, such as forecasts being only generated after demand has already exceeded expectations. Potentially ordering too late and ordering larger quantities to compensate for the lack of forecast-based ordering. Employees base their forecasts on a range of factors. One of these is historical demand (demand of the last three months). To quantify this 'forecast', we consider the average of these months and the expected demand. However, this is a very generous quantification, as forecasts are subject to judgemental errors and are not used consequently in this manner. This forecast is a benchmark for this research and cannot be considered the exact representation of the current state of art. The time bucket of the current estimates could be regarded as one month, but since Vivochem strives to supply customers within one week, it is better to use the time bucket of a single week. From a delivery perspective, it is not necessary to have an even smaller time bucket, as lead times are not measured in days.

Other factors, as described in *2.1.2 Other Data Sources a*re generalised and visualised in Figure 3. The importance of these factors differs per SKU. For example, the demand for one SKU might be influenced by gas prices, while another might be affected by weather conditions. Other factors are production costs and chemical availability. It is unknown how much these influence demand. Incorporating these data sources with forecasting might be valuable.

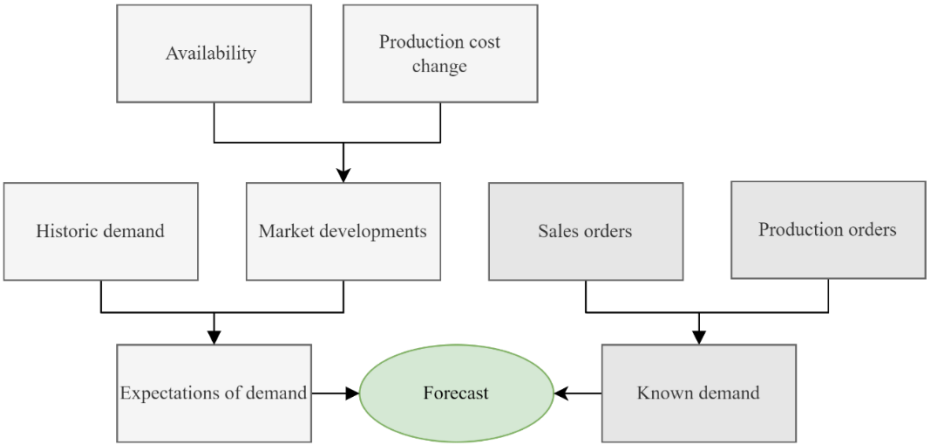

Figure 3: Forecast and its decision variables.

It is assumed that known demand has already been ordered. The resulting forecast is quantified into the purchasing department's reorder levels and order quantities. No calculations are made for these parameters; they are set to what the corresponding employee expects to be suitable.

A new forecasting system should satisfy the following conditions: firstly, it must be able to forecast over three weeks as this is the lead time plus a single week. Secondly, the time bucket is to be set to one single week. Thirdly, the new forecasting system needs to be optimised not solely for accuracy; inventory and operational capacity should be measured to balance the internal operations. Lastly, the forecasting system should show improved accuracy and reduced inventory costs compared to the benchmark while maintaining the service rate.

2.3.2 Current Forecasting Performance

It is necessary to give a quantitative performance indication of the current forecasting performance. As stated, the historical demand is based on the market for orders in the last three months. If we take the average of these three months and consider it the forecast, it is possible to estimate the performance. To evaluate the overall performance of the forecast, the *Symmetric Mean Absolute*

Percentage Error (sMAPE) and bias can be calculated with equations 1 and 2, respectively. The period is indicated with t in time units of one week. The demand occasionally is very low (near zero demand) which can result in extreme outliers due to measuring demand in weight. Therefore, we changed all demand points of less than 100 to zero, which constituted less than 1% of the data points. This change allows us to get a better understanding of performance.

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n} 200 * \left|\frac{A_t - F_t}{A_t + F_t}\right| \quad (1)$$

$$BIAS(\%) = \frac{1}{n}\sum_{t=1}^{n} \frac{F_t - A_t}{A_t} \quad (2)$$

The calculated sMAPE and bias values can be seen in Figures 4 and 5, respectively. Bias values seem are all positive indicating a risk of overstocking when this forecast would not be adjusted. The sMAPE values are large, which indicates the forecasts are inaccurate. Purchasing does not solely rely on this estimation to account for the upcoming demand, but the sheer size of sMAPE values is reason enough to seek improvement. This indicates the significant dependency on purchase and sales employees to readjust this inaccurate forecast.

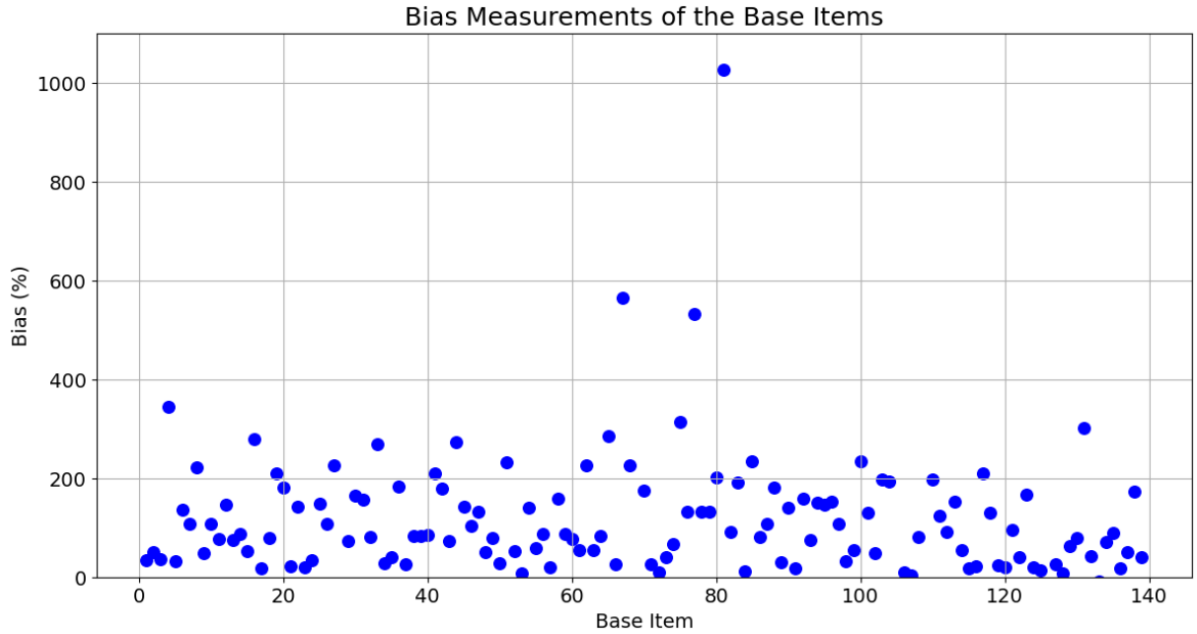

Figure 4: Bias percentage value of all base items by using the three month average.

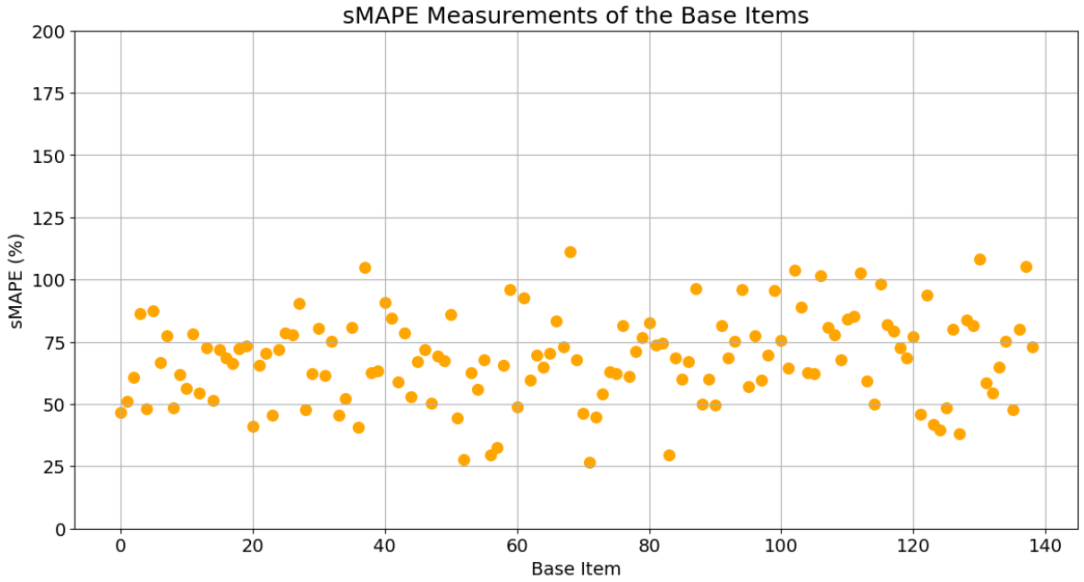

Figure 5: sMAPE percentage value of all base items by using the three month average.

The benchmark (three-month past average) is easily improved by removing the impact of the most significant demand fluctuations. A new forecast was made, which excludes historical peaks (demand four times larger than the past three months average) and troughs (demand ten times smaller than the past three-month average). These values were determined via trial and error. This is not intended as a systematic solution but illustrates the possible improvement by accounting for the pattern. Below, in Figure 6, we show the benchmark and the new forecast (with a one-week time bucket) for a random example. The adjusted forecasts are lower than the benchmark after a peak has occurred. These peaks seem to lead to overestimation by the benchmark method. The adjusted forecast in Figure 6 had an reduction of 11 percentile points for the sMAPE and 33 percentile points for bias compared to the benchmark, indicating a significant improvement. Considering the cumulative effect of forecast errors on the cycle inventory, one can hypothesise the reduction in inventory cost following this adjustment. A significant improvement could be made if a new model predicts these fluctuations.

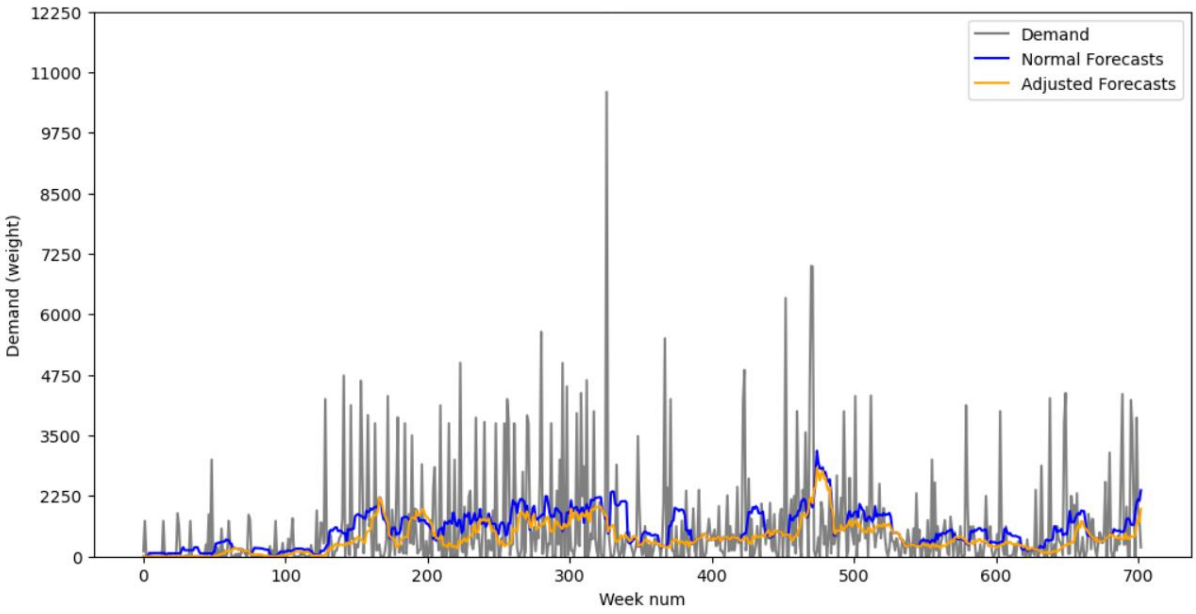

Figure 6: The current (three month average) forecast and the adjusted forecasts compared with the demand.

2.4 Conclusion of the Context Analysis

This section will answer each research question in this chapter, concluding our context analysis.

- RQ1.1 What data are available that can be considered relevant to the expected demand?
  - We found that different data types can represent demand, all related to the orders from Vivochem's customers. The following dates are possible: order date, requested delivery date, shipping date, and delivery date. We identified some potential external data sources to collect to represent economic, production cost, energy costs, and temperature that influence chemical demand.
- RQ1.2 What is the quality of available historical data?
  - From an assessment of the data quality, we decided to use the shipping date to represent the demand in a time series manner. The alternative dates are either not automatically updated or always collected (requested delivery date) or less related to the actual demand occurrence moment (order and delivery date). Therefore, these are less reliable.
- RQ1.3 What is the quantity of available historical data?
  - This data can be sourced from up to three years from the current ERP system and with an additional ten years from the previous ERP system. Adding up to more than 12 years of historical demand data.
- RQ1.4 Which criteria can be used to select SKUs with the highest potential for cost savings?.
  - 29% (constituting 139 base items) of base items were found to relate to 80% of inventory costs. We focus our research on this selection, as the most improvement is to be achieved by introducing new forecasting models for these base items.
- RQ1.5 How can SKUs be classified?
  - We classified base items based on the variance and intermittent periods in their demand patterns. Classification resulted in 35 lumpy base items, 12 intermittent base items, 84 smooth base items, and eight erratic base items.
- RQ1.6 How are forecasts currently produced and used?
  - We have described that forecasts are produced only after a predetermined inventory level is reached. After this trigger, the average demand for the past three months is generated and subject to judgemental interpretation. This forecasting method can be improved by removing significant outliers (peaks and troughs) within the data: a reduction of 11 percentile points for the sMAPE and 33 percentile points for bias. Indicating the potential of incorporating these phenomena with new models.
- RQ1.7 What are the requirements for a forecasting system?
  - We established the requirements for a new system: a time horizon of the lead time plus one week, a time bucket of a single week, and improved accuracy with reduced costs while maintaining the current service rate (compared to the benchmark).

By answering each research question, we have analysed the current state and thereby answered RQ1 (*What is the current state of forecasting and inventory?)*, providing sufficient information to conduct a literature review. The literature should focus on methods incorporating external data sources and demand pattern phenomena. Hereby, the third research phase, *Analysing the Problem*, has ended.

3. Literature Review

This chapter starts the fourth research phase: Formulating Solutions and answering RQ2 (*Which forecasting methods and measurements are available in the literature?*). In section: 3.1 Related Works, we present the relevant literature. In section: 3.2 Concept of Forecasting, we introduce the basics of forecasting. In section: 3.3 Forecasting Models, we discuss appliable forecasting models. In the last section: 3.4 Forecast Performance, different methods for measuring forecasting performance is discussed.

3.1 Related Works

In this section we start by introducing all relevant research. (Axsäter, 2015) and (Silver et al., 2016) are general inventory management books covering forecasting. Both provide an introduction to the basics of forecasting. In (Silver et al., 2016) A framework is proposed to incorporate judgmental human-based and quantitative forecasting. However, comparing results from forecasting competitions: M1 (with only statistical models) (Makridakis & Hibon, 1979) and M2 (with judgemental forecasting) (Makridakis et al.) Did not empirically show any added benefit to judgemental forecasting. In this first competition, statistical models were compared, such as trend-model (Holt, 1957), trend-seasonal model (Winters, 1960), and the ARIMA model (Box & Jenkins, 1970). Later, with the third competition (Makridakis et al., 2000) more variations of these models were compared. In the fourth competition, new types of models, *machine learning* (ML) models, were added; these showed potential (not necessarily performing best). According to (Ni et al., 2020) ML has increased in recent years, and this development can be seen in the fifth competition (Makridakis et al., 2022a) where ML models, for the first time, outperformed the statistical models. The best-performing model is discussed in (In & Jung, 2022).

Many error measurements exist for forecasting. In (Shcherbakov et al., 2013) a summary of the forecasting methods discusses the benefits and disadvantages of using a measurement. However, (Koutsandreas et al., 2022) suggest there is little to no difference in rankings for the different performance measures. According to (Goltsos et al., 2022) and (Kourentzes et al., 2020) there is a discrepancy between forecasting and inventory management, (Petropoulos et al., 2019) proposes a measurement to bridge this gap.

