LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT

School of Engineering Science

Master of Science in Technology – Business Analytics

Sergey Zakrytnoy

COMPARATIVE STUDY OF CLASSIC AND FUZZY TIME SERIES MODELS FOR DIRECT MATERIALS DEMAND FORECASTING

Examiners: Pasi Luukka, Professor

Jan Stoklasa, Post. Doc

Supervisor: Sammeli Sammalkorpi, CEO Sievo

Abstract

LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Engineering Science
Master of Science in Technology – Business Analytics

Sergey Zakrytnoy

Comparative study of classic and fuzzy time series models for direct materials demand forecasting

Master's thesis
2021
71 pages, 21 figures, 21 tables
Examiners: Professor Pasi Luukka, Post. Doc Jan Stoklasa

Keywords: time series forecasting, direct materials budgeting, demand forecasting, supply chain forecasting, fuzzy time series, SARIMA, Holt-Winters

In many industries, direct materials budgeting is an essential part of financial planning processes. In practice, it implies predicting quantities and prices of dozens and hundreds of thousands of different materials that will be purchased by an industrial enterprise in the upcoming fiscal period. Lack of collaborative processes over the length of the supply chain, distortion effects in demand projections and overall uncertainty cause the enterprises to rely on internal data to build their budgets.

This research addresses the need for a scalable solution that would use mathematical models to reveal intrinsic patterns in historical purchase quantities of direct materials and generate automatic forecast suggestions. Business context and limitations are explored, and relevant time series forecasting methods are shortlisted based on existing practice described in academic research. Furthermore, anonymized datasets of direct materials purchases from three industry partners are used to evaluate predictive performance of the shortlisted methods. Quantitative part of the study reports an improvement in prediction accuracy of up to 47% compared to the currently used naïve approach, with fuzzy time series models being most appropriate for the intermittent time series in question.

By means of a comparative study, the research demonstrates that it is feasible to apply univariate models in direct materials budgeting processes, and suggests further topics such as implementation complexity that need to be explored prior to taking those models into use.

Acknowledgement

I would like to thank my supervisors, Pasi Luukka and Jan Stoklasa, for the world-class guidance on the way to completing this research.

I also want to express gratitude to my colleagues at Sievo – for giving an opportunity to work on such an exciting and challenging topic, and our valued customers – for granting permission to use their data, which made the research possible in the first place.

Finally, but equally important, I want to thank my family and friends from LUT for providing valuable feedback and everlasting motivational support.

Contents

List of Abbreviations

ACF	Autocorrelation function
ADF	Augmented Dickey-Fuller (test)
AI	Artificial intelligence
AIC	Akaike information criterion
ANFIS	Adaptive neuro fuzzy inference system
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
BI	Business intelligence
CFAR	Collaborative forecasting and replenishment
COGS	Cost of goods sold
CPFR	Collaborative planning, forecasting and replenishment
DF	Dickey-Fuller (test)
ERP	Enterprise Resource Planning
FLR	Fuzzy logical relationship
FLRG	Fuzzy logical relationship group
FTS	Fuzzy time series
GSI	Group seasonal index
HOFTS	High-order fuzzy time series
HW	Holt-Winters
INARMA	Integer autoregressive moving average
ISI	Individual seasonal index
KMO-MSA	Kaiser-Meyer-Olkin measure of sampling adequacy
LHS	Left-hand side (of a fuzzy logical relationship)
MAE	Mean average error
MF	Material Forecasting (Sievo solution)
MRP	Material Requirements Planning
MSE	Mean squared error
PACF	Partial autocorrelation function
PWFTS	Probabilistic weighted fuzzy time series
PWHOFTS	Probabilistic weighted high-order fuzzy time series
RHS	Right-hand side (of a fuzzy logical relationship)

RMSE	Root mean squared error
RNN	Recurrent neural network
S2P	Source-to-Pay
SARIMA	Seasonal autoregressive integrated moving average
SCF	Supply chain forecasting
SKU	Stock-keeping unit
SVM	Support vector machine
UOM	Unit of measurement
VMI	Vendor managed inventory
WFTS	Weighted fuzzy time series
WHOFTS	Weighted high-order fuzzy time series

1 Introduction

1.1 Background

This Master thesis research addresses the challenge presented by Sievo Oy, a Finnish SaaS company that provides data-driven Procurement Analytics solution. The value proposition of the company consists of raw data extraction from a variety of corporate information systems, including Enterprise Resource Planning (ERP), Material Requirements Planning (MRP), Source-To-Pay (S2P) and others; data cleansing that includes collaborative classification of spend to the standardized category hierarchy, normalization of suppliers using external enrichments to identify parent-child relationships, automatic translations and currency conversions; and a business intelligence (BI) application that combines visibility dashboards and advanced AI-driven opportunity identification features. Sievo Procurement Analytics ecosystem consists of Spend Analysis, Savings Lifecycle, Contract Management, Procurement Benchmarking and Material Forecasting (MF) solution areas.

This research is aimed at improvement of Material Forecasting solution. Material Forecasting, as currently offered by Sievo, is a highly customizable cloud-native tool that allows for tracking budget goals and maintaining quantities and prices outlooks for direct material purchases, combining data from multiple source systems. In industry, direct materials are defined as items used in the production of end-products, such as raw materials or packaging. In an average manufacturing company, based on Sievo industry experience, the proportion of direct material purchases in the total cost of goods sold (COGS) amounts on average to 80%. With many materials being subject to market price volatility, direct material purchases impose high risk to gross profit margins, thus urging the enterprises to manage the outlooks proactively in order to have better visibility over future profitability.

The two key elements for direct material forecasting are the expected prices and quantities. The latter tend to come directly from planning processes, MRP systems – if in use. On the other hand, expected future prices could be gathered from procurement experts responsible for specific purchasing categories. Finally, financial department is the key stakeholder in managing the forecasts and estimating the impact on profitability.

More specifically, monthly MF process implies having above-mentioned values on a material and stock-keeping unit (SKU) level, separately for each plant or production unit, and, sometimes, supplier. In terms of forecast horizon, the outlook is typically required for

the next fiscal year or quarter, depending on the budget round setup in the finance department of the organization. In reality, we see that complex organizational structure, multitude of ERP and MRP systems and limited data quality in those, on the one hand, and different units of measurement (UOM), purchasing currencies and dispersed procurement knowledge, on the other, does not allow for automatic collection of a consistent dataset of price and quantities series for MF purposes.

With these limitations, Sievo as a solution provider leverages its expertise in data extraction and cleansing, completing its vision of a Procurement Information Hub. Alongside with the data coming from customer's MRP, it feeds information from previous forecasts and historical purchasing data into the forecasting engine (Fig. 1), providing initial setup for manual entry and adjustments. Once the manual entry is finalized, the consolidated outlook on material, plant, supplier level, is visualized in an interactive reporting environment.