Research on forecasting in the chemical industry is limited. (Bundgaard-Nielsen, 1972) discusses the use of (and need for) regression models to forecast the demand or prices of chemicals in the 1970's. Additional (Broeren et al., 2014) Identifies factors related to chemical consumption in an environmental context. (Estrada et al., 2020) researched different performance measures for forecasting in the context of the chemical industry, providing valuable insights.

Research on sporadic demand patterns is more extensive as not only the chemical industry is subject to this type of demand. (Gamberini et al., 2010) researched using traditional models in the context of sporadic demand with and without trend and seasonality. More complex data patterns were better forecasted using the (S)ARIMA model than the Holt-Winters methods. ML methods were compared with traditional methods in the context of sporadic demand by (Adur Kannan et al., 2020). (Nikolopoulos et al., 2016) proposes the addition of a supervised nearest neighbour for forecasting sporadic demand.

3.2 Concept of Forecasting

This section introduces the concept of forecasting. A forecast refers to a prediction made by forecasting. The forecasting activity can be either quantitative or qualitative (often done separately but possibly combined). Quantitative forecasting can only be done when three conditions are met. (Makridakis et al., 1998). First condition: data or information on past demand is available. Second condition: the information is quantifiable. The last condition is the assumption of continuity (past patterns will continue in the future).

Forecasting is necessary because of two phenomena: uncertainties and economies of scale (Axsäter, 2015). Uncertainty in demand (from customers) and lead time (from suppliers) results in the

need for safety stocks to maintain the ability to serve customers. Due to the economies of scale, there is a financial incentive (or even necessity) to order in batches.

3.2.1 Forecasting in the Chemical Industry

According to (Bundgaard-Nielsen, 1972) Economic growth is indeed related to chemical consumption, thus confirming the suspicion of the purchasing department (see: *Other Data Sources).* In addition to the additional data sources stated in the section: *2.1.2 Other Data Sources,* The literature suggests that there are more variables related to demand. (Broeren et al., 2014) proposes that cost prices are related to demand.

3.3 Forecasting Models

This section will present the different forecasting techniques and answer RQ2.1 (*Which methods are available in the literature?*). We categorise them into judgemental, time series, regression, and artificial intelligence (AI). We will use the M-competitions as a guide to navigate the vast domain of forecasting literature. They offer insight into the historical developments in forecasting methods, providing a valuable context.

According to (Makridakis et al., 1998) and (Silver et al., 2016) demand patterns can include the following components: level, trend, seasonality, cyclic and irregular. Level refers to the mean of a pattern. A trend is the long-term increase or decrease in demand. Seasonality refers to the fluctuations in demand due to seasonal factors, such as quarter of the year, month, or day of the week. Cyclical patterns are comparable to seasonality but do not have a regular periodic occurrence or length. These patterns are all visualised in Figure 7. Forecasting models try to mathematically model these components to improve their forecast accuracy. For our research, we must find models capable of incorporating a irregularity.

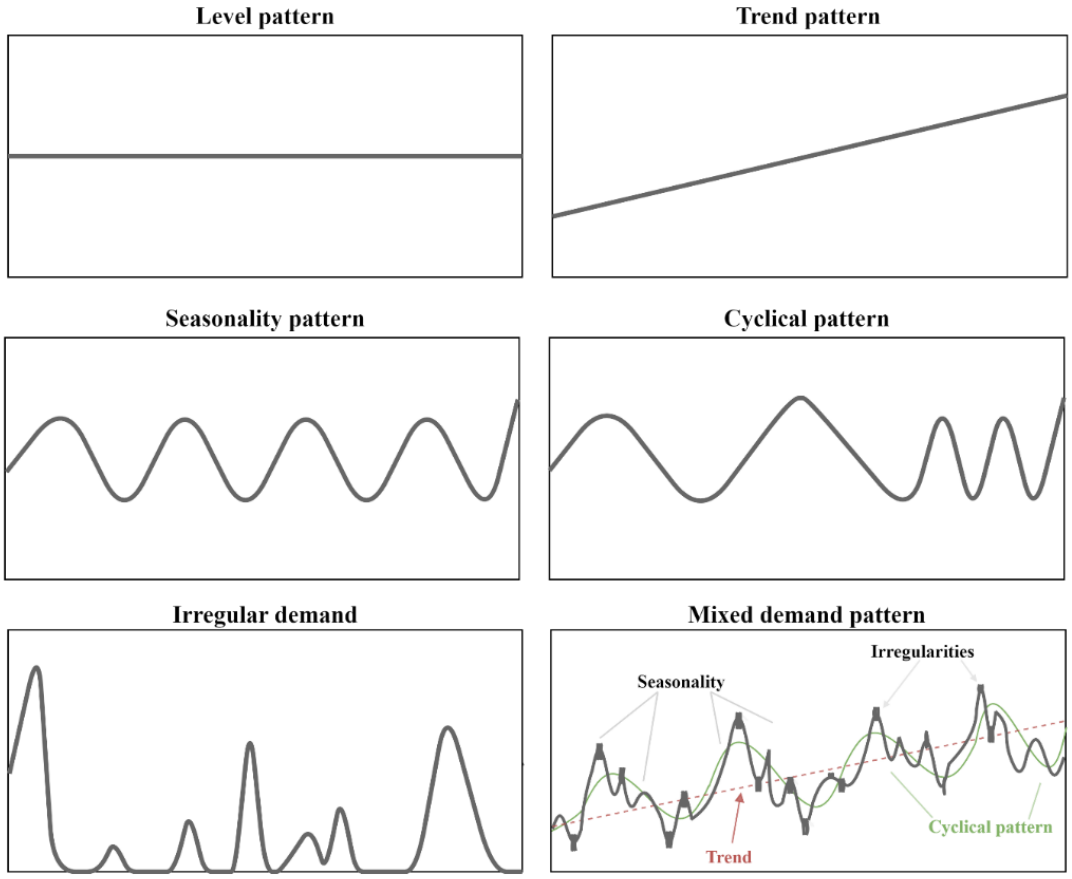

Figure 7: Six demand patterns visualised as a time series.

3.3.1 Judgemental Forecasting

We will discuss the first category of forecasts based on human forecasters' intuition and expertise. This is the earliest form of forecasting and is still prevalent throughout many businesses. The complexity of mathematical models is one reason many businesses rely on human forecasting. According to (Silver et al., 2016) there are advantages to human reasoning: many factors cannot easily be incorporated into mathematical models. Examples of these factors are the impact of promotions, competitor reactions, macroeconomic conditions, and innovations. As stated in the section: *2.1.2 Other Data Sources,* these are also prevalent in the chemical distribution industry (Bundgaard-Nielsen, 1972), thus, human forecasting could be beneficial. (Silver et al., 2016) proposed a framework (Figure 8) on combining human input with mathematical modelling.

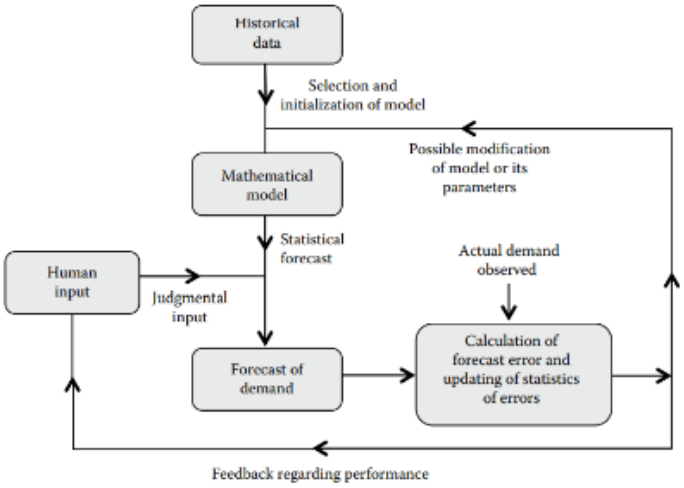

Figure 8: 'Framework: incorporating human input with statistical forecasting' (Silver et al., 2016).

Contrary to the concept behind the framework of (Silver et al., 2016), results from (Makridakis et al., 1993) showed no indication that human-based forecasting improved forecasts. It achieved this by comparing results from competition M2 (Makridakis et al., 1998) where judgemental input was possible with competition M1 (Makridakis & Hibon, 1979) where only pure statistical approaches were used. In (Fildes et al., 2023) manual adjustments to forecasts were researched. It was determined that these adjustments are often not beneficial. Positive adjustments showed a greater than 50% chance to worsen a forecast, while negative forecasts showed a greater than 50% chance to better the forecast. It should be noted that it could not be determined if the negative adjustments were significantly better than a random adjustment.

3.3.2 Time Series Models

The second category, time series models, is a subset of the statistical models. A time series refers to a historical data set of a sequence of observations over time. Time series models use this data set to predict upcoming demand. The primary statistical models are the constant model, trend model and trend-seasonal model (Axsäter, 2015).

The constant model assumes demand is stable over time with deviations due to random independent variables (Silver et al., 2016). The trend model of (Holt, 1957) assumes demand systematically changes over time in addition to deviations due to random independent variables. A trend variable is added to the constant model for the trend model, which is the systematic increase or decrease in a period. The trend model of (Holt, 1957) was expanded into the trend-seasonal by (Winters, 1960) by incorporating seasonality.

All previous models are based on a deterministic theory. Stochastic demand has been modelled in the ARIMA (short for Autoregressive Integrated Moving Average) model popularised by (Box & Jenkins, 1970). Many variations of the model exist; it is described with ARIMA (p, d, q) with the notation: p = order of autoregressive part (AR), d = degree of first differencing involved (I), q = order of the moving average

part (MA). The parameters are often determined via a measurement that balances the model's fitness and complexity. Many criteria exist, but Akaike Information Criteria (AIC) (Akaike, 2011) and Bayesian Information Criteria (BIC) are commonly used methods (Mondal et al., 2014). AIC can be considered the better alternative according to (Yang, 2005) as it does not punish more complex models as severely as the BIC.

The ARIMA $(p, 0, q)$ model can be described by equation 3. the variables are: a = demand level, b_t= coefficient of past demand, ε_{t-1} = error of past forecast and c_t= coefficient of the past forecast errors.

$$x_t = a + b_1 x_{t-1} + b_2 x_{t-2} + \cdots + b_p x_{t-p} + \varepsilon + c_1 * \varepsilon_{t-1} + c_2 * \varepsilon_{t-2} + \cdots + c_q * \varepsilon_{t-q} \quad (3)$$

The variable x_t, represents demand at t. For $d = 1$ it changes into $x'_t = x_t - x_{t-1}$ and for $d = 2$ it changes into $x''_t = x'_t - x'_{t-1}$, repeating this pattern for higher levels of d. Therefore d represents the degree to which past demand is differed. This process, called differencing, enables the model to remove trends and seasonality from the data. Variations of this model exist, such as SARIMA (with added seasonality) or ARIMA with regression.

The various M-competitions have extensively evaluated statistical methods such as time series methods. Results from the first three competitions (Makridakis et al., 1993; Makridakis & Hibon, 1979; Makridakis et al., 2000) Showed that simpler statistical models generally outperform complex models (especially when more noise is involved). When the fourth competition introduced even larger data sets, where complex models had more promising results. According to (Adur Kannan et al., 2020) Simple forecasting performs better in sporadic demand when demand shows no trend or seasonality. The ARIMA model performed better when demand did contain these patterns.

3.3.3 Regression Models

The third category is regression models, another subset of statistical models. This model bases their forecast not only on its historic demand but also on other variables. These variables can be external factors (unrelated to direct demand) or be related to the demand patterns other SKUs. This model type is less used than the other models for demand forecasting.

The general linear regression model can be described by equation 4 (Axsäter, 2015). The variable $z_{t,j}$ represents the regression factor per time unit, b_j the effect of said factor, a the level, and ε_t the error. Adding more than one factor is possible with an identifier j.

$$x_t = \mathrm{a} + \sum_{j=1}^{J} b_j z_{j,t} + \varepsilon_t \quad (4)$$

During the M5 competition, results (Makridakis et al., 2022a) showed the added value of incorporating variables into the forecast. However, this was not done with pure regression but with ML models. (Bundgaard-Nielsen, 1972) proposed using regression to forecast the demand for basic chemicals.

3.3.4 Artificial Intelligence Models

The last category is *artificial intelligence* (AI). An upcoming technology with more research publications each year (Ni et al., 2020). Starting with a definition: *"Artificial intelligence (AI) is the ability of a computer or a robot controlled by a computer to do tasks that are usually done by humans because they require human intelligence and discernment."* (Copeland, 2023).

It is, however, essential to note that there is no consensus among scientists on the definition (Collins et al., 2021). The domain of AI consists of many technologies, but only machine learning (ML) is used for forecasting. Many algorithms can be found under the umbrella of ML, including supervised learning, supervised learning, reinforcement learning, and deep learning. The advantage of these algorithms is that they can solve a problem by 'learning' how to solve a problem, thus mimicking/approaching human

intelligence (Choi et al., 2020; Maddula, 2021). Traditional non-learning algorithms may successfully solve a problem but only do so because they have been explicitly programmed for this purpose.

ML algorithms were first compared to traditional methods in the third competition. Results from the M3 competition (Makridakis et al., 2000) showed the relatively poor performance of a ML model. The M4 competition compared more models (both ML and statistical), and results from (Makridakis et al., 2020) showed that a hybrid (both ML and statistical) model was performing best; pure ML was still performing relatively worse. This all changed with the results of the M5 competition (Makridakis et al., 2022a) where ML algorithms were the top-performing methods, which might be due to the enormous data set involved (Makridakis et al., 2022b). A model called LightGBM (short for *light gradient-boosting machine*) (In & Jung, 2022) performed best. LightGBM is a model which sequentially constructs several regression trees (Barros et al., 2021). The model starts with a single tree trained on the data set; equation 5 represents the model at the starting stage. Where y is the prediction based on the first tree $F_1(\mathrm{x})$. Based on the loss function (for our context, this would either be MSE or MAE) from the first tree, it adds a second tree, which attempts to predict this error. Equation 6 shows how adding a second model is done, where $f_m(\mathrm{x})$ is the subsequent tree added to the previous model to construct the new model $F_m(\mathrm{x})$. The gradient is the collection of all residual from the previous tree, the new tree ($f_m(\mathrm{x})$) is subsequently fitted on these values. To summarise, the first tree (attempts to) predict 'y' based on 'x', and all subsequent trees (attempt to) minimise the error of the previous tree.

$$\mathrm{y} = F_1(\mathrm{x}) \tag{5}$$

$$F_m(\mathrm{x}) = F_{m-1}(\mathrm{x}) + f_m(\mathrm{x}) \tag{6}$$

These trees are iteratively added to improve the overall predictive power of the model until a specific condition is met. Examples of these stopping conditions are the number of trees or the number of iterations without improvement. After this condition is met, the model's training ends. In contrast to other more traditional gradient-boosting methods, lightGBM uses leaf-wise steps instead of level-wise ones. Level-wise growth expands all nodes (trees) when levels are expanded, while leaf-wise growth only expands the node with the steepest descent (in other words, with the best improvement). This principle is visualised in Figure 9. Level-wise growth allows for deeper trees within a shorter processing time. The large number of trees, which individually are very simple, collectively form a complex model. This complex model is able to capture complex non-linear relations in data. After training, the eventual forecast is generated by taking a weighted average from all trees. The weight assigned to each tree is decided by its learning rate, resembling the degree to which the new trees are part of the weighted average.

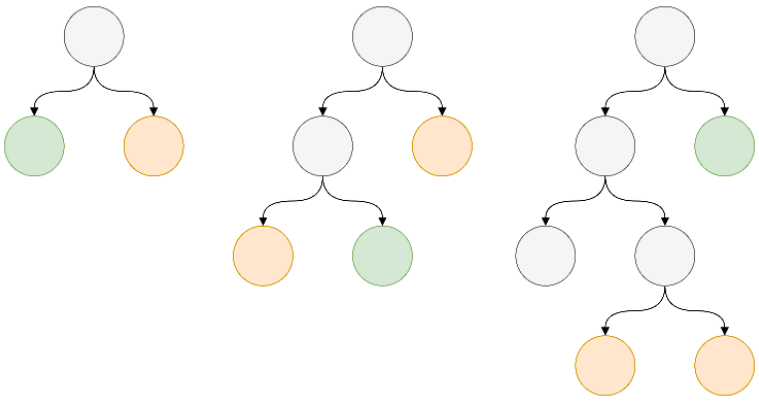

Figure 9: Visualisation of the leaf-wise principle.

The top five models from (Makridakis et al., 2022a) all used LightGBM, except the third best using a neural network (NN). A neural network is a ML learning algorithm mimicking the mechanisms of neurons in the human brain (Allende et al., 2002). It consists of several layers containing neurons; Figure 10 shows a network with three inputs and one layer with four neurons. Neurons in each layer are connected to neurons in adjacent layers, one side being the input and the other being the output. Each input and output has a corresponding weight assigned to it. By assigning training data to such a model, inputs and outputs can be trained based on a known output. These weights are optimised based on a loss function. This enables the model to 'learn' to produce forecasts.

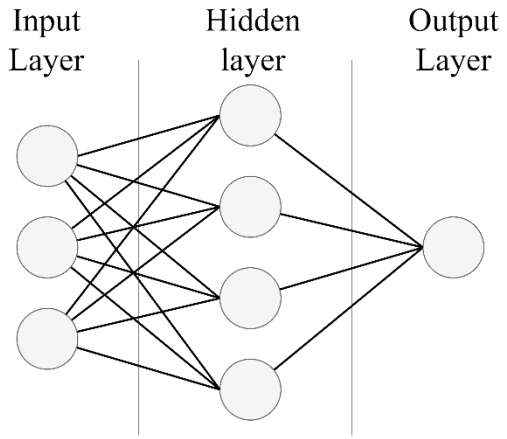

Figure 10: Visualisation of a neural network.

The performance of a neural network is tied to its topology (number of layers and neurons); a NN with more layers does not necessarily perform better (Kavzoglu, 1999; Tran et al., 2020; Wang et al., 1994). While many methods exist, there is no universal method to determine the topology (Ibnu et al., 2019). Findings from (Shen et al., 2021; Wang et al., 1992) showed that three-layered networks can model all non-linear relations; however, one must consider the specific NN types involved. (Uzair & Jamil, 2020) compared many neural networks from different research sources. Most networks used one to three layers; more layers did not improve accuracy but did increase computational time.

(Adur Kannan et al., 2020) showed that ML outperformed ARIMA models while simple (benchmarking) methods were still performing best. However, when data was aggregated, and exogenous data was added, forecasting combinations performed better. Once again, this suggests that combination methods are performing best. It should be noted that this research focused on supply-side forecasting but with similar sporadic data. (Nikolopoulos et al., 2016) found that adding the nearest neighbour algorithm to a forecasting model improves forecasting accuracy. The nearest neighbour is argued to perform best for industrial data where there is a limited quantity of data. An improved version: 'k-nearest neighbours' (KNN), was proposed by (Hasan et al., 2024), which can incorporate zero values. The KNN algorithm for regression predicts values based on the 'k' nearest neighbours. The distance to these 'neighbours' is based on their corresponding variables. In equation 7, we show how this distance (in an Euclidean manner) is calculated between two points, point ***a*** and point ***b***, for several variables/features related to these points. This distance is then used to determine the closest 'k' neighbours. Data should be normalised beforehand to counter differences in scale.

$$d_{a,b} = \sqrt{\left(x_{a,1} - x_{b,1}\right)^2 + \left(x_{a,2} - x_{b,2}\right)^2 + \cdots + \left(x_{a,n} - x_{b,n}\right)^2} \tag{7}$$

After determining the 'k' nearest neighbours, it computes the weighted average based on the relative distance. In the context of forecasting, this weighted average constitutes the forecasts generated for this point based on the corresponding variables.

3.4 Forecast Performance

This section will discuss the various measures used to evaluate forecasting performance and their advantages/disadvantages. We will also determine a specific measure for the effect of forecasting performance on inventory. Thus, in this section, we will answer RQ2.2 (*How can the forecasting accuracy be measured?*) and RQ2.3 (*How can the inventory performance be measured?).*

3.4.1 Forecasting Measures

There are many performance measures available for forecasting. Findings from (Koutsandreas et al., 2022) suggest that the chosen error measure has little to no impact on the ranking of forecasting methods. Considering practical limitations, we will make a selection of measures. For selection purposes, findings from (Koutsandreas et al., 2022; Shcherbakov et al., 2013) will be used.

Forecasting measures for the chemical industry have been researched by (Estrada et al., 2020). Different statistical methods were researched. None of these showed significant differences in subsequent rankings of methods. Scale-dependent measures calculate error from the difference between the observed and forecasted values. In this category, we choose the *mean squared error* (MSE) and *mean absolute error* (MAE); both are easy to interpret compared to other measures. MSE is a commonly used measure but is sensitive to outliers, which is why we use MAE as well. We will also use a symmetric error measure to combat possible non-symmetry in outliers*:* sMAPE. Equations of MSE, MAE and sMAPE are 8, 9 and 1, respectively. We include bias to evaluate the over/under forecasting (equation 2).

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2 \tag{8}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{9}$$

3.4.2 Inventory Measures

According to (Goltsos et al., 2022), there is a discrepancy between forecasting and inventory management. In research, forecasting is often assessed with error measures, but none of these are representative (at least not directly) of the effect on inventory (Kourentzes et al., 2020). To solve this problem (Petropoulos et al., 2019) proposes a measurement to assess forecasts not for accuracy but for inventory management referred to as *root mean square* (RMS). The root mean square is taken from variables representing the financial, operational, and service aspect of inventory management. It can be seen in equation 10, where: C_h are the inventory costs (cost price per day of storage) in euro's, v_0 is the variance of order quantity (to the supplier), and α is the service level (fill rate). All are divided by the mean aggregated between forecasting methods. Exception is the service level which is to be maximised, therefore, this part is inversed.

$$RMS = \sqrt{\frac{1}{3}\left(\frac{C_h}{C_h{}^{mean}}\right)^2 + \frac{1}{3}\left(\frac{v_0}{v_0{}^{mean}}\right)^2 + \frac{1}{3}\left(\frac{\alpha^{mean}}{\alpha}\right)^2} \tag{10}$$

This equation allows for assessing the effect of forecasting on inventory by evaluating the relative effect models have on these variables. It will be necessary to adjust the variables to have the measurement fit our research context. It is essential to choose an inventory policy for simulation reasons. Currently, no systematic approach is used; the *continuous Order Op To* (OUT) system was used in (Petropoulos et al., 2019).

3.5 Conclusion of Literature Review

In this section, we answer each research question in this chapter, concluding our literature review.

- RQ2.1 Which methods are available in the literature?
  - We researched relevant literature based on the requirements and context established in chapters one and two. in the section: *3.3 Forecasting Models.* ML methods, such as NN, KNN and LightGBM, have shown a promising ability to forecast sporadic/complex demand patterns. These methods can incorporate exogenous data. Regression models were found to be relevant for the chemical industry. However, the ability of ML to incorporate exogenous data makes them arguably irrelevant. Traditional methods have, up until recent years, performed better than ML methods, depending on the complexity of the demand patterns. More complex time series methods, such as ARIMA, perform better with more complicated patterns. In contrast, methods such as Holt perform better for less complex patterns.
- RQ2.2 How can the forecasting accuracy be measured?
  - Forecasting measurements were discussed in the section: *3.4 Forecast Performance.* It was determined that different accuracy measurements do not result in (significantly) different rankings. We consider the following accuracy measurements: sMAPE, MSE, MAE and bias to offer a diverse selection of types of error measures.
- RQ2.3 How can the inventory performance be measured?
  - To measure the impact of forecasting on inventory, we found the RMS method from (Petropoulos et al., 2019) to be suitable.

We have answered RQ2 (*Which forecasting methods and measurements are available in the literature?*). With this conclusion, the fourth research phase, *Formulating Solutions*, ends. The next chapter should define how these methods can be translated into appropriate models.

4. Methodology

This chapter starts the fifth research phase: *Choosing a Solution*. In this chapter, we answer RQ3 *(How can the performance of the forecasting methods be evaluated?).* Within *4.1 Data Collection*, we discuss our data collection and preparation choices. We elaborate upon the input variables that are derived from these data sets. In section: *4.2 Model Parameters,* we discuss which parameters are to be set and the method to determining the optimal setting. Within the *4.3 Performance Assessment* section, we explain how we will assess the performances of the various models. Within the *4.4 Validation & Verification* section, we explain how we ensure that our models are verified and their results can be validated.

4.1 Data Collection

This section explains the data sourcing and cleaning procedures, answering RQ3.1 (*How is data prepared and sourced?*). Some parts of the data collection have already been discussed in the section: *2.2 Demand Patterns*, these parts will be repeated shortly. We will discuss new data collection in more detail. We distinguish between endogenous (directly related to demand) and exogenous (indirectly related to demand) data and elaborate upon them separately.

4.1.1 Endogenous data

ERP data represents the historical demand data (measured in kilograms), specifically the shipment date of products to customers. The data was sourced from Navision (ERP system) via Power-BI. For a full elaboration on the data sourcing and preparation, we refer to the section: *2.1.1 Historic Data.* From the ERP data, multiple input variables can be derived that could improve ML models, which we can refer to as feature engineering. The statistical models will not be able to make use of these features. We will explain all input variables and elaborate upon their hypothetical/expected value. The first one is *Total demand* (measured in KG), which is the observations of total demand across all base items; the overall trend in chemical consumption might be captured better across all base items than within a single base item.

The second set of input variables, *Moving average forecast* and *Moving average error* belong to the moving average metric. These are the lagged forecasts and corresponding errors from the benchmark method. Thus, an ML model only gets historic forecasts that would have been known at any specific moment. This could allow ML models to 'learn' from the benchmark and its errors, thus improving the current state of the art.

The next set of input variables belongs to the metric of demand patterns. These are added to indicate certain demand pattern phenomena explicitly. The first input variable *Peak*, indicates that an observation of demand in a week was five times larger than the average demand (up until then). This could allow models to predict or account for extreme outliers. The input variable: *Dip* indicates if demand in a week is a dip, defined as when observed demand is less than a quarter of the (up until then) average demand. These two phenomena have shown the potential to improve forecasts in the section: *2.3.2 Current Forecasting Performance*. We have added the input variable for weeks with zero demand: *Zero*. These three are all binary variables, indicating whether a corresponding phenomenon is observed. These models could help a ML model learn to predict these demand phenomena by capturing underlying relationships.

The last metric is time, for which we added the month and quarter numbers as input variables. This could assist an ML model in recognising seasonality in these different time buckets. We summarise all the input mentioned above variables in Table 2. Input variables from the metric demand, moving average, and Demand pattern are lagged by three weeks (the lead time plus one week). This is necessary as these values are not known three weeks in advance.

Table 2: Endogenous variables.

Metric	Variables	(Sourced) time bucket	Period	Data source
Demand	Total demand	Weekly (lagged by three weeks)	2012-2024	ERP system
Moving average	Moving average forecast Moving average error	Weekly (lagged by three weeks) Weekly (lagged by three weeks)	2012-2024	Calculated from Demand metric
Demand pattern	Peaks Dip Zero	Weekly (lagged by three weeks) Weekly (lagged by three weeks) Weekly (lagged by three weeks)	2012-2024 2012-2024 2012-2024	Derived from Demand metric
Time	Monthly Quarter	Monthly Quarter	2012-2024 2012-2024	Power Bi training set (Singh, 2024)

4.1.2 Exogenous data

We collected exogenous data from various sources summarised in Table 3. We searched for publicly available data based on the hypothesised influential variables. The hypothesised relation between these metrics (or variables) and demand has been elaborated upon in the section: *2.1.2 Other Data Sources*. However, we will explain how the data sets represent the variables.

We include three input variables for the metric temperature: *Daily average temperature, Daily Minimum temperature,* and *Daily maximum temperature* from a single measurement station in the Netherlands. We included the minimum and maximum temperatures to incorporate the full spectrum of temperatures. For example, the temperature might fall below freezing point during the night (minimum temperature), but the average temperature might not capture this. This could have impacted the demand for chemicals to combat these conditions (example: usage of salts during the winter). We assume that using only one measurement station is sufficient to capture the relation between demand and temperature, as the majority of customers are located in the Netherlands or border regions of neighbouring countries.

For the second metric, economic development, we include various input variables such as the *GDP* of the *EU, China,* and the *USA*. We attempted to collect these in as tiny time buckets as possible; however, this data was not measured in larger time buckets. Despite the more significant time buckets, we assume it would capture (a degree) macroeconomic influence on chemical demand. We also included the input variables: *Production price index* and *Consumer price index* of the Netherlands. We attempt to model the (potential) relation between costs and demand with these variables.

For the last metric, energy cost, we include the input variables, such as *Crude oil prices* (Europa) and *EU gas prices*. We intend to model the (potential) relationship between demand and energy costs with these two input variables. Including oil and gas cost prices will cover a large part of industrial energy consumption, thus enabling the potential relationship with demand.

These data sets have been sourced from official (or related to) governmental sources. There is no uniformity in quality and quantity since data had to be sourced from different sources, as no single data bank held all data sets. The exogenous data sets were combined and aggregated into equal time buckets with the endogenous data. For smaller time buckets: we aggregated the daily data points of into weekly time buckets. For the larger time buckets: all weeks belonging to the corresponding larger time bucket contain the value of the larger time bucket. For example: *EU gas prices* belong to a monthly time bucket, but all four weeks belong to this time bucket will contain the value of the whole month. This data is then lagged such that only data points for previous month are available. For the data sets that have yet to be updated up to 2024, we need to lag the data. This is done by shifting the data to the latest data for the last observation point. The variables are summarised in Table 3.

Table 3: Exogenous variables and data sources.

Metric	Variables	Time bucket	Time period	Data source
Temperature (de Bilt, Netherlands)	Average temperature	Daily	2012-2024	*(KNMI, 2024)*
	Minimum temperature	Daily	2012-2024	
	Maximum temperature	Daily	2012-2024	
Economic development	GDP EU	Quarterly	2012-2023	*(Eurostat, 2024)*
	GDP China	Yearly	2012-2022	*(FRED, 2024)*
	GDP USA	Quarterly	2012-2023	*(BEA, 2024)*
	Production price index	Monthly	2012-2023	*(CBS, 2024b)*
	Consumer price index	Monthly	2012-2024	*(CBS, 2024a)*
Energy cost	Crude Oil Prices (Europa)	Daily	2012-2024	*(EIA, 2024)*
	EU Gas prices	Monthly	2012-2024	*(IMF, 2024)*

Data needs to be split into training and testing data. Training data will be used to train models, and testing data will be used to validate and estimate the performance of models against unknown but similar data. This is a widely used and accepted method of evaluating forecasting performances. A typical split is 80% training data and 20% test data (Joseph, 2022), which we also use. We split the data on a weekly basis, such that the first 80% of weeks (week 1 to 504) are used to train models and tested on the last 20% of weeks (week 504 to 630).

4.2 Model Parameters

This section will answer RQ3.2 (*Which methods are to be developed into models?*) and RQ3.3 (*Which parameters need to be set for the models?*). First, we will define which methodologies we intend to apply. In the literature research in the section: *3.3 Forecasting Models.* We found several applicable forecasting methods that have shown promising capabilities. Within the category of statistical models, we identify the Holt and ARIMA methods. These methods are widely used and have, up until recent years, outperformed more complex methods. They show no indication of being specifically capable of modelling our demand patterns but are included for their wide usage and to offer an additional benchmark to the other models. The selected ML models are the neural network, k-nearest neighbour and LightGBM. These ML models have started outperforming statistical models when given large data sets and exogenous data for complex demand patterns.

For the Holt model, we need to set the smoothing parameters. For the ARIMA model, we need to set the parameters: order of autoregressive (p), degree of first differencing (d), and order of the moving parameters (q). We need to determine the NN model's topology (layers and layer sizes) and the algorithm which determines the weights between neuron connections. For the LightGBM model, we need to determine the learning rate, max bin size, max number of leaves, boosting algorithm and the number of boosting rounds. We base our parameter search on literature researching parameter settings for LightGBM (Barros et al., 2021). For the KNN model, we need to determine the number of nearest neighbours to consider. In addition to these models, we use the approximation of the current state of forecasting as a benchmark, as described in the section: *2.3.2 Performance.* We will refer to this model as a benchmark. In the next chapter, we will configure these models and elaborate more extensively on their workings. These parameters will be configured in chapter 5. For configuration, if any parameter can be easily set per individual base item, this is the preferred option. However, this may not be achievable for all parameters. If needed, we generalise the parameter across all base items.