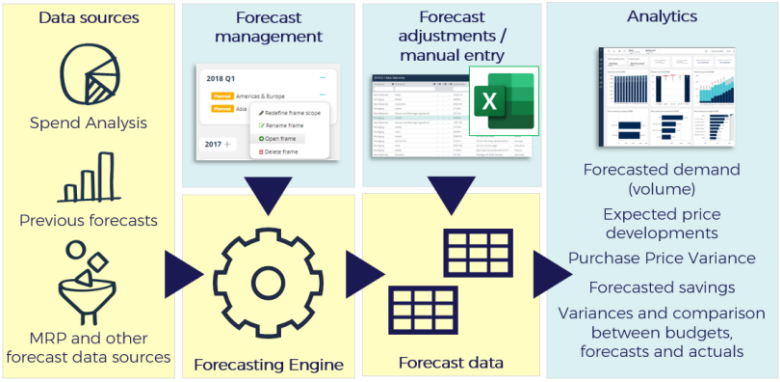

Figure 1. Data flow in Sievo Material Forecasting solution

The focus of this research is to explore and validate opportunities to enhance the workflow with predictive models. It is worth mentioning that manual entry remains an important channel of correcting forecasted values. While high-volatility or high-profit impact materials may remain subject to manual review by category experts, it is expected that for non-critical materials that would likely comprise the long-tail of purchases, predicted values can be accepted without mobilizing additional human resource. Furthermore, the ability to forecast internal demand for material items based entirely on historical spend data will enable the use of MF solution area without building integration processes with client's MRP, which presents a strategic advantage in the field.

The quantitative part of this research includes suitable predictive models from classic and fuzzy econometrics domains to support digital transformation of budgeting function in large

industrial enterprises. The business context, observed data limitations, computational requirements and the need for scalability are carefully considered when drawing conclusions and recommendation.

1.2 Research questions and limitations

The core research questions of the thesis and the underlying subtasks are described below.

1. **What are the suitable algorithms to tackle the material forecasting problem?**

To address this research question, we use existing knowledge to describe a longlist of applicable methods, and put those into context of existing data limitations. The initial selection of methods is based on previous research around resource planning automation, statistical inference for demand forecasting and time series forecasting models in general. The subtasks that need to be completed include:

   a. review previous work related to supply chain forecasting;
   b. identify available data points and perform data extraction;
   c. evaluate data quality and prototype model implementation.

2. **With selected methods of time series forecasting, is it possible to outperform the baseline (naïve) approach, currently used in Sievo solution?**

The second research question of the thesis refers directly to its quantitative part. It is essential to not only implement applicable methods for direct materials demand projection, but also provide a meaningful comparison to status-quo using available industry references. For the comprehensive evaluation, we undertake to complete the following subtasks:

   a. design the experiment and metrics for comparability;
   b. test the proposed methods on extracted datasets.

Working on real data enables us to report results that are easy to interpret and are highly relevant for business decision support. At the same time, it introduces a number of limitations to the research. Lack of data points that would represent external factors of direct materials demand is a crucial constraint that affects models selection, essentially limiting those to the univariate kind. In addition to this, the dataset in use is subject to outliers, corrupted values, and other types of noise.

The time series containing purchasing information of different direct materials are various by their characteristics: time period, intermittency, recency, which calls for appropriate delimitation of the quantitative research. Data pre-processing and filtering based on logical rules is used to scope the model evaluation process.

1.3 Structure of the thesis

The text of this thesis is structured as follows. In Chapter 2, we present theoretical background for the research, including related work in supply chain forecasting domain as well as a high-level overview of econometric and machine learning methods tested previously for a range of demand forecasting research topics. Chapter 3 contains a formal description of data exploration, dimensionality reduction and time series forecasting methods shortlisted as applicable to the quantitative part of the research.

Chapter 4 opens the empirical part of the thesis with exploration of the analyzed datasets. Specifically, eligible time series are picked and transformed for testing purposes, and exploratory data analysis is performed. Based on the knowledge obtained from early stages of the analysis, time series forecasting experiments are designed and listed in Chapter 5, and results are reported in Chapter 6. Finally, conclusions are drawn and recommendation is given with regards to applicability of the analyzed methods in Sievo MF solution.

2 Related work

In this chapter, we introduce the definition of supply chain and positioning of supply chain forecasting (SCF) in the operational landscape of a modern enterprise. Unless specified otherwise, the study of knowledgebase in SCF domain is based on the comprehensive invited review (Syntetos et al., 2016).

In further sections, we move on to an overview of statistical and machine learning techniques that have been applied to SCF in the past. Considering the objectives and limitations of the present research, we critically evaluate applicability of the mentioned techniques to the domain of direct material forecasting, and conclude the literature review with a list of identified research gaps as well as key points to consider in method selection and experiment design phases.

2.1 Supply chain definition and physiology

In broad terms, a supply chain encompasses all decision-making units involved in fulfilling a customer demand for a certain commodity (Copra & Meindl, 2012). Within any supply chain, it is common to distinguish flows of goods, services, information and money; different elements of the chain can be addressed separately depending on the purpose of the analysis. In its length, a complete supply chain would stretch from raw material suppliers, through wholesalers and retailers to the final customer. Edge-node demand for final goods and services is propagated throughout the chain with a series of sequential purchase requisitions and information inference as shown in Fig. 2.

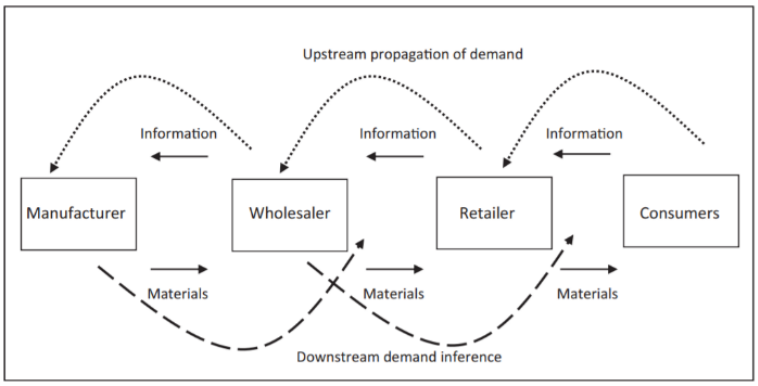

Figure 2. Demand inference. [Syntetos et al., 2016, p. 3]

Modern digitalization technologies introduce possibilities for shorter delivery times than previously, which adds up to the factors of market competitiveness. With consumers of final goods and services being the starting point of any supply chain, elevated expectations urge

the suppliers to introduce intelligent forecasting models to the conventional manufacturing processes, traditionally prone to lags resulting in the cycles of overproduction.

Scholars (Chopra & Meindl, 2012) state that the objective of any supply chain is maximization of overall value generated, where value is defined as the difference between sales revenue and total incurred costs throughout the chain of involved decision-making units. The resulting optimization problem would be a trivial part of enterprise operational analysis given complete information; otherwise, total costs of filling the demand at each node of the supply chain is an increasing function of the uncertainty associated with the upstream of information. This brings us to the reality where the choice and implementation of relevant SCF methods becomes a key factor to the performance of economic agents in a competitive landscape.

In order to find the suitable approach to SCF, we decompose the physiology of a supply chain into dimensions: length, depth and time. We then describe various phenomena, concepts and research gaps that can be attributed to each one of these perspectives. Definition of those in mathematical context is of utmost importance, as it dictates the selection of methods and preceding data processing steps.

2.1.1 Length dimension. Bullwhip effect

One property of a supply chain that characterizes the number of parties involved in the value generation is its length. Naturally, the length of a supply chain is an increasing function of the complexity of the final product or service.

When considering upstream propagation of demand in SCF process, i.e. how information about demand for final product is transition through the nodes over the length of a supply chain, it is important to recognize the complexity of a supply chain in question and the variety of factors – both internal and external – that can influence or distort the projections. The amplification of demand variance that takes place as the value proceeds through the chain nodes was defined as “Bullwhip Effect” (Lee et al., 1997). In the original work, the four root causes of the effect were given as demand signal processing, rationing and shortage gaming, batch ordering and price fluctuations. These can be summarized as operational inefficiencies and external factors that affect the deviation between expected and realized demand quantities.