4.3 Performance Assessment

In this section, we explain how we assess our forecasting models, thereby answering RQ3.4 (*How can we assess model performance?*). Performance measures have been extensively discussed within the literature review section: *3.4 Forecast Performance.* We briefly state why the measurements were chosen and explain how they suit our purposes.

4.3.1 Forecasting measures

We use the bias, MSE, MAE, bias, and sMAPE measurements to evaluate the forecasting performance. We use the commonly used MSE as it is very easily interpretable. However, we also

included the MAE measurement due to MSE being sensitivity to outliers (as result of squaring the errors). We also include sMAPE as it weighs negative and positive errors equally while providing a relative error measure. We will include bias to evaluate the tendency of over/under forecasting, but will not use it to determine improvement as it does not indicate the accuracy of a model. As supported by the literature, more measures do not necessarily offer more insight; see literature section: *3.4.1 Forecasting Measures*. Research pointed out that many measures do not affect the relative ranking of forecasting methods. Thus, we feel justified in using these measurements. The error measurements are calculated from the test data and collected for each model and base item. We have classified each base item according to its demand pattern. All error measurements are individually assessed but aggregated into their respective classes. Allowing us to identify possible differences in the fitness of models for different demand patterns. We assess the forecasts based on their error values with forecast horizon of three weeks but will also extend the forecast horizon to investigate if the methods perform differently with different horizons.

4.3.2 Inventory measures

We want to evaluate the performance of these forecasts from an inventory perspective. This is done using the RMS measurement, as shown in equation 10. We slightly changed the formula to meet the specific needs of this research. The changes can be seen in equation 11; first of all, we changed the weights assigned to the variables. We consider the variance of orders to be less important than the service rate and inventory costs; thus, we have decreased its weight. The variables have indices to indicate the corresponding base item (i) and forecasting method (j). The variables: inventory cost ($I_{i,j}$), order (to suppliers), variance ($v_{i,j}$), and service rate ($\alpha_{i,j}$) are the respective variables for the forecasting method (j) for that base item (i). The inventory cost variable ($I_{i,j}$) shown in equation 12, calculated with the cumulative inventory level ($I_{n,i,j}$) where 'n' represents the week number. Calculating the inventory cost would only be possible if we discriminate in packaging, allowing us to identify the space taken up by products. The significant complexity would not result in much value. Simplicity in the model is preferred, and we could consider less inventory generally equal to less inventory costs. Additionally, since the values are normalised, we feel justified in this change. The order variance variable ($v_{i,j}$) is shown in equation 13, calculated with the order quantity ($q_{n,i,j}$) at week 'n' and the mean order quantity q_{mean}). Equation 14 shows how we calculated the service rate (fill rate) variable ($\alpha_{i,j}$), based on the fulfilled demand ($D^{F}{}_{i,j}$) and total demand ($D_{i,j}^{total}$). The variables from equations 12, 13, and 14 are calculated from the inventory system we work out in the next chapter.

$$RMS_{i,j} = \sqrt{\frac{2}{5}\left(\frac{I_{i,j}}{I_i{}^{mean}}\right)^2 + \frac{1}{5}\left(\frac{v_{i,j}}{v_i{}^{mean}}\right)^2 + \frac{2}{5}\left(\frac{\alpha_i{}^{mean}}{\alpha_{i,j}}\right)^2} \tag{11}$$

$$I_{i,j} = \sum_{n=1}^{N} I_{n,i,j} \tag{12}$$

$$v_{i,j} = \frac{1}{N}\sum_{n=1}^{N} q_{n,i,j} - q_{mean} \tag{13}$$

$$\alpha_{i,j} = \frac{D^{F}{}_{i,j}}{D_{i,j}^{total}} \tag{14}$$

These variables are normalised with the corresponding means across all forecasting methods, thus allowing us to compare the values. These variables are inventory ($I_i{}^{mean}$), order variance ($v_i{}^{mean}$), and service rate ($\alpha_i{}^{mean}$). A lower RMS value represents a relative improvement across its three variables. Since the RMS values are normalised by including mean values, they offer insight into the relative performance. They allow us to compare the effect of different forecasting methods on inventory. This measurement is valuable to bridging the gap between forecasting and inventory. For example, it could be possible that a more positively biased model would yield a more favourable RMS value (higher service rate as a result of more inventory). At the same time, forecasting measures would indicate

another less biased model having comparable error measure values. It is important to note that the measurements indicate the potential improvement based on the chosen inventory policy, which is not fully representative of the inventory decision-making at Vivochem. However, it still indicates the relative performance of inventory based on different forecasting models.

4.4 Validation & Verification

In this section, we discuss how we verify the models and validate the results produced by our models to ensure reliability and reproducibility. Thereby answering RQ3.5 (*How can we validate and verify the models?).*

4.4.1 Verification

We want to ensure that the models used are verified. In our literature review, we researched forecasting methodologies and considered their applicability to strongly fluctuating demand. The literature research has shown us that these methods are appropriate for forecasting. Thus, the models we intend to develop can also be considered appropriate. We use literature, if possible, for the parameters and describe how we configure them. We consider our models verified based on the literature research indicating their widespread usage as forecasting models.

4.4.2 Validation

We want to ensure that we can validate the results of our models. For this purpose, we researched performance indicators for both forecasting and inventory see section: *3.4 Forecast Performance,* All of these measurements enable us to validate the model's results when comparing them with the benchmark model. When the models perform relatively better according to these measurements compared to the benchmark method, we consider their performance valid.

4.5 Conclusion of the methodology

In this section, we will answer each research question in this chapter. With this, we conclude our methodology.

- RQ3.1 How is data prepared and sourced?
  - All data are aggregated into equal time buckets, providing one large data set. This data set is then split, 80% is used to train the models and 20% to test the models.
- RQ3.2: Which methods are to be developed into models?).
  - We intend to develop and apply the statistical methods Holt and ARIMA, as well as the ML methods NN, KNN, and LightGBM. Other methods have been excluded as they have not shown promising results for our research context during our literature research.
- RQ3.3 Which parameters need to be set for the models?).
  - For the holt model we need to set its smoothing parameters. For the ARIMA model, order of autoregressive (p), degree of first differencing (d), and order of the moving parameters (q). For the NN model the number of layers, neurons, and the algorithm to set weights. For the lightGBM we need to set the need to set the learning rate, max bin size, max number of leaves, boosting algorithm and the number of boosting rounds. For the KNN we need to set the 'k' nearest neighbours number.
- RQ3.4 How can we assess model performance?
  - We decided to use bias, MSE, MAE, and sMAPE to assess forecasting performance. We decide to use an adjusted RMS measurement to assess the inventory performance.
- RQ3.5 How can we validate and verify the models?
  - We consider our models verified as they are based upon methods found in our literature research. Our models can be validated by comparing the results of the models against the benchmark model with the error measurements.

Hereby, we have answered RQ3 *(How can the performance of the forecasting methods be evaluated?).* The next chapter should define the specific parameters of the models.

5. Model Designs

In this chapter, we develop the forecasting models and inventory model. In section: 5.1 Forecasting Models, we explain how we set the parameters for the models. In section: 5.2 Inventory Model, we will explain the inventory system we use. Thereby, answering RQ4 (*How are the models designed?*).

5.1 Forecasting Models

This section will answer RQ4.1 (*How should the forecasting models be configured?)*. We chose Python as the programming language as it is widely used and many available open-source libraries are programmed with it. We program the models as much as possible in accordance with the literature. However, many parameters have less applicable literature and are too time-consuming to optimise manually. To still be able to develop all models, we use auto-determination parameters via specialised methods wherever possible. We consider this the better approach.

5.1.1 Benchmark model

The benchmark method: average of the observed demand in the last 12 weeks. Forecasts are generated in runs of three weeks, meaning we cumulatively generate forecasts for periods of three weeks (since the maximum lead time is two weeks) for the test data set. The model has a forecast horizon of three weeks with a time bucket of one week. This model is fully self-developed. This model only uses the historical demand data as input. It generates forecasts for each base item. The bias of this model on the training data was small, and therefore, we have not adjusted the forecasts on bias.

5.1.2 Holt model

The "*ExponentialSmoothing*" model from the "*statsmodels"* library (Perktold et al., 2024) is used as the Holt model. We automate the determination of its smoothing parameters. This is necessary as determining smoothing constants for all base items is a time-consuming task that is impossible within this research period. We apply the *least squares* method (Simoncelli & Daw, 2003; Zemkoho, 2022). The method searches for the parameters that minimise the squared errors from forecasts generated on the training set. We are justified using this method as squared errors are very much related to all the forecast measures we use, meaning that it is highly likely that the optimal smoothing parameters will be realised. We apply this optimisation method for each base item, giving each item a unique model with parameters fitting to their demand pattern. Giving each base item a unique model fitted to their specific demand pattern. This model only uses the historical demand data as input. The model has a forecast horizon of three weeks with a time bucket of one week. The bias of this model on the training data was small, and therefore, we have not adjusted the forecasts on bias.

5.1.3 ARIMA model

For an ARIMA model, the "*auto_arima_model*" from the "*pmdarima*" library (Smith, 2023) is used. We use an auto-ARIMA model which determines parameters automatically in a stepwise manner with criterium: AIC (Akaike, 2011). This criterion selects for the best-fitted model with the least parameters. This is in accordance with literature stating that complex forecasting methods do not necessarily outperform simpler ones (section: *3.3 Forecasting Models*). Thus, this auto-determination is an acceptable method, considering that the manual determination of optimal parameters is very time-consuming. In literature (section: *3.3 Forecasting Models)*, we found that this is the preferred method out of the two commonly used methods. We apply this parameter configuration method to a model for each base item, giving each base item a model with parameters fitted to its demand pattern. This model only uses historical demand data. It has a forecast horizon of three weeks and a time bucket of one week. We adjusted the forecasts generated by this model with the average bias from the training data.

5.1.4 Neural Networks model

We use the following model: "*MLPRegressor" f*rom the "*sklearn.neural_network*" library (scikit-learn developers, 2023b). As the possible configurations for a NN models are numerous, tuning these parameters for all base items too time consuming. We need to use a general model. To find one, we base our search on the theory that most problems can be solved with either one to three hidden layers, as stated in the section: *3.3.4 Artificial Intelligence* on NN. It is essential to choose a balanced number of hidden layers that causes no overfitting (due to too many hidden layers) while still capturing the

underlying relationships. Thus, we experiment with these hidden layers to find which fits our context best. This was done by evaluating the performance of different layers across all base items. Table 4 shows the results, where a negative percentage indicates a lower error and vice versa.

Table 4: Average percentile difference in sMAPE and bias of hidden layers two and three compared to a single hidden layer of neural network across all base items.

	One hidden layer	Two hidden layers	Three hidden layers
Average MAE	-	1%	2%
Average MSE	-	2%	5%
Average sMAPE	-	1%	1%
Average Bias	-	-1%	-1%

We do not intend to optimise the entire topology of the NN, as this could be a research effort on its own. The two and three hidden layers result in only minor differences in the error values compared to the single layer, except for the three layers having a larger average MSE. Due to the many minor percentage differences, we conclude that performance between using one or two layers is the same. We chose to use the simplest model with a single hidden layer. This is in accordance with literature stating that more layers do not necessarily equal better performance (referring to section: *3.3.4 Artificial Intelligence*). The models contain one neuron for each of the input layers. We use the *Limited memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS) method solver to determine weights for the relations between neurons. This solver is a gradient-based optimiser; its limited memory property makes it worthwhile for ML applications to spare computational power. According to (scikit-learn developers, 2023b) it is the best-suited solver for smaller data sets. Its limited memory property and applicability for smaller data sets make it a justified method for determining the weights. We apply a model with this solver to each base item, giving each base item a unique NN model fitted to their demand pattern. This model can use the complete data set. The model has a forecast horizon of three weeks with a time bucket of one week. We adjusted the forecasts generated by this model based on the average bias from the training data.

5.1.5 K-Nearest Neighbour model

We use the model: "*KNeighborsRegressor*" from the "*sklearn.neighbors*" library (scikit-learn developers, 2023a). The model uses a brute force approach, meaning all neighbours have their respective distances calculated (in an Euclidean manner) and are weighted accordingly. The distances are based on all input data related to a data point. This data is normalised via min-max scaling, negating the effect of different feature value ranges. This means the 'k' nearest neighbours (in other words, most similar data points) are used to generate the forecast. The brute force approach is computationally executable for our data set. Thus, there is no reason not to use it. The number of neighbours considered by the algorithm is left to be determined. For this purpose, we will compare the average bias, MAE, MSE, and sMAPE across all base items for several different k-values. We decided to include all measurements as we observed a stronger inconsistency between these compared to other models. We apply: k=3, k=5, k=10, k=20, k=30, k=100, and k= 200 to cover a large range. The results are shown in Table 5; a negative percentage indicates an improvement for that measurement.

Table 5: Average percentile difference in measurement values of different k values across all base items compared to k = 5.

	k=3	k=5	k=10	k = 20	k= 30	k=100	k=200
Average MAE	-	-3%	-6%	-8%	-9%	-11%	-12%
Average MSE	-	-11%	-15%	-19%	-20%	-23%	-27%
Average sMAPE	-	2%	3%	4%	4%	5%	10%
Average Bias	-	5%	69%	52%	51%	51%	51%

Table 5 shows that sMAPE and bias increase with a higher 'k' while the other measurements decrease. We consider the bias increase (5%) at k=5 negligible against the MSE decrease (-11%) at k=5. Thus, we chose k=5 as the relative improvement in MSE and MAE of more neighbours (at k<10) to be offset the significant increase (more than 50%) in bias at these levels. This model can use the complete data set. We apply a KNN to each base item, generating forecasts based on the nearest neighbours (based on the available input data) of a specific base item's specific data point. The model has a forecast horizon of three weeks with a time bucket of one week. Due to the nature of the KNN model, we cannot generate a forecast on the training data, and thus, we do not adjust the estimates on any bias value.

5.1.6 LightGBM model

We use the opensource LightGBM model from the '*lightgbm'* library (Microsoft, 2023). The amount of parameters forces us to set the parameters based on the literature (see section: *4.2 Model*) to ensure the research stays within the given timeframe. We use the *'Dropouts meet Multiple Additive Regression Trees'* (DART) boosting algorithm. The boosting method prevents overfitting as it 'drops' a random number of trees (explained in section: *3.3.4 Artificial Intelligence)* after each iteration and is recommended by (Barros et al., 2021). We do not limit the model's depth. The other parameters are the number of leaves, learning rate, number of rounds, and max number of bins. The number of leaves determines the complexity of the model. It represents the number of end nodes each tree has and, thus, how often decisions based on the data are made before generating a prediction. Too many leaves can result in overfitting and vice versa. The learning rate represents the degree to which a newly generated tree is used in the overall model, with higher values 'weighing' new trees 'more'. The number of rounds refers to the number of new trees generated to improve the model. The number of bins represents the number of bins in which input data is placed. For the parameters: number of leaves, learning rate, number of rounds, and max number of bins, we apply a grid search. Parameters settings are equal, or one or two levels higher and lower than those recommended by (Barros et al., 2021) which can be viewed in Table 6. Except we do experiment with higher numbers of boosting rounds as the training of these models takes significant time due at greater number of rounds. Experimenting with all base items for all possible settings is too time-consuming (taking almost one hour per setting) for this model; therefore, we measure the average performance values across 14 randomly selected base items (constituting 10% of the total) instead of all base items.

Table 6: grid search settings for LightGBM.

Number of leaves	50	40	30	20	10
Learning rate	0,01	0,05	0,1	0,15	0,2
Number of rounds	100	200	500	1000	2000
Max number of bins	200	300	400	500	600

These settings resulted in 625 experiments, we chose the following settings: 50 number of leaves, 0,2 learning rate, 200 max number of bins, and 500 boosting rounds. This setting had a decrease of 15% in the average bias with only a 1% increase in average MAE and average MSE and smaller than 1% difference in average sMAPE. All other settings either do not improve significantly or decrease other error measures while increasing other error measures. Therefore, using this setting, which significantly improves bias while having minimal impact on the other error measures.

We train a LightGBM model for each base item, giving each base item a unique LightGBM model fitted to its demand pattern. This model can use the full data set. The model has a forecast horizon of three weeks with a time bucket of one week. We adjusted the forecast from this model based on the median bias from the training data; we found this adjustment to be the better option.

5.1.7 Data Selection

The ML methods can be trained on the whole training set, including all endogenous and exogenous variables. However, we only want to use the data of added value in the models. We decided to input the same data across all base items for these data sets to prevent a significant variation of training data. We acknowledge that it risks adding unnecessary noise (especially for the endogenous data). However, if we had differentiated each data set for each model, it would mean that comparing the different data sets would be asymmetric, as each data set would be unique.

To determine which data should be left out of our models, we use an NN model. We trained the model as we have described in the section: *5.1.4 Neural Networks model*. For the historical data set, we experimented with inputting up to 12 weeks of past data (since this is the same as done with the benchmark method). The results can be viewed in: *Appendix A: Lagged historical data.* We determined that only inputting the most recent demand observation is the best input. For the endogenous and exogenous data, we summed all (absolute) weights to each neuron for each input variable since we only use a single hidden layer (see section: *5.1.4 Neural Networks model)* we can determine the total impact of a variable across base items. Since the weights are only relevant in their relativity to one other, we select the highest weights to be kept. From the endogenous data set in Table 2, we selected the 5% highest ranked variables. These are listed in: *Appendix B: Selected Endogenous Variables.* From the exogenous data set in Table 3, the following input variables were ranked significantly high: the GAS index (EU Gas prices), the Oil index (Crude Oil Prices (Europa)), and the PPI (Production price index). Only these three were ranked higher than the average, indicating their relatively high impact. Therefore, we include these input variables.

5.1.8 Summary of Models

We make one last addition to the ML models; we create three models for each ML model to evaluate the possible difference in performance when adding more data. The first ML model type only uses historical demand data. The statistical models, too, only use the historical demand. The second ML model type additionally includes the exogenous (input variables: GAS index (EU Gas prices), Oil index (Crude Oil Prices (Europa)), and PPI (Production price index)). The third ML model type additionally includes endogenous data (see section: *Appendix B: Selected Endogenous Variables*). The difference in input variables will likely result in different performances, allowing us to evaluate the usefulness of adding more (different types of) data. The models are summarised in Table 7.

Table 7: method, dataset, and parameters for all models we have created.

Method	Model	Data set	Parameters
Benchmark	*n/a*	Historical demand	Moving average = 12 weeks
Holt-trend	*ExponentialSmoothing*	Historical demand	Set with least squares method
ARIMA	*auto_arima_model*	Historical demand	Set with AIC method.
Neural network model one	*MLPRegressor*	Historical demand	Hidden layers = 1 Number of neurons = 1
Neural network model two	*MLPRegressor*	Historical demand and exogenous data	Hidden layers = 1 Number of neurons = 6
Neural network model three	*MLPRegressor*	Full dataset	Hidden layers = 1 Number of neurons = 45
K-Nearest Neighbours one	*KNeighborsRegressor*	Historical demand	K = 5
K-Nearest Neighbours two	*KNeighborsRegressor*	Historical demand and exogenous data	
K-Nearest Neighbours three	*KNeighborsRegressor*	Full dataset	
Gradient boosting one	*LightGBM*	Historical demand	Number of leaves: 50 Learning rate: 0,2 Number of rounds: 500 Max number of bins: 200
Gradient boosting two	*LightGBM*	Historical demand and exogenous data	
Gradient boosting three	*LightGBM*	Full dataset	

5.2 Inventory Model

This section will answer RQ4.2 (*How should the inventory model be designed?)*. We estimate the effect of forecasting on performance by creating a system that uses forecasts as input to determine inventory levels. This system enables us to use the RMS performance measure (equation 11). We use the same test data range to evaluate the performance of the forecasting methods from an inventory perspective. As done in: (Petropoulos et al., 2019): a continuous order up to system (OUT) policy with lost sales will be used to simulate inventory levels over the test data. The safety stock (s) is based on past forecast errors, and the order up to level (S) is based on the forecasted demand over the lead time. Thereby integrating the inventory system. The simplicity of the model is the main reason for its use. Assuming lost sales is done to eliminate the skewed effect that strong positive forecasting biases could have on the service rate. Time units for inventory decisions are single weeks and i indicates the week number. At the beginning of a week, an order may be placed following the policy. Equation 15 shows how we calculate the order up to quantity (S_i), where x_{i+3}, x_{i+2} and x_{i+1} indicate the forecast values over the lead time (two weeks) plus the review period of one week and ss_i the safety stock. This value is updated for each week, thereby integrating the forecasts (x_i) with the inventory policy.

$$S_i = x_{i+3} + x_{i+2} + x_{i+1} + ss_i \tag{15}$$

Safety stock is also continuously updated based on past forecast errors; see equation 16. Where Z indicates the safety factor (z score from the normal distribution) based on a cycle service level rate of 95%, L is the lead time (two weeks), R the review period, and σ_i indicates the standard deviation of past forecast errors.

$$ss_i = Z * \sqrt{L + R} * \sigma_i \tag{16}$$

After generating the safety stock and order up to levels for each week within the testing data, we model the order quantities. Order quantity is in units of 24000 (kg) as this is the average weight of a tanker truck delivery, and the vast majority of chemicals are delivered this way. The number of tanker truck-sized deliveries (K) is determined based on the difference between the order up to level (S_i) and current inventory level at the beginning of the week ($I_{i,b}$) minus the incoming order quantities ($q_{i-1}$and q_{i-2}). See equation 17 for its calculation.

$$q_i = \begin{cases} 24000 * K & if\ \ S_i > (I_{i,b} + q_{i-1} + q_{i-2}) \\ 0 & if\ \ S_i < (I_{i,b} + q_{i-1} + q_{i-2}) \end{cases} \tag{17}$$

After placing an order, the order takes two units (two weeks) to arrive. Orders are added to the inventory at the start of a week, and the demand is subtracted at the end of the week (if possible). The fulfilled demand is calculated with equation 18. Subtracting the fulfilled demand ($D_{i,fullfiled}$) from the inventory level ($I_{i,b}$) results in the inventory at the end of the week ($I_{i,e}$) as shown in equation 19. Equation 20 shows the inventory at the start of the week ($I_{i,b}$) is the inventory at the end of the previous week ($I_{i-1,e}$) with the added quantity of a (possible) order (q_{i-2}).

$$D_{i,fullfiled} = min(I_i,\ D_i) \tag{18}$$

$$I_{i,e} = I_{i,b} - D_{i,fullfiled} \tag{19}$$

$$I_{i,b} = I_{i-1,e} + q_{i-2} \tag{20}$$

This inventory system integrates the forecasted values from each forecasting model into a inventory policy. The demand levels, inventory levels, and order quantities according to this inventory system will be the input for the RMS measurement (equation 11), enabling us to evaluate the inventory performance of the forecasting models.

5.3 Conclusion of the Model Design(s)

In this section, we will answer each research question in this chapter, concluding our model design.

- RQ4.1 How should the forecasting models be configured?
  - The benchmark method is developed into a model with a moving average of the past 12 weeks. For the Holt method, we created a model with the least squares method, allowing us to set smoothing parameters for each base item. We use a model that optimises the parameters based on AIC for the ARIMA method. Selecting the least complex model with the best fit for each base item. We saw no significant difference between using more/less layers for the Neural network method. Thus, we chose to use the one-layer model for it is the simplest model, with one neuron per input variable. The weights assigned to each connection are set via the LBFGS algorithm. Each neural network is individually trained per base item, but its topology remains the same across base items. For the KNN method, our data set is small enough to allow for a brute-force approach, using k=5 as the best setting. After we applied a grid search, we decided to set the parameters of the lightGBM model as such: number of leaves: 50, Learning rate: 0,2, number of rounds: 500, and max number of bins: 200. All these models are summarised in Table 7.
- RQ4.2 How should the inventory model be designed?
  - In section: *5.2 Inventory Model,* we show how we construct an order up to a system that uses the forecasts of the models to determine the safety stock (ss) and order up to level (S), with orders in units of 24 tons (representing tanker truck deliveries).

Thus, we have answered RQ4 (*How are the models designed?*). The next chapter presents the results from the models defined in this chapter.

6. Results

In this chapter, we analyse the performance of the forecasting models, thereby answering RQ5 (*What is the performance of the models?).* In the first section: *6.1 Forecasting Performance,* we analyse the forecasting performance of our models. In the second section: *6.2* Extending the Forecast Horizon*,* we analyse the forecasting performance when we increase the forecast horizon. In the third section: *6.3 Inventory Performance,* we analyse the inventory performance of our models. In the last section: *6.4 Best Model(s),* based on all results, we determine which model or combination of models should be implemented and estimate the effect its inventory reduction.

6.1 Forecasting Performance

In this section, we will analyse the results from the forecasting measures: bias, MAE, MSE, and sMAPE. We will start with the general performances before diving into more specific results.

6.1.1 General performance

In this section, we will answer RQ5.1 (*Which models perform best according to the forecasting performance?*). We calculated the number of improved base items by comparing the measurement results of the models to those of the benchmark model. In Figure 11, we show the number of improved base items according to each measurement for each model. We observe a degree of inconsistency between the measurements. The MSE shows a relatively low number of improved base items for the LightGBM and KNN compared to the other measurements. We suspect this (low number of improved base items according to MSE) to be due to the squaring of errors punishing larger errors, which might occur only a few times. MAE and bias deviate positively from the ML models. The sMAPE deviates the least from other measurements across most models. The bias is improved for more base items by the ML models compared to the statistical models. Correctly forecasting intermittent periods could reduce bias. This is impossible for the statistical and benchmark models, whereas ML models do not have this drawback, possibly explaining the difference. The inconsistency (of MSE, MAE and sMAPE) indicates variability in the size of errors and is a testament to our choice to include these multiple measurements.

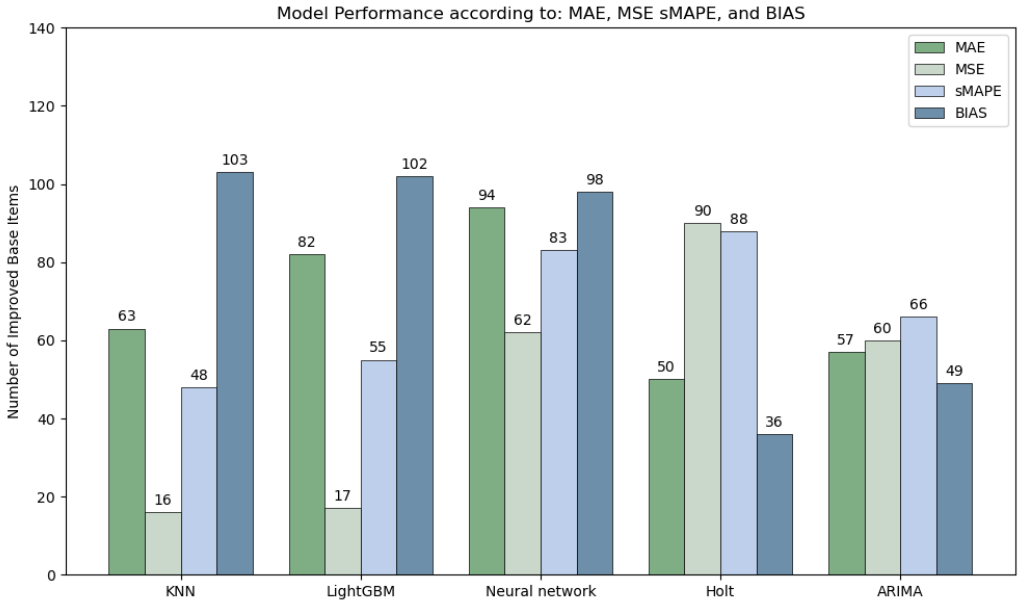

Figure 11: Number of base items improved by models according to MAE, MSE, sMAPE, and bias compared to the benchmark model.

We need to decide when a model can be considered an improvement compared to the benchmark model. Luckily, for most models, at least two measurements correlate with how many base items are improved (shown in Figure 11). Therefore, we consider a model to improve forecasting for a base item when at least two of the three error measures, MSE, MAE and sMAPE, are less (and therefore indicate an improvement) than those of the benchmark model. We do not include bias as it solely indicates

systematic over/under forecasting and not the error size of forecasts. The results are shown in Figure 12, where the number of improved base items is shown at different thresholds. We will explain the graph with an example: the bar on the left in Figure 12 illustrates that the KNN model improves 48 base items (under the requirement of at least two measurements being improved). Of these 48 base items, 29 have at least two measurements, which are improved by a minimum of 10%, indicated by the purple part of the bar. We chose the 5% and 10% thresholds to illustrate the degree of improvement.

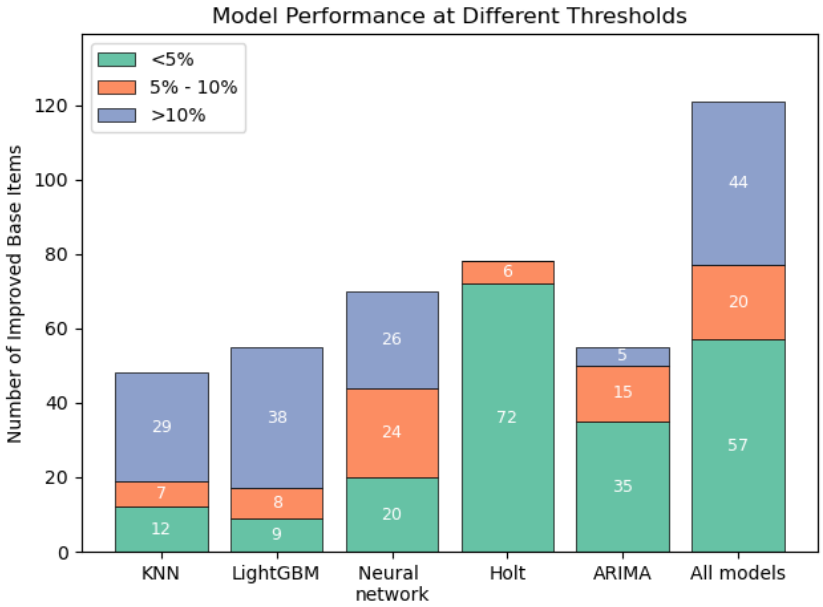

Figure 12: Number of improved base items at different thresholds.

We can see that we have a better model for 121 base items, as shown by the bar on the far right in Figure 12. The Holt model improves the highest number of base items. However, the Holt and ARIMA models improve significantly fewer base items for the highest threshold than the ML models. This indicates that the ML are very specialised for certain base items. The difference between the number of base items improved across all models (indicated by the bar on the far right) and the number of improved base items for each individual model is quite large. Therefore, we will not find a universally applicable model. Possibly, a characteristic of certain base items that are better forecasted by the ML models, which we discuss in the next section: *6.1.2 Demand type*. Another possibility is that the incorporation of additional data sources might allow these models to model certain relations found in these specific base items, which we discuss in the later section: *6.1.3 ML performance.* For now, we have shown how the models perform compared to each other, answering RQ5.1 *(Which models perform best according to forecasting performance?*).

6.1.2 Demand type

In this section, we will answer RQ5.2 (*Which models perform best according to forecasting performance for the different types of demand?*). Again, we will consider a model to improve forecasting when at least two measurements indicate an improvement. Figure 13 shows the number of improved base items for each demand type by each model. For example, KNN improves forecasting for ten base items of smooth demand, as shown by the dark blue bar on the far left in Figure 13. We see that the Holt model improves for the most base items belonging to the smooth demand followed by the Neural network. For the intermittent class, we see a tie between the lightGBM, NN and Holt models. For the lumpy demand, all ML models improve more base items than the statistical models. The Holt model also improves the most for the erratic class. However, since we saw a significant difference in improvement numbers at different thresholds (Figure 12), we need to evaluate the level of improvement.

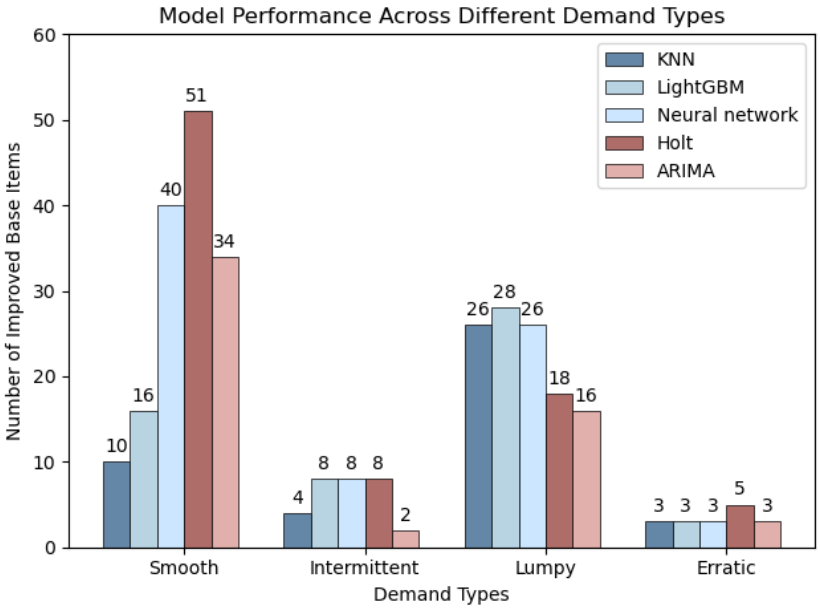

Figure 13: The number of base items improved per demand type by each Model.