Incomplete information and its increasing distortion effect on demand projections and, consecutively, on operational efficiency of manufacturers in the value chain, leave room for collaborative concepts that imply sharing information for overall value maximization (Chopra & Meindl, 2012). Collaborative forecasting and replenishment (CFAR) concept suggests interchange of decision-support models and strategies to facilitate forecasting processes (Raghunathan, 1999). Since then, a range of similar concepts have emerged both as research items and marketed digital solutions. Those include Collaborative planning, forecasting and replenishment (CPFR), Vendor managed inventory (VMI) information systems (Syntetos et al., 2016).

It has been recognized that long-term collaborative efforts are often hindered by the non-transparency that is normally attributed to strategic activities of commercial organizations. First listed in the work (Premkumar, 2000), success criteria that need to be fulfilled for such collaboration to thrive include aligned and non-competing business interests, competent team engaged in the joint project, transparent performance indicators, incentive systems and others. Some scholars (Davis and Spekman, 2004) state that these conditions have rarely been addressed as part of pilot implementations which may explain substandard outcomes. The latter have been shown and analyzed with statistical methods by several groups of researchers, concluding that many of the attempts to introduce collaborative supply chain forecasting systems actually yielded negative dynamics in the performance, widening the bullwhip effect and otherwise burdening the procurement function (Thonemann, 2002; Heikkilä, 2002). Finally, digital technologies and near real-time analytical platforms highlight the benefits of an agile procurement landscape and strengthens the position against long-term commitments (Vakharia, 2002; Yusuf et al., 2004).

In the absence of proper collaborative mechanisms for SCF, it is becoming increasingly relevant to understand the options that modern business has for autonomous forecasting of demand. Moreover, the importance of decision-support system gets higher as we move towards the upstream end of the supply chain, i.e. as more parties get involved in the process (Carbonneau et al., 2008).

When it comes to the length dimension in present research, given transactional dataset from the industry partners, we have full visibility to the first-tier suppliers of different stock-keeping units (SKUs), i.e. we have the information on the quantities, prices and terms of each separate purchase that created financial liability in the source system; however, there is

no extended view on adjacent nodes of the supply chain, which is considered in the selection of feasible forecasting methods.

2.1.2 Depth dimension. Cross-sectional aggregation

Depth of a supply chain is a dimension that describes the complexity of organizational structure within procurement and distribution functions. It represents different levels of aggregation that are available for historical data analysis and supply forecasting activities. In broad terms, SCF can be viewed as a hierarchical concept that is designed to provide information to various stakeholders, from category buyers to executive management (Syntetos et al., 2009).

Key element that needs to be defined in preparation of SCF execution or analysis is the target level of aggregation – e.g. whether forecasts will be used for the organization-wide overview, in a specific location or with regards to a particular subset of suppliers or product groups. This definition is bound by the availability of the data and/or operational processes in place; i.e. higher granularity requires corresponding level of detail in the historical data, while different options for aggregation are related to the availability of respective fields as dimensions in the original dataset.

While forecasting output is subject to variability depending on the level of decision-making hierarchy, the forecasting input is commonly driven by existing data structures (Syntetos et al., 2016). This research is based on anonymized historical purchasing data from a number of industry partners which provides transaction line level granularity with various dimensions that include SKU (material, item), material group, GL account, cost center, plant (location), legal entity, vendor and others. We therefore possess a holistic view on the purchases that allow for all kinds of cross-sectional aggregation. Sievo MF best practice configuration offers forecast adjustments on a supplier-SKU-plant dimension level, thus providing both the aggregated view and the possibility to drill-down to a particular plant, material or supplier.

In quantitative research aimed at evaluation of different forecasting methods, the dimensionality of cross-sectional aggregation needs to be defined so as to maximize its pattern recognition potential. When it comes to industrial time series, seasonality is an important aspect (Hyndman & Kostenko, 2007) that contextualizes the trade-off between aggregation level and sample size. It is likely to have more lengthy demand quantity time

series when aggregated from a number of individual materials or items by e.g. product group or location. Depending on the business context of the supply chain in focus, geographical aggregation can also enhance seasonal patterns (e.g. specific tourist destinations, agricultural zones); same holds true for certain groups of products, particularly evident for consumer goods (e.g. ice cream).

Individual seasonal indices (ISI) or group seasonal indices (GSI) are derived depending on whether aggregation of any kind has been applied to the original items. Different methods include estimation of seasonal component directly from aggregated series (Wirthycombe, 1989), linear combination of individual seasonal indices derived from each item (Dalhart, 1974), or adjusted variations of the above (Chen & Boylan, 2007).

It was shown by researchers (Ouwehand et al., 2005) that, while holding the potential to lengthen and enrich the original time series, aggregation of those yields very different performance depending on the grouping mechanism. Master data attributes that come from the source ERP systems need to be statistically validated. As an alternative to the latter, an unsupervised approach has been recently introduced (Boylan et al., 2014). It was demonstrated that K-means clustering provided a viable alternative to the product groupings available in the analyzed data. In chapter 4 of this research, we evaluate both the possibility to use original product dimensions as grouping factor and algorithmic approaches for dimensionality reduction.

2.1.3 Time dimension. Temporal aggregation

In absence of collaborative mechanisms, temporal patterns remain the key driving factor for automatic predictions. It is therefore important to determine the level of granularity that would, on the one hand, comply with the available dataset and, on the other hand, open possibilities to reveal intrinsic patterns.

It is not uncommon that demand time series are intermittent; with degree of intermittency increasing alongside with the level of granularity. Selecting an appropriate forecasting model makes it an essential pre-processing step to distinguish between periods of more saturate, continuous data series and those with presence of zero observations. The research gap in this domain was characterized as “urgent” (Gardner, 2011).

Temporal aggregation, a process of aggregating original time series to lower frequencies, is one possible solution to the intermittency issue (Syntetos, 2014). The two main types of

temporal aggregation are non-overlapping and overlapping, the latter being a sliding window moving average without loss of observations. Non-overlapping temporal aggregation, on the other hand, typically ends up in a significant shortening of the original series. It is therefore a trade-off between potential increase in uncertainty, stemming from loss of more granular information on demand, and intermittency of the resulting series that needs to be considered. Additionally, it was shown (Nikolopoulos et al., 2011) that having the degree of temporal aggregation aligned with the lead time in a production process resulted in a statistically significant improvement in forecasting accuracy, ceteris paribus.

Agility of supply chains and dynamical structure of the inventory portfolios often mean that the demand time series may not only be intermittent but also short in number of observations, which needs to be addressed by the appropriate data manipulation and method selection. We have mentioned cross-sectional aggregation and composition of group seasonal indices as methods designed to enhance predictive power of the dataset. Other approaches utilizing the properties of time dimension include supplementing the original demand series with values of similar or preceding, outdated versions of the same product or incorporating expert judgement to bring in additional information. The latter gets increasingly important as we extend the forecasting horizon compared to the length of available historical data. Naturally, it is more likely that a purchasing process experiences structural change or additional external factors that would affect its quantitative representation in demand series values as we look further into the future. On the other hand, short-term forecast based on sufficient amount of historical data can potentially be scaled to large number of stock-keeping units without human interaction. As shown in Fig. 3, optimal real-life applications will likely land in a combination of statistical methods and expert adjustments.

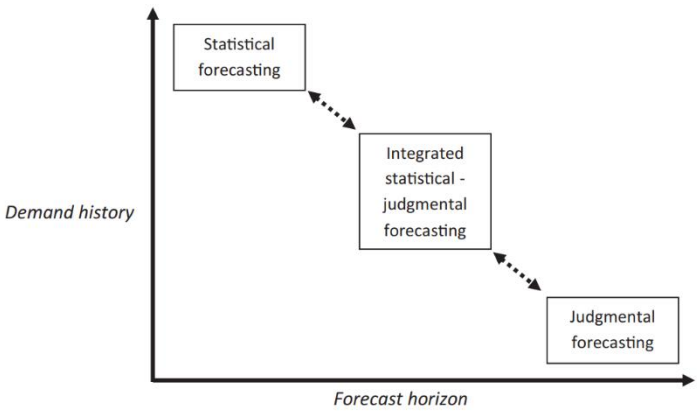

Figure 3. Types of forecasting. [Syntetos et al., 2016, p. 14]

Judgmental adjustments can take forms of simple forecast values override (currently implemented in Sievo MF), human-validated method selection or manual entry of missing or preceding observations to lengthen the learning base for statistical inference. Described as a very common practice in industry (Fildes & Goodwin, 2007), it has been demonstrated by multiple research teams that a collaborative approach involving human experts tends to improve forecasting accuracy (Fildes et al., 2009) for wider forecasting steps (significant directional corrections on monthly/quarterly level) while decreasing the quality and amplifying variance at higher frequency (weekly). Clearly, judgmental adjustment appears to be an adopted practice in the industry, lacking scientific formulation in many different ways: unaddressed topics include the effect of forecast adjustments on overall supply chain efficiency, in-depth analysis on different reasons for made adjustments etc. This research focuses on the quantitative validation of statistical methods for automated forecasting, leaving the collaborative component outside of the scope.

2.2 Statistical methods in supply chain forecasting

In this section of the literature review, we proceed to compile a list of various statistical methods that have been applied to supply chain forecasting problem in the past. Most recently available in this domain, such classes of models as feed-forward neural networks, recurrent neural networks (RNN), auto-regressive integrated moving average (ARIMA), exponential smoothing have been summarized against the selected benchmarks: naïve forecast, averages, trend forecast and multiple linear regression. The main source for this review is the research item (Carbonneau et al., 2008) and some of its references, most relevant to the business context of this thesis.

2.2.1 Benchmark methods

In a comparative study of predictive methods, there tends to be a selection of non-sophisticated techniques rendering easily interpretable results. These results are further used as benchmarks for comparative assessment. Assuming time series X_t with availability of historical observations up until X_{t-1}, the most common benchmark methods (Carbonneau et al., 2008), (Nikolopoulos et al., 2011), (Spithourakis et al., 2011) and the underlying logic for the forecasts, i.e. prediction of value X_t, are presented in Table 1.

Table 1. Benchmark forecasting methods

Method	*Formula*
Naïve	$X_t = X_{t-1}$
Average	$X_t = {\sum_{i=1}^{t-1} X_i}/{t-1}$
Moving average, with window length m	$X_t = {\sum_{i=t-m}^{t-1} X_i}/{m}$

2.2.2 ARIMA processes in supply chain

Temporal patterns within supply chain forecasting domain are often attempted to be approximated with autoregressive processes (Rostami-Tabar et al., 2013; Mohammadipour and Boylan, 2012). ARIMA is a common generalization of such models that accounts for possible non-stationarity of the original series and encompasses both autoregressive (AR) and stochastic (moving average: MA) elements. Industry-specific characteristics have been reflected in integer modifications to the ARIMA framework, translating into integer autoregressive moving average (INARMA) process (Mohammadipour and Boylan, 2012).

Previous research dedicated to identification of ARIMA processes within supply chain time series provides numerical metrics with regards to representation of different models, as well as rules and patterns that can be used to estimate optimal specifications. The research (Ali et al., 2012) showed that 30% of the analyzed SKUs followed the AR(1) process, defined as $X_t = c + \varphi_1 X_{t-1} + \varepsilon_t$. Transformation of stochastic processes through different types of temporal aggregation has been addressed separately. For non-overlapping temporal aggregation of an ARIMA (p, d, q) process, it was shown (Weiss, 1984) that the resulting series can be represented as ARIMA (p, d, r) where $r = [(p(m-1) + (d+1)(m-1) + q)/m]$, m denoting the aggregation level. Similarly, for overlapping aggregation the resulting process is ARIMA (P, d, Q) where $P \leq p$ and $Q \leq q + m - 1$ (Luiz et al., 1992).

Few studies have gone beyond accuracy-type metrics, including estimate impact of forecasting methods on overall supply chain optimization. As opposed to moving average and naïve forecasts, arguably inducing the bullwhip effect (Dejonckheere et al., 2003), ARIMA-based forecasting was shown to diminish demand signal distortion (Chandra & Grabis, 2005).

In many of the above-mentioned studies, the researchers acknowledge the data limitations and utilize the ARIMA framework to merely describe properties of the stochastic process. In the forecasting efforts, preference is given to alternative methods, exponential smoothing being the dominant one in the observed literature. In the quantitative part of the thesis, we apply ARIMA model on aggregated purchase quantity series as part of the comparison study.

2.2.3 Neural networks

Neural networks and recurring neural networks represent another class of methods commonly used in time series analysis. Neural networks are multilayer computational mechanism designed to approximate complex non-linear functions through error back-propagation and optimization (Rumelhart et al., 1986). They are known for the ability to reveal hidden patterns in the data resulting in high predictive capability. Performance of neural networks tends to follow a direct relation with the complexity of its architecture (number and order of neurons, intermediary activation functions) and amount of available data for training.

Recurrent neural networks differ from the standard feed-forward type with a specific composition of layers, as shown in Fig. 4. Within-layer feedback loop is providing additional capacity to isolate temporal patterns, dictating also a specific type of training called "back-propagation through time" (Werbos, 1990).

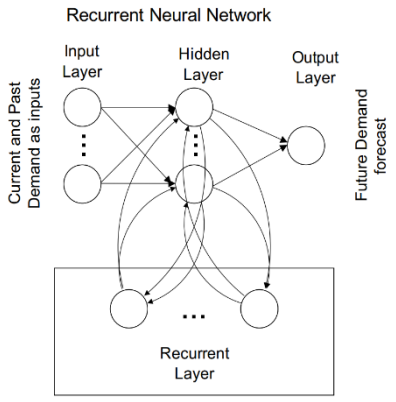

Figure 4. Recurrent neural network. [Carbonneau et al., 2008, p. 1144]

The research (Carbonneau et al., 2008) utilized a dataset of demand quantity from Canadian steel industry to test the capabilities of neural networks against benchmark methods specified above. Both recurrent and feed-forward specifications were tested, alongside with other machine learning methods such as support vector machines (SVM) and linear regression. It was shown that, while nominally outperforming the benchmark methods by mean average

error (MAE), the metrics deviated insignificantly. The gain in performance was characterized as marginal. Given the complexity of the implementation and absence of collaborative mechanisms that would contribute to the more comprehensive dataset, it was concluded that such advanced machine learning methods as neural networks did not correspond with the maturity of the business challenge that was addressed.