To consider the level of improvement, we again visualise in Figure 14 the number of base item classes per demand type, but now at an improvement threshold of 5% for two out of the three measurements. The Holt model no longer improves the most base items for the smooth demand type. The NN model is the best for smooth demand at this higher threshold. The lightGBM improves for the most base items at this threshold for the lumpy and intermittent demand type and is tied with the NN for the erratic demand type. At this higher threshold, the ML models strongly outperform the statistical models in the number of improved base items, especially the lightGBM and NN models.

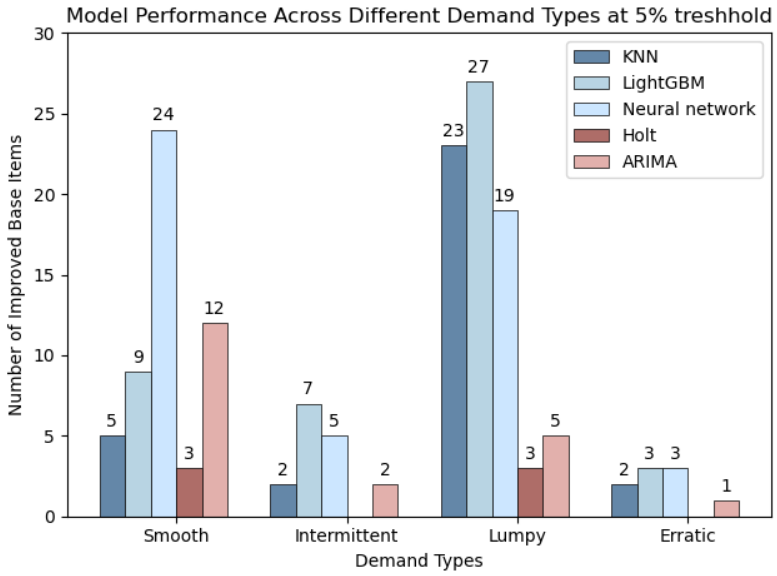

Figure 14: The number of base items improved per demand type by each model at a 5% threshold.

In the previous section (*6.1.1 General performance)*, we found no universal model, and now we observe a difference in the performance of models across demand types (Figures 13 and 14). These demand types could explain why we found no universal model. Holt improves the most base items for the smooth demand type, but the NN model improves more base items at a higher threshold. The specialisation for demand types is most prevalent for the ML models. These are more specialised for the intermittent and lumpy demand, especially at the 5% threshold. These results indicate that the benchmark model could be easily replaced with either lightGBM or NN for many base items belonging to the intermittent and lumpy classes. For the smallest class, the erratic demand type, we see that ML

improve more base items than the statistical models. With this, we have answered RQ5.2 (*Which models perform best according to forecasting performance for the different types of demand?*).

6.1.3 ML performance

In this section, we will answer the following question: RQ5.3 (*Does more data improve forecasting for ML models?*). Again, we will consider a model to improve forecasting when at least two measurements indicate an improvement. We calculated the number of base items improved by each model when trained on different data sets. We will refer to the data sets as dataset one (containing only historical demand), dataset two (containing historical demand and exogenous input variables) and dataset three (containing historical demand, exogenous input variables, and endogenous input variables). These data sets only contain the selection of variables we determined to have a significant impact (described in section: *4.1 Data Collection*)

In Figure 15, we visualise the number of improved base items by the ML models when trained on different data sets. For example, the blue column on the far left shows that KNN improves 34 base items when trained on data set one. When LightGBM and NN are trained on the first data set (containing only historical demand), they improve for the most base items and almost twice as many base items as the other data sets. This indicates that the historical demand data set is the better option and that the other data sets offer improvement for fewer base items. For the KNN model, training it on data set three improves the highest number of base items. However, the difference between data sets is minimal for the KNN model.

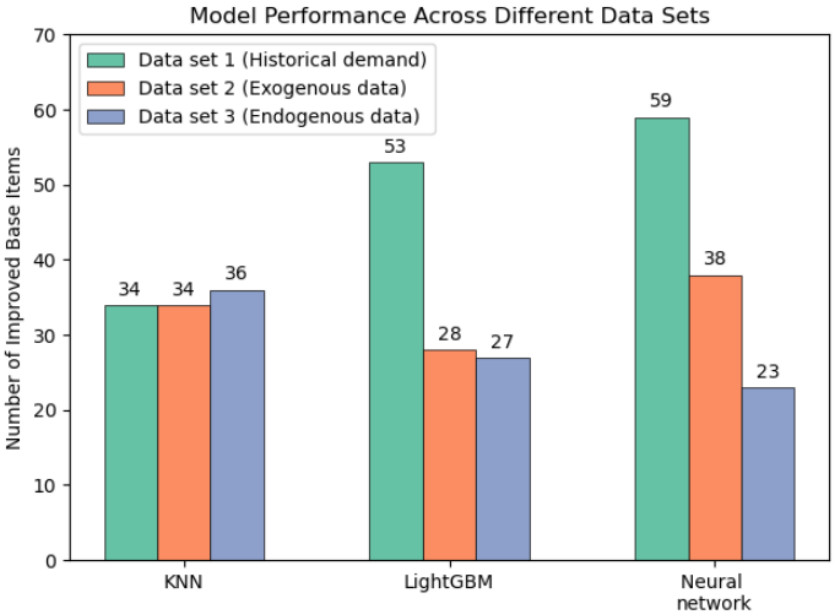

Figure 15: Number of base items improved by models trained on different data sets.

However, Figure 15 might be misleading; data sets might not contain the same base items, and therefor their added value might be overlooked. Additionally, forecasting for base items might be improved more by adding data. For this purpose, we calculated the number of improved base items from a model trained on the second data set, which is not improved when trained on data set one or is improved most by the second data set (meaning at least two measurements indicate such). Furthermore, we calculated the number of improved base items by a model trained on the third data set, which is not improved when trained on the first or second data set or is improved most with the third data set (meaning at least two measurements indicate such). We chose to compare the data sets in this manner to evaluate the (potential) benefit of adding more data. We visualise this in Figure 16, where the orange bar on the far left indicates that 15 base items are more improved by training KNN on the second data set, compared to the first data set. Figure 16 shows that the number of improved base items when trained on data set two is 15 and 20 for KNN and NN, respectively. Including the exogenous input variables (part of data set two): 'GAS index', ‘Oil index', and 'PPI' for the KNN and NN models

improves forecasting for these base items. We will not disclose which products belong to this selection; however, we will try to find why we observe an improved forecasting ability by adding this data. According to the quality department, there is no shared chemical property such as being produced from hydrocarbon sources (which would relate to the oil index). From a production perspective, there is no higher sensitivity to price fluctuations compared to the norm (which could have been related to gas prices) for this selection, according to the purchasing department. From a customer perspective, there is no known application for these chemicals which are (predominantly) related to these variables. From a demand perspective, the distribution of base items in this selection was comparable (with no more than 5% deviation) to the general distribution, finding no explanation in the demand classification. There are three possible likely explanations for why we cannot explain why these base items benefit from the exogenous data. First of all, it could be possible that the improvement is simply coincidental, however, considering that we require at least two measurements to indicate an improvement we estimate this to be less likely. Secondly, there could be an unknown relation between these variables and the product demand, which we are unaware of, which would explain the improvement. Lastly, considering the generality of the exogenous variables, it could be the case that the relation is valid for all chemicals but only correlates with this selection because their historical data better reflects a broader trend in chemical demand. For data set three, the number of improved base items is very low across all models. Indicating that the additional endogenous input variables do not offer much improvement except for a tiny selection of base items. The low number could indicate that these variables either do not model demand patterns accurately and, therefore, offer no valuable information. Alternatively, the variables could describe demand patterns that the model already accounts for based on the historical data, thereby only adding noise to the model.

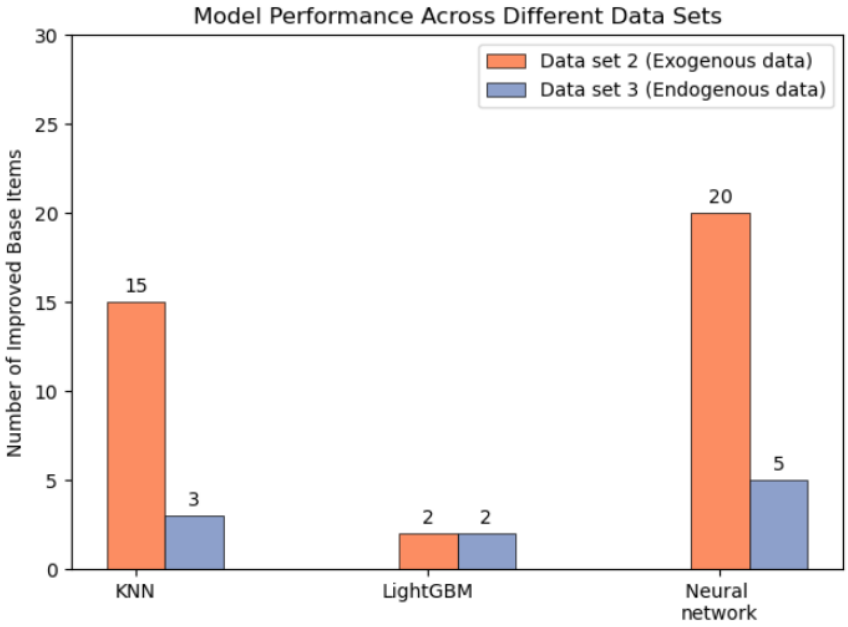

Figure 16: Number of base items improved by data set two and three.

We observe that training solely on historical demand (data set one) results in the most significant number of improved base items for all models. The exogenous input variables (part of data set two): the 'GAS index', 'Oil index', and 'PPI', offer added value for a smaller selection of base items when incorporated into the NN and KNN models. Adding endogenous data (data set three) only benefits a tiny number of base items. Thus, we have answered RQ5.3 (*Does more data improve forecasting for ML models?*).

6.2 Extending the Forecast Horizon

In this section, we will answer RQ5.4 (*How does a longer forecast horizon affect forecasting performances?*). We decided to extend the forecasting horizon (from three weeks) to six weeks and 26 weeks. With this range, we evaluate the robustness of the forecasting models regarding the extension of the horizon. Once again, we consider a model to improve a base item when it improves at least two

measurements compared to the benchmark model. Figure 17 shows the number of improved base items for the different forecasting horizons. For example, the three bars on the far left show that the KNN model improves forecasting for 46 base items at a horizon of three weeks, 95 base items at a horizon of six weeks, and 101 base items at 26 weeks.

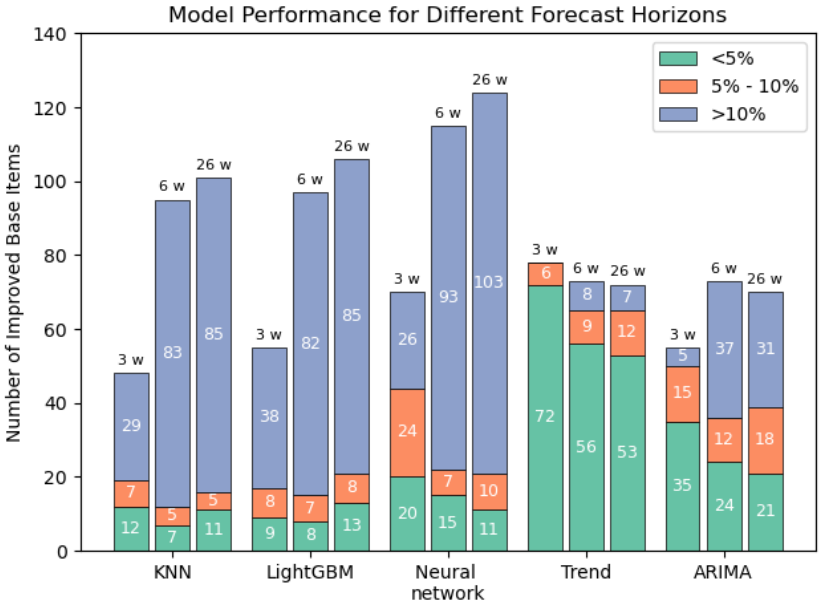

Figure 17: Model performance for different time horizons: three, six and 26 weeks.

Figure 17 shows a significant increase in improved base items for the ML models with a longer horizon, especially at the highest threshold. For the statistical models, we see a minor increase in the number of improved base items in the higher thresholds for the longer horizons. The relative significant increase for the ML models compared to the statistical models could be because the statistical models do not differ much from the benchmark model. The ML might capture patterns in the demand which it can use to forecast more accurately over longer horizons. Figure 18 shows the number of improved base items for the different forecasting horizons when trained on different data sets under the same conditions as in Figure 17. This will allow us to evaluate if adding data is more beneficial on longer horizons. For data set two, we observe an increase in the number of improved base items for the longer horizons, especially for the KNN model. The increase for date two for the KNN and LightGBM is comparable to the relative increases seen in Figure 17 for the total number, indicating that the data can be more valuable for longer horizons. For data set three, we observe a decrease for the KNN and NN model at the extended horizons but an increase for the LightGBM.

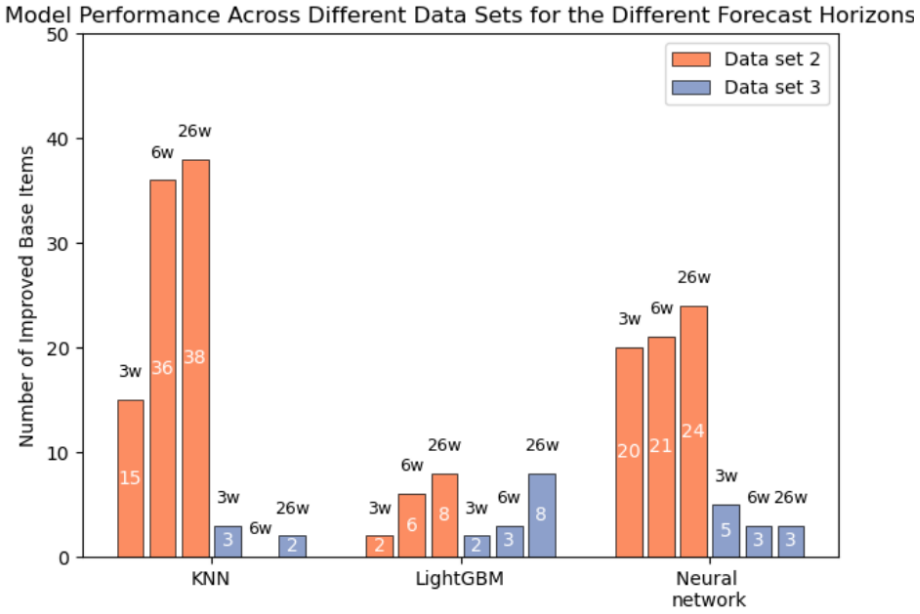

Figure 18: Model performance for different time horizons

We observe the robustness of the models in maintaining their performance compared to the benchmark when a longer forecast horizon is used. The ML models perform significantly better than the benchmark at this longer horizon. Regarding the data sets, we observed that more base items are improved when adding the exogenous variables (data set two) at the longer horizons compared to the standard horizon. Using data set three still improves low numbers at the extended horizons. As a result of this, we have answered RQ5.4 (*How does a longer forecast horizon affect forecasting performances?*).

6.3 Inventory Performance

In this section, we will answer the following question: RQ5.5 (*Do models forecasting performances differ according to inventory performance?*). In Figure 19, we show the same data as in Figure 121; however, we now include the RMS measure (represented by the yellow bars). RMS is calculated according to equation 11 after integrating the forecasts into the inventory system as described in the section: *5.2 Inventory Model.*

The ML models improve more base items according to the RMS than the statistical models. We observe the LightGBM model improving the inventory performance for the highest number of base items. This is logical if we consider that the number of base items improved across all the error measures is relatively high to the other models (except for the NN model performing very well across metric too). The RMS measurement does not correlate with any specific forecasting measurement but does result in comparable percentages for at least one or two different measurements. In Figure 19, we can observe that RMS has numbers comparable to the bias measurement for the ML models. Reducing the bias could result in better inventory performance as less over/understocking would occur. However, we see that for the statistical models, there is a less comparable number between the RMS and bias. However, a relatively lower number of base items, according to the bias, correlates with a lower number of improved base items, according to the RMS. We suspect that inventory performance is most strongly correlated to the bias and, to a lesser degree, with the other measurements.

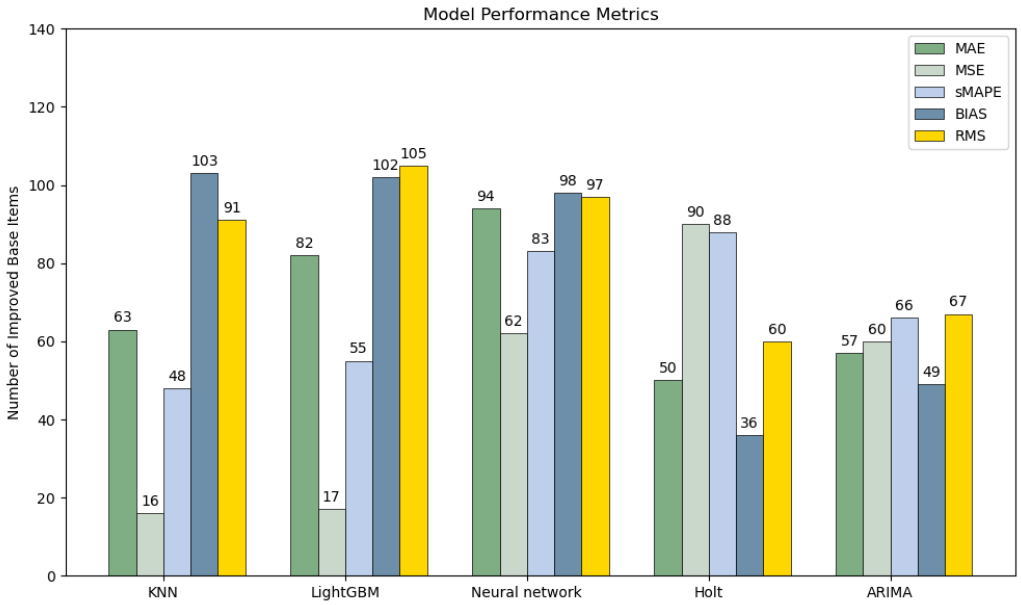

Figure 19: Percentage of base items found to be best forecasted by each model according to RMS and the error measures.

We conclude that the forecasting performance of the models is converted to inventory performance, especially for the ML models. This is what we would expect: an improvement in forecasting performance to improve inventory performance. As a result, we have answered RQ5.4 (*Does the forecasting performance of models differ compared to inventory performance*).

6.4 Best Model(s)

In this section, we combine all results into a solution and answer RQ5.5 (*Which model(s) should be implemented based on our results?*). Since we found no single model to be clearly dominant, we will evaluate if multiple models could be the best solution. We try to minimise the number of different models for practical considerations; introducing more models would require more resources, which is not preferred unless significantly beneficial. We will evaluate several solutions (primarily) based on the results discussed in the sections: *6.1.1 General performance*, *6.1.2 Demand type*, and *6.3 Inventory Performance*. We will estimate the reduction in inventory costs by analysing the inventory levels we have calculated, as described in section: *5.2 Inventory Model.*

The NN model performed relatively well; it improved the greatest number of base items at a threshold of 5% or higher (Figure 12). According to the RMS measurement, it also had the second-highest number of improved base items and performed relatively high across the other measurements (Figure 19). Regarding robustness, it had the highest increase in the number of improved base items across all models compared to the benchmark (Figure 17) at the extended horizons. The NN model performed best for the smooth class (Figure 13). For the erratic class, it ties with the lightGBM model as the best model (both improving the same number of base items (Figure 14)), and for the intermittent class, it is slightly outperformed by the LightGBM model (Figures 13 & 14). For the lumpy class, lightGBM improved more base items (Figure 13). Considering all this, we chose to evaluate the NN model and see if using the LightGBM model for the Lumpy class is of added benefit. Another possible solution is using NN with Holt for the smooth class as Holt improved more items at the lower threshold. Additionally, as LightGBM is one of the better-performing models (Figure 12), we evaluate it as an alternative to the NN model. Regarding data, we observed no benefit in using additional data for the LightGBM model (Figure 16), and therefore, we would only use data set one to train this model. We observed the second data set (comprised of exogenous data) to increase the number of improved base items for the NN model (Figure 16). Therefore, we evaluate if selectively training (based on the RMS) NN on the second data set reduces inventory for the base items and improves.

To summarise, we evaluate five possible solutions: only NN, NN with LightGBM for lumpy demand, NN with data set one and two, NN with Holt for smooth demand, and only lightGBM. These solutions

represent the best-performing combinations. These are summarised in Table 8; we show the number of base items that have a reduced inventory (using equation 12) while having an acceptable service rate of at least 95% or maintaining the same (or higher) service rate (using equation 14) as the benchmark model. The average percentage reduction across the improved base items is shown, as well as the percentage difference between the best solution and each solution regarding total inventory reduction.

Table 8: Possible solutions and their effect on inventory.

	Solution description	Number of base items with reduced inventory	Average % inventory reduction across improved items	% difference in reduction compared to the best solution
1	NN (data set one)	65	22,6%	-
2	NN (data set one) with LightGBM for lumpy class.	66	17,9%	-0,1%
3	NN (both data set one and two)	65	21,7%	-1,2%
4	NN (data set one) with Holt for the smooth class.	54	9,3%	-23,4%
5	LightGBM	61	19,6%	-15,4%

The greatest reduction of inventory is achieved with the first solution: the NN model trained on data set one. Reducing an average of 22,6% of inventory across 65 base items while maintaining the service rate. All other solutions reduce inventory less than this solution, additionally, it improves the second greatest number of base items. Of the 65 base items, 19 base items even show an improved service rate compared to the benchmark model, with an average four percentile point improvement. In this group of improved base items, five erratic, four intermittent, 13 lumpy, and 43 smooth base items are found. Considering the ratio of the 65 base items from this selection, the largest difference between this selection and the general distribution for the smooth type, where NN overperformed. However, percentage wise it is only 6% difference. Indicating that this selection is quite representative and features no strong over/under representation of any demand type. There is a discrepancy between the number of base items for which we reduce inventory with the NN model and the number of base items for which an improved RMS value was found (Figure 19). This can be explained by the requirement of maintaining the service level, while for the RMS value, a lower service level can be compensated by a reduced inventory or less deviation in ordering. Therefore, this difference is to be expected.

6.5 Conclusion of the Results

In this section, we will answer each research question in this chapter and conclude our results.

- RQ5.1: Which models perform best according to the forecasting performance?
    - We observed a discrepancy between the forecasting error measures. Therefore, we decided to consider a model to improve forecasting if two error measures are lower than those of the benchmark model. We observed that there is single model to clearly perform better than others. We did observe that the ML models improve significantly more base items at a higher threshold than the statistical models. With the best statistical model (ARIMA) improving 15 less base items compared to the worst ML model (KNN) at the higher thresholds.
- RQ5.2: Which models perform best according to forecasting performance for the different types of base item classes?
    - For base items belonging to the lumpy and intermittent demand types, the LightGBM improved the highest number of base items at both thresholds (28 items), with NN and KNN both being a close second (26 items). For base items belonging to the smooth demand, the Holt model improves the most items (51 items), but the NN improves significantly more base items at a higher threshold (24 items compared to three for Holt). For base items belonging to the erratic demand type, the Holt model again

improves the most models (eight items), but at the higher threshold, we observe a tie between the NN and LightGBM (three items compared to zero for Holt).

- RQ5.3: Does more data improve forecasting for ML models?
    - We observed that ML models trained solely on the historical demand data improve the highest number of base items, except KNN, which improves around 35 base items across data sets. The KNN and NN models improved a significant number of base items when given exogenous data (15 and 20 respectively), which were not improved without this data. The third data set, with endogenous data, only improved a tiny number of base items (with a max number of five for the NN model). Therefore, adding more data does not necessarily improve forecasting but can benefit a small selection of base items for the KNN and NN models.
- RQ5.4: How does a longer forecast horizon affect forecasting performances?
    - Increasing the forecasting horizon to six and 26 weeks increased the number of improved base items for the ML models, especially at the higher thresholds (almost doubling for all models). The number of improved base items at different horizons was comparable to the statistical models, but the number of base items at higher thresholds increased for longer horizons (also at least doubling). Regarding data, the second data set improved many more base items for the KNN model at higher horizons (again more than doubling), while the third data set remains of little use.
- RQ5.5: Do model performances differ according to inventory performance compared to forecasting performance?
    - The number of base items improved according to the RMS value (representing inventory performance) did not strongly correlate with the MSE, MAE, sMAPE measurement. We observed a correlation between the high number of improved base items for the bias and RMS for the ML models. According to the RMS, the NN model had the second-highest number of improved base items and had a high ranking across all forecasting measurements. Indicating that improving forecasting performance results in improved inventory performance.
- RQ5.6: Which model(s) should be implemented based on our results?
    - We determined that using the neural network model trained on historical data is the best solution. This solution results in the greatest reduction of inventory levels, reducing inventory by an average of 22,6% across 65 base items (constituting 47% of the base items). Even improving the service level (fill rate) by an average of four percentage points for 19 base items.

In this chapter, we have answered RQ5 (*What is the performance of the models?*). This concludes the research phase*: choosing a Solution.* In the next chapter, we will discuss the implementation of this solution.

7. Implementation

In this chapter, we present a proposal for implementing the NN model. This chapter constitutes the last research phase: implementation. We will formulate which steps are needed to successfully implement the solution, thus answering RQ6 (*How can the most appropriate model(s) be implemented?*).

7.1 Forecasting Process

As described in the section: *2.3.1 Process of Forecasting*, forecasts are currently only generated when inventory levels drop below a certain point and are subject to judgemental interpretation showing that employees are not used to systematic forecasting. Therefore, we advise not to replace the judgemental forecasting process in its totality but to replace the current quantitative forecasts (moving average of the past three months) for the base items improved with our solution.

The NN model must be incorporated in an easy-to-use application. The development of this application should be done in collaboration with the IT, sales, purchasing, and continuous improvement departments to ensure a correct and easy-to-use application. The development should be done under the supervision of the continuous improvement department. The purchasing and sales department should advise on the accessibility and layout of the tool. The technical support should be provided from the IT department. The model should be set according to the parameters we have defined in the section: *5.1.6 Neural Networks model.* Customer orders with shipment data should be sourced from the ERP system, loaded in the application, and aggregated into weekly demand. This data should be used to train the model (as we did in this research); however, if the reliability of the requested delivery date is increased, we suggest using this date (elaborated further in the section: *8.3 Recommendations*). Vivochem needs a suitable platform to integrate the models. We suggest exploring the possibility of visualisation of the forecast within a Power-BI dashboard. Employees are already used to this application, as these dashboards are widely used throughout Vivochem. Power-BI can be programmed with Python, and since this research effort also used Python, it will be suitable. In this dashboard, an employee should be able to select a base item and see both the generated forecasts and additional information. Additional information is necessary as forecasts are currently subject to judgemental interpretation and are based on more information. The purchasing department should be consulted on which information they value. Additionally, we suggest showing the past demand of a base item within the dashboard to allow employees to understand the context of the forecasts. We estimate the integration of this model to last less than one month, considering debugging and data preparation.

To improve the forecasting process, we advise generating forecasts automatically without considering inventory levels, changing from a reactive to a proactive purchasing process. Considering the delivery and lead times, we recommend that the models be (re)trained weekly; this will allow forecasting models to be up to date without unnecessarily constraining the ERP system's capacity. We estimate the computational time to be less than an hour for a regular operating system, an acceptable duration that can efficiently run on weekends when automated, thereby not restraining the ERP system or having employees waiting for forecasts to be generated. Additionally, since these processes are subject to lead times of one or two weeks, generating forecasts more than once per week will not be necessary. The forecasts should be generated for a forecast horizon of three weeks to cover suppliers' lead times. We have established that the NN model maintains performance at longer forecast horizons; therefore, we would advise developing the option to extend the forecasting range within the application.

7.2 Training

After the application has been developed and tested, the purchasing department's employees should be trained. During the training, how to use the application, how forecasts are generated, which data are used, and when new forecasts are generated should be explained. Additionally, the employees should be informed of the forecasting system's limitations, preventing them from misinterpreting forecasts. The necessary information can be derived from this research effort, and the training should be the responsibility of the project manager of the continuous improvement department. We estimate

the training of employees to be done in a single session of one to two hours (provided that the employees are already aware of the project).

7.3 Measurement & Monitoring

We advise the project manager to carefully monitor the forecasts generated by our model, especially in the first months of implementation, to ensure the successful transformation of the solution into an operational process. To measure the effectiveness of our new forecasting system, we advise using key performance indicators (KPIs). We suggest using KPIs to measure both the forecasting and inventory performances.

We suggest using the three forecasting measurements we have used in this research to estimate forecasting performance, preferably integrating these into the dashboard itself. This will allow Vivochem to continuously measure the forecasting model's performance and recognise when forecasts are less reliable (for example, if a fundamental change happens to a demand pattern). Additionally, it might allow for improvement of the forecasting model by evaluating how the real-time conditions affect forecast performance. Since this research aims to lower inventory while maintaining the service rate, we recommend measuring the inventory performance. Since purchasing orders are strongly influenced by economic decisions which cannot be translated into forecasting systems, we recommend collecting data to measure how well the forecasts are translated into purchase orders. This would additionally allow Vivochem to evaluate its purchasing department's decision-making performance, potentially in the form of continuous measurement of the throughput speed of its inventory. Another measurement we advise is the measuring of systematic over and understocking. Should the forecasting models improve the decision-making ability of the purchasing department, these performance indicators will quantify the improvement. Tracking forecasting and inventory performances should be communicated with the employees who influence these indicators.

8. Conclusion

In this chapter, we will bring this research to an end. We answer our main research question in the section: *8.1 Conclusion*. Next we will discuss the limitations of our research in the section: *8.2 Discussion*. Given our conclusion within the context of the limitations, we formulate recommendations for further research in the last section: *8.3 Recommendations*.

8.1 Conclusion

Throughout this research effort, we have answered many sub-research questions, which allow us to answer the main research question: *"How can the sporadic chemical demand be forecasted while maintaining an acceptable service rate and reducing inventory?"*.

We determined that the best solution is using a NN model (trained on historical demand) with a single hidden layer with three neurons to forecast demand. This model reduces the inventory of 65 base items by an average of 22,6% while maintaining an acceptable service rate and increases the service rate for 19 of these base items by an average of four percentile points. We have not reduced inventory for the remaining base items while maintaining the acceptable service rate. With this, we have answered the main research question.

The improved forecasting and inventory performance will enable Vivochem to better align its inventory with customer demand. There are many benefits associated with this improvement. First, it will reduce inventory for these chemicals by 22,6%, freeing up warehouse space and reducing inventory costs. Secondly, reducing costs makes the associated financial resources available for other purposes. Thirdly, for 19 base items, we observed a four percentile increase in the service rate (fill rate), which improves reliability and customer experience. Lastly, the purchasing department can improve its decision-making as its demand expectation is improved due to the higher accuracy of the new forecasting model. This enables more economically responsible decision-making and decreases the number of over and understocking occurrences.

8.2 Discussion

In this section, we will discuss our research's limitations and how these impact the reliability and reproducibility of our results. This section aims to establish an understanding of what we can and cannot conclude from our results.

8.2.1 Data

During our research, we used shipment data to represent the historical demand. Shipment data could be lagged as it does not equal the exact moment of demand. However, since Vivochem strives to supply its customers within two weeks, the skewness of this data is estimated to be minimal on the large scale of 12 years of data. We considered it the best available data, and it was even necessary to use this data due to the worse data quality of the alternatives.