There is also evidence of successful implementations of machine learning methods in supply chain forecasting (Efendigil et al., 2009). They combined the efforts in neural network architecture and fuzzification of the data, resulting in an adaptive neuro fuzzy inference system (ANFIS). Different specifications of the system, including membership functions and neurons architecture, were tested. The overall conclusion claimed ANFIS to be a superior method as opposed to the regular feed-forward neural network. Later, fuzzy inference systems were utilized in a comparative study of its applications to demand series of different properties (Efendigil & Önüt, 2012).

2.3 Research gap

It is notable that in most of the research items, the selection of methods was driven by availability of the data. One of the main contributions of this thesis is to obtain a realistic performance measurement of selected methods on historical purchasing data received from partners in different industries.

In terms of the methods that have been mentioned as applicable by the community, we list multiple benchmark solutions including naïve forecast, average, simple and exponential moving average; autoregressive models ARIMA and INARMA; as well as more advanced machine learning concepts embracing artificial feed-forward neural networks, recurrent neural networks and artificial neuro fuzzy inference systems. More than two distinct approaches are rarely combined in scope of a single research, which presents another opportunity for contribution.

Data processing and feature extraction are heavily underrepresented in existing studies. Recognizing the importance of implementation complexity factor in business applications, we believe there is room for improvement with the basic methods, built on top of preprocessed data. Additionally, dimensionality reduction techniques have not been considered for large-scale forecasts.

The selection of methods and design of experiments in this thesis will consider the prior knowledge in the field while trying to bridge the identified research gaps.

3 Methods

In this chapter, we provide formal definition of the methods used in the quantitative part of the research. Starting with unsupervised techniques for exploratory data analysis and dimensionality reduction, we move on to the description of the forecasting models that have been shortlisted for the comparative analysis.

3.1 Within-group correlation measures

With a comprehensive dataset like the one utilized in this research, it is important to explore the potential of available dimensions, or attributes, to ensure the optimal selection of forecasting methods. Determining intrinsic patterns within demand quantity series of materials belonging to the same master data attribute is a key data exploration problem, having two major implications: revelation of the predictive capacity of the holistic dataset and options for dimensionality reduction. The latter is separately considered as a success criterion in final evaluation and conclusions.

3.1.1 Pearson correlation

Key concept for measurement of intrinsic patterns and similarities within groups of materials is correlation, defined as a statistical relationship between random variables. Arguably the most common variation of this measure is Pearson correlation coefficient (Pearson, 1895), which estimates the linear interdependency of two variables and is calculated for a sample as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

where x_i and y_i are values of the variables, $\bar{x}$ and $\bar{y}$ are their means.

3.1.2 Multiple correlation

Various concepts have been offered to extend the principle of Pearson correlation to analyze datasets containing three or more variables. Multiple correlation provides the framework to estimate linear correlation between a selected dependent variable and its approximation by a linear combination of the remaining independent variables. It is estimated through an auxiliary linear regression – one dependent variable against the rest of them as regressors – as the square root of coefficient of determination

$$R_{y\cdot(x_1..x_n)} = \sqrt{R^2} = \sqrt{1 - \frac{RSS}{TSS}} \quad (2)$$

where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares and $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares, $\bar{y}$ depicting the simple mean of observed y_i values (Draper & Smith, 1998).

If the model fit is better than mean, the resulting measure takes values from 0 to 1, representing scale between the potential for perfect prediction of the basis variable from the set of regressors (value of 1) and the situation when no linear combination of the regressors can render a forecast that would be more accurate than the simple mean of the observations of the target variable (value of 0; basic intuition for R^2 metrics).

3.1.3 KMO Measure of Sampling Adequacy

Kaiser's index, more commonly known in its modification called Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO-MSA), is one concept, closely linked and to a large extent based on the correlation matrix, that is designed to evaluate the fitness of the data for factor analysis or similar groupings. It can be derived from a dataset as

$$KMO = \frac{\sum_{i \neq j} {r_{ij}}^2}{\sum_{i \neq j} {r_{ij}}^2 + \sum_{i \neq j} u_{ij}} \quad (3)$$

where r_{ij} is the correlation and u_{ij} is the partial covariance between series i and j (Kaiser, 1974).

In our analysis, the measure is applicable to both evaluation of within-group correlation inside material groups and overall estimation of intrinsic patterns in the data.

3.1.4 Multirelation

Drezner (1995) introduces the multirelation concept which is designed to measure the degree of linear relation among all the vectors Y_i for $i = 1, \dots, k$ in the dataset. It is claimed to provide better representation of the interdependency within a dataset than the Kaiser's index. It measures how close a set of points in a k-dimensional space can be embedded into a $(k - 1)$-dimensional space, calculated as

$$r(x_1, x_2, \dots, x_k) = 1 - \lambda(R) \quad (4)$$

where $\lambda(R)$ is the least eigenvalue of the correlation matrix R_{xx}. The higher the calculated coefficient, the more related the vectors are in a given dataset.

3.2 Clustering techniques

Clustering is the task of dividing objects into homogeneous groups without prior knowledge on correct categorization. It therefore belongs to the unsupervised learning class of problems, which means that the data that we use do not have the target labels for different groups.

3.2.1 Clustering method types

The variety of clustering methods can be subdivided in two major categories: (1) model-based clustering methods and (2) distance-based clustering methods.

Model-based clustering methods assume that the data is a combination of groups sampled from different statistical distributions. The parameters of the distributions are unknown, and the model is trying to determine a specified number of groups with certain distribution characteristics that would explain the variance in the observations.

Distance-based clustering aims at finding such grouping of the objects in the dataset that would ensure minimal distance (~ highest similarity) between the objects within one group, while keeping maximal distance (~ lowest similarity) to the objects belonging to other groups. The latter is a fundamental principle of an efficient clustering outcome.

K-means (MacQueen, 1967) remains a very popular algorithm when it comes to clustering of large datasets. Its main advantages include simplicity of implementation and interpretation, linear time complexity w.r.t. sample size. The prerequisite to applying K-means approach is that we need to know the exact target number of clusters (commonly denoted as k). Then, the procedure iterates as follows:

1. Centroids of k clusters are initialized randomly;
2. Distance is calculated between each data point and cluster centroid (various distance metrics may be used, including Euclidean, Manhattan, Minkowski etc.);
3. Objects receive labels of their belonging to clusters based on the closest centroid;
4. Centroid coordinates are recalculated with new cluster members.

Steps 2-4 are repeated until no data point changes its label, i.e. the algorithm converges. An example of the K-means clustering technique applied to the Iris dataset (Fisher, 1936) and visualized in the dimension of the two principal components is presented in Fig. 5.

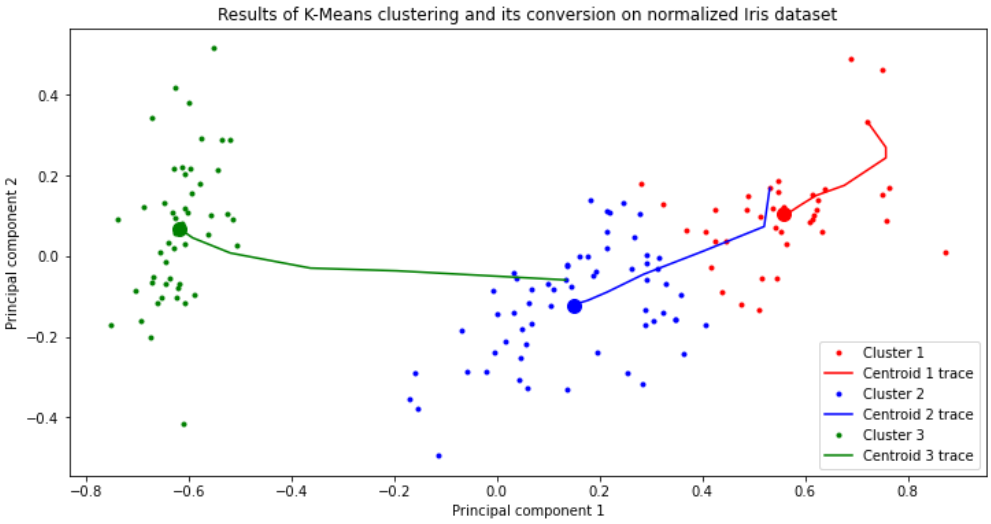

Figure 5. K-means clustering results on the normalized Iris dataset

The K-means approach, while being commonly adopted, has a number of limitations that need to be acknowledged. Those include:

- The result is dependent on initial centroid coordinates – it is better to iterate the whole process more than once;
- Sensitivity to outliers;
- Inability to handle non-linear datasets;
- Certain distance metrics will cause bias on unscaled values – features need to be normalized.

3.2.2 Elbow method for optimal number of clusters

Certain clustering methods (such as K-means, described earlier) require prior knowledge on the number of clusters to be sought. Several methods have been derived to evaluate different outcomes of the clustering processes.

The elbow method is a common visualization technique that allows to determine the optimal number of clusters. Clustering problem is repeatedly solved using predefined values of parameter $k \in \{1, 2, \dots, k_{max}\}$; within sum-of-square measure, also referred to as inertia, calculated as the sum of squared distances of each sample to its closest cluster center, or distortion, the average of the same squared distances, are calculated and stored as measures of quality for each value of k. Typically, distortion is preferred over inertia, as it removes the bias caused by different number of elements in different clusters.

The selected metrics is visualized against number of clusters in the experiment in a linear chart, of inertia showing a strictly decreasing curve. The extremes would be $k = 1$, resulting in the highest inertia/distortion measures, calculated as sum/mean of square distances from each sample to the global centroid, and $k = n$, where n is the size of the sample, resulting in 0 distance from each sample to itself representing a separate cluster. The intuition behind elbow method is to visually determine the value of k at which the improvement in the quality metrics is no longer significant, compared to previous. An example of Elbow method application to the normalized Iris dataset (Fig. 6) suggests 2 or 3 clusters as k.

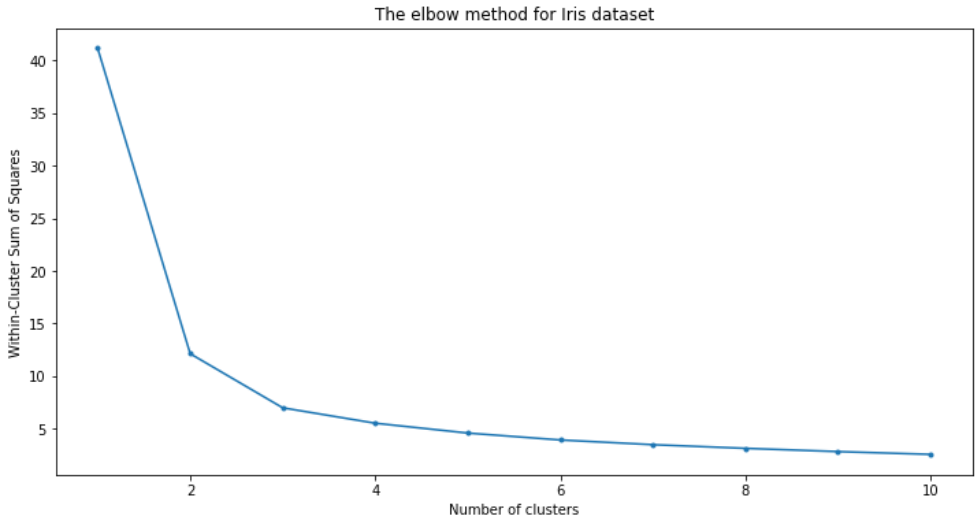

Figure 6. The elbow method for Iris dataset

3.2.3 Silhouette score

Silhouette score is an alternative quality measure of clustering outcome that does not require for human evaluation, as opposed to the Elbow method. The intuition behind the silhouette score reflects the fundamental principle of clustering process: the objects need to be as similar as possible within one cluster while remaining as different as possible from the samples outside of this cluster.

The Silhouette score is calculated for each element and averaged across the sample for overall evaluation. The calculation uses the mean intra-cluster distance and mean distance from a sample to other cluster centroids, as follows:

$$S = \frac{\sum_{i=1}^{n} \frac{b(i)-a(i)}{\max(a(i),b(i))}}{n} \quad (5)$$

where $a(i)$ is the mean intra-cluster distance, $b(i)$ is the mean distance from sample i to its second-nearest cluster.

Silhouette score values close to 1 represent highest clustering quality. Values around 0 indicate overlapping clusters, while scores below 0 mean that the sample has been clustered sub-optimally, as there is at least one cluster centroid that is closer to the sample than the cluster of its current assignment.

3.3 Dimensionality reduction

Dimensionality reduction is the transformation of data into lower-dimensional space that preserves the properties of original data, relevant in the context of a particular data analysis problem. The motivation for dimensionality reduction can be described from two perspectives:

1. Processing of data in low-dimensional space is less computationally complex, which allows for more agile model selection and makes the selected method easier to adopt in business context;
2. Interpretation of the data and its pattern recognition potential is easier to reveal via e.g. data visualization techniques.

Both specified elements of motivation hold their relevance in this research: forecasting model would require less resources should it be run on a reduced number of time series. Performed as part of exploratory data analysis, dimensionality reduction would provide the insights to intrinsic patterns which would also affect the ultimate selection process of the forecasting models.

Factor analysis is a statistical technique used to explain covariance between a set of observed variables by a set of fewer unobserved (latent) factors and their weightings. The intuition behind factor analysis is that the original features of objects in the dataset can be represented as linear combinations of latent factors, unobserved in the original data, but estimated numerically. Then, the variance in each feature would be partially explained by the estimated factors, and partially unique (specific to the original feature and error term).

Latent factors model can be represented in matrix form as

$$Y(n, v) = F(n, f) \cdot P^T(f, v) + \varepsilon(n, v) \quad (6)$$

where $Y(n, v)$ is the matrix of original observations, with n rows and v columns; $F(n, f)$ is the matrix of factors, represented by n values in f-dimensional space; $P^T(f, v)$ is the matrix of loadings; and $\varepsilon(n, v) \sim N(0, \delta^2)$ is a matrix error terms, in optimal case assumed to follow the normal distribution with variance δ^2.

Most commonly, the decomposition is achieved via maximum likelihood method, which is obtained by minimizing the fitting function, given as

$$F_{ML} = \ln\left|\tilde{\delta}\right| - \ln|\tilde{\mathrm{s}}| + \mathrm{tr}\left(\tilde{\mathrm{s}} \cdot \tilde{\delta}^{-1}\right) - p \tag{7}$$

where $\ln\left|\tilde{\delta}\right|$ is a logarithm of the determinant of variance matrix, $ln|\tilde{\mathrm{s}}|$ is the logarithm of the determinant of variance-covariance matrix of the sample, $tr\left(\tilde{\mathrm{s}} \cdot \tilde{\delta}^{-1}\right)$ is the trace of the ratio of the above mentioned matrices, and p is the number of the observed variables.