8.2.2 Model Parameters

We used methods for the statistical models to determine the optimal parameter settings. However, we tested a range of possible settings for the ML models experimented and estimated the performance based on the average percentile differences in the error measurements. While this is an acceptable method, we cannot be certain that we have reached optimal settings. Therefore, the possibility remains that a model could have performed better if a greater range of settings was researched. Thus, we do not rule out that future efforts may improve the model. However, we deemed the range of settings we researched sufficient and the possible leftover improvement minimal. If the remaining optimisation is indeed minimal, we reason that further optimisation would not lead to a different solution but would only improve the performance of the proposed NN model.

8.2.3 Consistency in Measurements

One important observation we made is the inconsistency between the forecasting measurements. We have theorised an explanation for this phenomenon: a small number of relatively large errors are the cause. However, we expected to observe different results from the measurements, so we chose

these measurements (MSE, MAE, and sMAPE). We negate the drawbacks of individual measurements by using distinct error measurements with all their sensitivities. We negated the inconsistencies by requiring two error measurements to be improved (compared to the benchmark) to consider a base item to be improved by a new model. Additionally, we differentiated the level of improvement by introducing different thresholds. Doing so allowed us to weigh the degree of improvement with the interpretation of the results. This approach increases the trust in our conclusions since they are based on multiple measurements and the level of improvement.

8.2.4 Sensitivity Analysis

We have researched increasing the forecast horizon but have not considered optimising the model parameters for these alternative conditions. If parameters were optimised for this specific setting, models could have performed significantly differently. However, we extended this horizon to evaluate the model's robustness. We consider the possibility of further optimisation at this out-of-scope but recommend further research. With this, we identified that the effective range of the models is greater in comparison to the benchmark for a great number of base items. It enables Vivochem to consider the possibility of using the new forecasting model beyond the standard settings.

8.2.5 Demand Types

We observed certain types of demand that were better forecasted than others. Should Vivochem ever want to expand forecasting to base items that are not part of this research, we advise comparing the benchmark model's performance with the NN before assigning a model to a new base item. This preliminary comparison prevents redoing the whole research but still choosing the best model for a new base item. Another potential problem could be if a significant shift in demand occurs and a base item starts exhibiting another demand pattern, which could mean another forecasting model is a better fit. This could happen regardless of which model our research proposes. We consider this a possibility we cannot foresee and covered by our proposal to monitor the forecasting models after implementation.

8.2.6 Inventory Model

One major obstacle for our research was the non-existence of a current systematic inventory policy. Orders are subject to economic and operational limitations. The order policy, we designed does not fully represent Vivochem’s ordering process. However, we were forced to approximate it due to the necessity of estimating the inventory reduction. Therefore, the exact inventory reduction is most likely to differ, but we expect a comparable reduction as we have applied the same inventory policy to the benchmark model.

8.2.7 Literature performance

Our final solution consist of a neural network, due to it having the greatest inventory reduction. While literature and particularly the M-competitions state the LightGBM model is one of the best forecasting models. This gap between literature and can be due to various reasons. First of all, LightGBM has been performing well for larger data sets and might not perform similarly for our data set. Additionally, the irregularities in our data set might be more significant than in the literature. Secondly, LightGBM might not have shown the greatest reduction in inventory it could be argued that it would be second best option as it too performed relatively well across metrics (which is accordance with literature). Thus, it could be simply the case that the inventory model is better suited to translate the forecasts of the NN into inventory performance than the forecasts of the LightGBM. Therefore, literature might indicate that LightGBM is a well performing forecasting model it does not necessary be the best option for this research context.

8.3 Recommendations

In this section, we give several recommendations to Vivochem on how to improve upon this research:

- The execution of our implementation plan to integrate the forecasting models for the 65 base items we have identified to reduce inventory costs while maintaining an acceptable service rate. Thereby reducing overstocking, reducing inventory costs, freeing up the associated financial costs, improving customer service and enabling purchasing to improve their decision-making.
- Researching the possible application/development of a systematic inventory policy, which is currently non-existent. Our models will improve inventory performance, but their impact will remain limited if inventory decisions are not made systematically. We would especially recommend

researching if base items for which we have not observed any improvement in forecasting performance should be set to 'order on demand' due to (possibly) unpredictable demand.

- Determining if an extended forecast horizon could be valuable for the purchasing department. We have observed that the majority of models (especially the NN model) maintain their performance with an extended forecast horizon (compared to the benchmark). In that case that economic benefit can be achieved at a greater horizon, then Vivochem should optimise the settings of the NN model and use this model in addition to the one from this research.
- Researching the accuracy of aggregated forecasts from different models. Different models might capture different parts of demand patterns, and combining these forecasts might be more accurate.
- Researching the possibility of developing hybrid models and their performance. Some literature sources suggest that hybrid models are good-performing models. However, not as much literature is available as for pure models. While this has been left out of this research effort, a hybrid solution might be a possible improvement to the current solution.
- Adding more data to the forecasting models:
    - Adding the daily time bucket to the forecasting input data to potentially capture more detailed demand patterns. Our forecasting models were based on the weekly demand, as this is the current time bucket used by Vivochem in inventory decisions.
    - Another possible improvement is incorporating the data from the whole company group. This could improve the forecasting model as the data would represent a larger share of the chemical market and more accurately represent demand.
- Research whether our solution models can be used to forecast product-specific demand. Instead of forecasting at the base item level, forecasting at the product level could enable even greater reduction in specific inventory levels, aligning inventory more with customer demand. This would have the additional benefit of improving the efficiency of the dispensing process.
- Replacing the shipment date with the requested delivery date by increasing reliability via automation. If the requested delivery date is continuously updated and sufficiently reliable, forecasting accuracy will be improved as the demand data will be more representative of actual demand.

9. Bibliography

Adur Kannan, B., Kodi, G., Padilla, O., Gray, D., & Smith, B. C. (2020). Forecasting spare parts sporadic demand using traditional methods and machine learning-a comparative study. SMU Data Science Review, 3(2), 9.

Akaike, H. (2011). Akaike's information criterion. International encyclopedia of statistical science, 25-25.

Allende, H., Moraga, C., & Salas, R. (2002). Artificial neural networks in time series forecasting: A comparative analysis. Kybernetika, 38(6), [685]-707.

Axsäter, S. (2015). Inventory control (Vol. 225). Springer.

Barros, F. S., Cerqueira, V., & Soares, C. (2021). Empirical study on the impact of different sets of parameters of gradient boosting algorithms for time-series forecasting with LightGBM. PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part I 18,

BEA. (2024). Gross Domestic Product Federal Reserve Economic Data (FRED). https://fred.stlouisfed.org/series/GDP

Box, G. E. P., & Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day Inc.

Broeren, M., Saygin, D., & Patel, M. (2014). Forecasting global developments in the basic chemical industry for environmental policy analysis. Energy Policy, 64, 273-287.

Bundgaard-Nielsen, M. (1972). Forecasting in the chemical industry. Industrial Marketing Management, 1(2), 205-210.

CBS. (2024a). Jaarmutatie consumentenprijsindex; vanaf 1963 CBS. https://www.cbs.nl/nl-nl/cijfers/detail/70936NED

CBS. (2024b). Producentenprijzen; ProdCom afzet en verbruik, index 2015=100 2012-2023 CBS. https://www.cbs.nl/nl-nl/cijfers/detail/83935NED?q=ppi

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. Translational vision science & technology, 9(2), 14-14.

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of Information Management, 60, 102383.

Copeland, B. (2023, 11-10-2023). artificial intelligence. Britannica. Retrieved 12-10-2023 from https://www.britannica.com/technology/artificial-intelligence

EIA. (2024). Crude Oil Prices: Brent - Europe Federal Reserve Economic Data (FRED). https://fred.stlouisfed.org/series/DCOILBRENTEU

Estrada, M., Camarillo, M. E. G., Parraguirre, M. E. S., Edgar, M., Castillo, G., Juárez, E. M., & Gómez, M. J. C. (2020). Evaluation of several error measures applied to the sales forecast system of chemicals supply enterprises. International Journal of Business Administration, 11(4), 39-51.

Eurostat. (2024). Gross Domestic Product for European Union (27 Countries from 2020) Eurostat. https://fred.stlouisfed.org/series/CPMNACSCAB1GQEU272020

Fildes, R., Goodwin, P., & De Baets, S. (2023). Forecast Value Added in Demand Planning. Available at SSRN 4558708.

FRED. (2024). Gross Domestic Product for China Federal Reserve Economic Data (FRED). https://fred.stlouisfed.org/series/MKTGDPCNA646NWDB

Gamberini, R., Lolli, F., Rimini, B., & Sgarbossa, F. (2010). Forecasting of sporadic demand patterns with seasonality and trend components: an empirical comparison between Holt-Winters and (S) ARIMA methods. Mathematical Problems in Engineering, 2010.

Goltsos, T. E., Syntetos, A. A., Glock, C. H., & Ioannou, G. (2022). Inventory–forecasting: Mind the gap. European Journal of Operational Research, 299(2), 397-419.

Heerkens, H., Winden, A. v., & Tjooitink, J.-W. (2017). Solving Managerial Problems Systematically (Vol. First edtion) [Book]. Noordhoff Uitgevers BV. http://ezproxy2.utwente.nl/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1491449&site=ehost-live

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. ONR Memorandum, 52.

Ibnu, C. R. M., Santoso, J., & Surendro, K. (2019). Determining the neural network topology: A review. Proceedings of the 2019 8th International Conference on Software and Computer Applications,

IMF. (2024). Global price of Natural gas, EU Federal Reserve Economic Data (FRED). https://fred.stlouisfed.org/series/PNGASEUUSDM

In, Y., & Jung, J.-Y. (2022). Simple averaging of direct and recursive forecasts via partial pooling using machine learning. International journal of forecasting, 38(4), 1386-1399.

Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(4), 531-538.

Kavzoglu, T. (1999). Determining optimum structure for artificial neural networks. Proceedings of the 25th Annual Technical Conference and Exhibition of the Remote Sensing Society,

KNMI. (2024). Daily mean temperature, Minimum temperature, and Maximum temperature in (0.1 degrees Celsius), De Bilt. <http://www.knmi.nl/kennis-en-datacentrum/achtergrond/centraal-nederland-temperatuur-cnt>

Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. International Journal of Production Economics, 225, 107597.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the selection of forecasting accuracy measures. Journal of the Operational Research Society, 73(5), 937-954.

Maddula, S. (2021, 12-10-2021). Domains of AI. https://suryamaddula.medium.com/domains-of-artificial-intelligence-8046d0778f1a

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. International Journal of Forecasting, 5-22.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: An empirical investigation. Journal of the Royal Statistical Society: Series A (General), 142(2), 97-125.

Makridakis, S., Ord, K., & Hibon, M. (2000). The M3-Competition. International journal of forecasting, 16(4), 433-436. https://doi.org/https://doi.org/10.1016/S0169-2070(00)00078-9

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. International journal of forecasting, 36(1), 54-74. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.04.014.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022a). M5 accuracy competition: Results, findings, and conclusions. International journal of forecasting, 38(4), 1346-1364.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022b). The M5 competition: Background, organization, and implementation. International journal of forecasting, 38(4), 1325-1336.

Makridakis, S., Wheelwright, S., & Hyndman, R. J. (1998). Forecasting: methods and applications. John Wiley & Sons.

Microsoft. (2023). Welcome to LightGBM's documentation! https://lightgbm.readthedocs.io/en/stable/

Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2), 13.

Ni, D., Xiao, Z., & Lim, M. K. (2020). A systematic review of the research trends of machine learning in supply chain management. International Journal of Machine Learning and Cybernetics, 11, 1463-1482.

Nikolopoulos, K. I., Babai, M. Z., & Bozos, K. (2016). Forecasting supply chain sporadic demand with nearest neighbor approaches. International Journal of Production Economics, 177, 139-148.

Perktold, J., Seabold, k., & Taylor, J. (2024). statsmodels 0.15.0 (+270). statsmodels-developers. https://www.statsmodels.org/dev/generated/statsmodels.tsa.holtwinters.ExponentialSmoothing.html

Petropoulos, F., Wang, X., & Disney, S. M. (2019). The inventory performance of forecasting methods: Evidence from the M3 competition data. International journal of forecasting, 35(1), 251-265.

scikit-learn developers. (2023a). sklearn.neighbors.KNeighborsRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html#sklearn.neighbors.KNeighborsRegressor

scikit-learn developers. (2023b). sklearn.neural_network.MLPRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. e. (2013). A survey of forecast error measures. World applied sciences journal, 24(24), 171-176.

Shen, Z., Yang, H., & Zhang, S. (2021). Neural network approximation: Three hidden layers are enough. Neural Networks, 141, 160-173.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2016). Inventory and production management in supply chains. CRC Press.

Simoncelli, E., & Daw, N. (2003). Least squares optimization. Lecture Notes, http://www. cns. nyu. edu/eero/teaching. html.

Singh, A. (2024). Calendar Power Bi. www.LearnPowerBI.com

Smith, T. G. (2023). pmdarima.arima.auto_arima. https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html

Tran, T. T. K., Lee, T., & Kim, J.-S. (2020). Increasing Neurons or Deepening Layers in Forecasting Maximum Temperature Time Series? Atmosphere, 11(10), 1072. https://www.mdpi.com/2073-4433/11/10/1072

Uzair, M., & Jamil, N. (2020). Effects of hidden layers on the efficiency of neural networks. 2020 IEEE 23rd international multitopic conference (INMIC),

Vivochem. (2024). Vivochem B.V. Retrieved 01-02-2024 from https://www.vivochem.nl/producten

Wang, Z., Di Massimo, C., Tham, M. T., & Morris, A. J. (1994). A procedure for determining the topology of multilayer feedforward neural networks. Neural Networks, 7(2), 291-300.

Wang, Z., Tham, M. T., & JULIAN MORRIS, A. (1992). Multilayer feedforward neural networks: a canonical form approximation of nonlinearity. International Journal of Control, 56(3), 655-672.

Williams, T. M. (1984). Stock control with sporadic and slow-moving demand. Journal of the Operational Research Society, 35, 939-948.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. Management science, 6(3), 324-342.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. Biometrika, 92(4), 937-950.

Zemkoho, A. (2022). A basic time series forecasting course with python. Operations Research Forum,

Appendix

Appendix A: Lagged historical data

Table A1 shows the percentile difference of different degrees of lagged data for the NN model with one neuron per input variable. A positive percentage indicates a larger average error.

Tabel A2: Number of lagged weeks and percentile difference compared to the one week of lagged data.

	Number of lagged weeks											
	12	11	10	9	8	7	6	5	4	3	2	1
Average MAE	33%	35%	26%	19%	19%	17%	14%	13%	10%	8%	4%	-
Average MSE	145%	384%	82%	59%	55%	49%	42%	39%	32%	29%	15%	-
Average SMAPE	13%	13%	13%	12%	12%	11%	10%	9%	7%	4%	1%	-
Average bias	63%	170%	15%	48%	103%	30%	58%	34%	48%	22%	-2%	-

Appendix B: Selected Endogenous Variables

Table A2 shows the top 5% weighted input variables from the endogenous data set.

Tabel A2: Input variables selected from the endogenous data set.

Input variable name
BASE ITEM 129 Moving Average
BASE ITEM 2 Error
BASE ITEM 121
BASE ITEM 38 Moving Average
BASE ITEM 138 Moving Average
BASE ITEM 131 Error
BASE ITEM 98 Moving Average
BASE ITEM 65 Moving Average
BASE ITEM 12 Moving Average
BASE ITEM 132 Error
BASE ITEM 7
BASE ITEM 38 Error
BASE ITEM 20 Error
BASE ITEM 98
BASE ITEM 60 Moving Average
BASE ITEM 130 Moving Average
BASE ITEM 90 Moving Average
BASE ITEM 130 Error
BASE ITEM 138 Error
BASE ITEM 121 Moving Average
BASE ITEM 128 Error
BASE ITEM 91 Moving Average
BASE ITEM 138
BASE ITEM 50 Error
BASE ITEM 137 Moving Average
BASE ITEM 121 Zero
BASE ITEM 60 Error
BASE ITEM 121 Error
BASE ITEM 28 Error
BASE ITEM 12
BASE ITEM 7 Moving Average
BASE ITEM 137 Error
BASE ITEM 106 Moving Average
BASE ITEM 91
BASE ITEM 28
BASE ITEM 50
BASE ITEM 121 Trough
BASE ITEM 137
BASE ITEM 28 Moving Average
BASE ITEM 50 Moving Average
BASE ITEM 60
BASE ITEM 12 Error