3.4 Stationarity

The following sections (3.4 – 3.5.3) dominantly reference "Forecasting: Principles and Practice" by Hyndman & Athanasopoulos (2018).

Some of the time series forecasting models require the series to fulfil the stationarity requirement. In broad terms, stationarity is the tendency of the series to preserve their statistical properties over time. Stationarity can be defined in a weak and strong form.

3.4.1 Weak stationarity

Weak form of stationarity implies that the time series hold constant mean, finite variance and constant autocovariance over time, formally defined as

$$\begin{cases} \forall t, E[x_t] = \mu \\ \forall t, E[(x_t - \mu)^2] < \infty \\ cov(x_u, x_v) = cov(x_{u+a}, x_{v+a}) \end{cases} \tag{8}$$

for any $u, v, a \in \mathbb{Z}$ where $u \neq v$.

Intuitively, it means that the series should not follow a trend but fluctuate around the mean value with constant variance. As opposed to the weakly stationary form (Fig. 7b), non-stationary series (Fig. 7a) are not expected to revert to the mean value, and their variance approaches infinity alongside with the time.

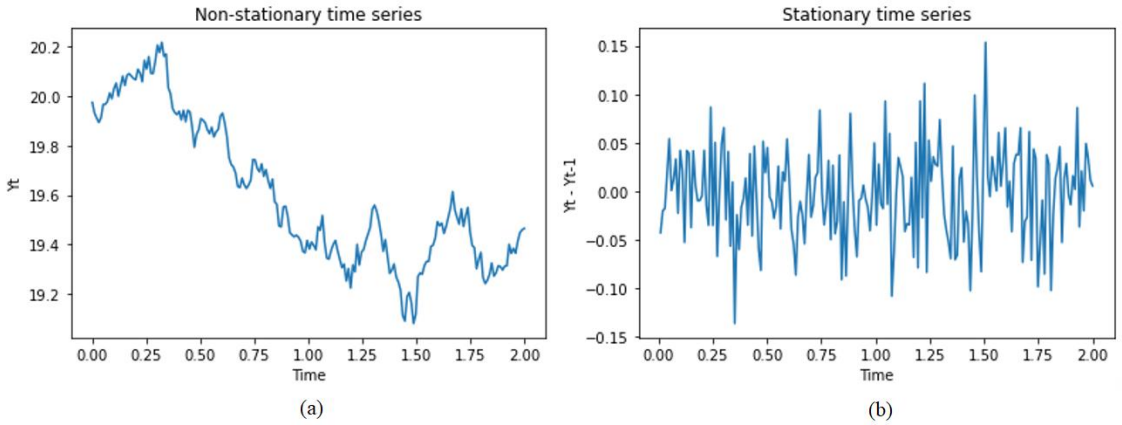

Figure 7. Simulated time series: (a) Non-stationary; (b) Differenced, weakly stationary.

3.4.2 Strong stationarity

Strong stationarity demands that the probability distribution of any sample from the time series is the same regardless of the time period or method with which the sample is drawn, formulated as

$$F_X(x_1, \dots, x_n) = F_X(x_{1+\tau}, \dots, x_{n+\tau}) \tag{9}$$

for any discrete step τ.

3.4.3 Dickey-fuller and Augmented dickey-fuller tests for unit root

One common type of time series that represent non-stationarity is the random walk process, which can be defined as a first-order autoregressive process AR(1) with unit root

$$y_t = \varphi y_{t-1} + \varepsilon_t \tag{10}$$

where $\varphi = 1$. Unit root implies that the process does not revert to its mean, i.e. it carries the shocks in (epsilon) forward with the autoregressive component. The absence of mean-reversion property does not fulfil the condition for weak stationarity, therefore a process with the unit root is considered non-stationary by definition. When subtracted y_{t-1} from both sides of the equation, we get

$$y_t - y_{t-1} = (\varphi - 1)y_{t-1} + \varepsilon_t \tag{11}$$

and see that the existence of unit root is equivalent to having zero coefficient $\varphi - 1 = 0$ in an equation of differenced time series against its lag. The significance of $(\varphi - 1)$ can be tested with t-statistics (Dickey-Fuller test, DF). The more common version, known as

Augmented Dickey-Fuller (ADF) encompasses p lags of the dependent variable, having auxiliary regression as

$$\Delta y_t = (\rho_1 - 1)y_{t-1} + \sum_{j=2}^{p} \rho_j(\Delta y_{t-j+1}) + \varepsilon_t. \quad (12)$$

We accept the null hypothesis of the existence of unit root with p-value below accepted confidence level α (t-statistics above relevant critical value for the Dickey-Fuller test).

3.4.4 Detrending and differencing

The solution to non-stationarity of time series depends on the type of their non-stationarity. The main solution to stochastically non-stationary time series, presented in a random walk process example, is differencing. Derived from (11), we see that for non-stationary time series with unit root, i.e. $\varphi = 1$, the differenced time series

$$y_t - y_{t-1} = \varepsilon_t, \varepsilon \sim N(0, \delta) \quad (13)$$

are strongly stationary given normal distribution of error term. The order of differencing, i.e. number of iterations that need to be undertaken to fulfil stationarity requirement, is determined by integration order, and obtained from repetitive ADF testing.

If the non-stationarity is observed in a form of trend, detrending is to be performed. A common approach to detrending is running an auxiliary regression of the series against their integer indices, in case of a linear trend $y_i \sim i, i = 1, \dots, n$ thus obtaining $\hat{y_i} = \alpha + \beta i$, and subtracting the predicted values from the original $y_{detrended} = y_i - \hat{y_i}$. Higher-degree polynomial trend types may call for more complex auxiliary regression specifications, such as $y_i \sim i, i^2, \dots, i^n$.

3.5 Time series forecasting

Time series is a series of numerical values indexed in time order. Usually, the observations are spaced in time at equal intervals; thus, the time dimension is discrete. Detail around time dimension, resampling and temporal aggregation are discussed in part 4.2 of this research.

The selection of methods for predicting future values (i.e. forecasting) of time series is vast. Depending on the underlying principle and data requirements, we can distinguish three main types of methods:

1. Explanatory model implies representing the dependent variable as a function of external factors (regressors), which often means that we need to determine causal

relationship in preparation for modelling. Error term accounts for unexplained variation;

2. Autoregressive time series models generate forecasts based on historical values of the focused series, excluding all possible external variables. The results from this type of models are less intuitive to interpret, but are more robust when causal relationships are not determined;
3. Mixed models contain both explanatory and dynamic components. These models are known as dynamic regressions, transfer function models, linear systems etc.

Lack of numeric data points representing external factors available in an independent enterprise supply chain forecasting predispose us to the autoregressive time series forecasting models, which we describe in more detail in the current section of the thesis.

3.5.1 Naïve benchmark

Naïve forecasting method is the basic estimation technique in which series value from last period is taken as the forecast for the next one, without attempting to adjust it or establish causal factors. Naïve method

$$y_{t+1} = y_t \tag{14}$$

is often used for comparison against more sophisticated models. Naïve method is indicatively implemented as the default forecasting mechanism for both quantity and price series in Sievo MF module, which justifies its selection as a benchmark in the quantitative part of this research.

3.5.2 Holt-Winters exponential smoothing

In this section, we cover selected examples of models from the exponential smoothing family. First proposed in 1950s, these models generate forecasts as weighted averages of previous observations, with the weights decreasing exponentially over time periods.

Holt-Winters (HW) seasonal method, also known as triple exponential smoothing, represents the kind of time series decomposition, in which the series estimation formula is split into three equations: level, trend and seasonality, bearing different smoothing coefficients, and being aggregated over the fourth, overall smoothing calculation, resulting in a system of simultaneous equations

$$\begin{cases} S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \\ b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \\ I_t = \beta \frac{y_t}{S_t} + (1-\beta)I_{t-L} \\ F_{t+m} = (S_t + mb_t)I_{t-L+m} \end{cases} \quad (15)$$

where y_t is observation of the series, S_t is the smoothed observation, b_t is the trend factor, I_t is the seasonal index, F_{t+m} is the forecast at m periods ahead; α, β and γ are smoothing parameters that are estimated so as to minimize the fitting error.

The resulting system is recursive with regards to its different components. The baseline value for trend is calculated as

$$b_0 = \frac{1}{L}\left(\frac{y_{L+1} - y_1}{L} + \frac{y_{L+2} - y_2}{L} + \cdots + \frac{y_{L+L} - y_L}{L}\right) \quad (16)$$

where L is the length of the season, y are observation series, while the initial season factor is calculated as

$$I_0 = \frac{\sum_{p=t}^{N} \frac{y_{t+pL}}{Ap}}{N} \quad (17)$$

where t is the time period, N is the number of complete seasons we have the data for, y are observation series and $A_p = \frac{\sum_{i=1}^{L} y_i}{L}, p = 1,2, \ldots, N$.

3.5.3 Seasonal Autoregressive Moving Average

Autoregressive Moving average (ARMA) family represents a univariate class of econometric models, consisting of autoregressive (AR) and stochastic (MA) components. Autoregressive component reflects the dynamic structure of the series, explaining its linear relation to the lags up to order *p*, while the moving average component is a linear combination of *q* lags of the error term. ARMA models, formulated as

$$y_t = C + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad (18)$$

where y is the estimated series, C, φ_i and θ_j are parameters to be estimated, and ε_t is an error, require the time series to fulfil weak stationarity requirements.

Seasonal autoregressive integrated moving average (SARIMA) model is an extension of traditional integrated ARMA, which activates the pattern recognition potential over seasons

by introducing a new set of parameters: orders of seasonal autoregressive component (P), seasonal integration (D) and seasonal moving average (Q) that are combined in an equation

$$y_t = C + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{k=1}^{P} \gamma_k y_{t-kL} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{r=1}^{Q} \mu_r \varepsilon_{t-rL} \qquad (19)$$

where, in addition to terms from (19), we introduce γ_k and μ_r as seasonal parameters to be estimated with the length of seasonal period L.

Selection of parameters for SARIMA model can be approached in several different ways. First, the decision on the parameters d and D (orders of simple and seasonal differencing) is to be made upon examination of the results of ADF test. The remaining parameters are orders of simple and seasonal autoregressive and moving average components.

One of the common ways to determine optimal combinations of parameters (p, q) and (P, Q) is to plot the values of autocorrelation function (ACF) and partial autocorrelation function (PACF) against number of lags. ACF quantifies the dependency of the time series on lags 1 to p, including the effects of all lags in-between. PACF represents correlation coefficient between the original series and itself with lag = p and excludes the effects of lagged series in-between.

The visual criteria for optimal model are mirrored in cases of AR and MA processes, thus guiding us to identify

- order p of AR as number of spikes in PACF with geometrically decaying ACF;
- order q of MA as number of spikes in ACF with geometrically decaying PACF.

For seasonal parameters, we should be observing similar behavior, as if the lags in-between seasonal spikes at regular interval of L periods would have no effect.

An alternative approach is to utilize one of the information criteria, Akaike information criterion (AIC) being one of the most commonly used. Calculated as

$$AIC = 2k - 2\ln(\hat{L}) \qquad (20)$$

where k is the number of parameters in a model and $\hat{L}$ is the value of log-likelihood function that the model would reproduce the observed values, it represents the loss of the model, i.e. unexplained variance in the series, and should be minimized to support the optimal combination. Additionally, the model can be tested for autocorrelation of residuals, which by definition of an autoregressive model should not be observable. The approach is more

scalable than visual identification, as it is fully logical and enables automated decision-making.

Forecasted values are obtained by shifting time indexing

$$\hat{y}_{t+1} = \hat{C} + \sum_{i=1}^{p} \hat{\varphi}_i y_{t-i+1} + \sum_{k=1}^{P} \hat{\gamma}_k y_{t-kL+1} + \sum_{j=1}^{q} \hat{\theta}_j \varepsilon_{t-j+1} + \sum_{r=1}^{Q} \hat{\mu}_r y_{t-rL+1} \qquad (21)$$

thus relying on existing observations or the ones predicted in earlier iterations.

3.5.4 Fuzzy time series

Fuzzy time series (FTS) is a concept from fuzzy data analysis domain, which is based on the fundamental concept of a fuzzy set. The latter was founded by Zadeh in 1965, and allows for a gradual membership $\mu_A(x), x \in U$ to a specified set A for every element x of a universe of discourse U, thus serving as a flexible mathematical way to model uncertainty. The degree of membership of each element $\mu_A(x) \in [0, 1]$ and is calculated from the membership function. The membership function also determines the shape of the fuzzy set, most commonly – triangular (Fig. 8a), trapezoid (Fig. 8b) and Gaussian bell curve (Fig. 8c).

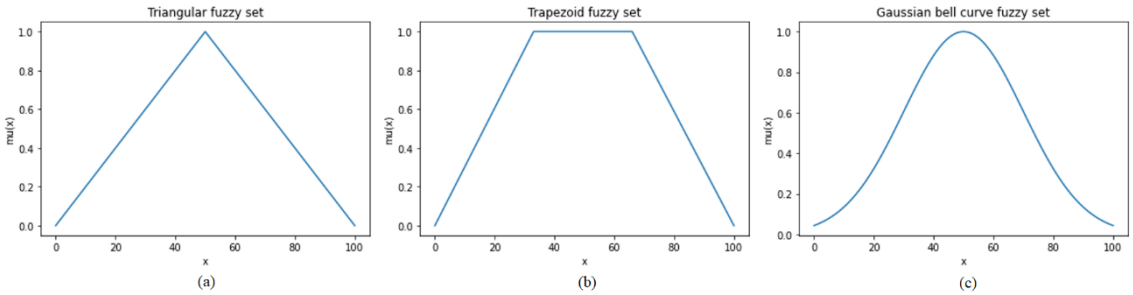

Figure 8. Examples of fuzzy set membership functions: (a) Trianguar; (b) Trapezoid; (c) Gaussian bell curve.

Fuzzy time series $F(t)$ on the subset of real numbers $Y(t)$ $(t = 0,1,2, \dots)$ implies that $F(t)$ consists of $\mu_i(t)$ $(t = 1,2, \dots)$. Real time series can be transformed into their fuzzy representation by dividing the universe of discourse (range of observed values) into equal intervals and assigning values of membership function of each original observations to the corresponding fuzzy set(s), resulting in $U = (u_1, u_2, \dots, u_m)$ where u_i are linguistic variables. Alternatively, the fuzzy sets can be obtained as a result of c-means clustering of original values, which will render non-uniform splitting of the universe of discourse.

Let's consider a simulated example of continuous time series simulated as $y_t = \sin(t) * (1 + r)$, $t \in \{0, 1, \dots, 14\}$ and $r \sim U[-0.2, 0.2]$ (Fig. 9a). The naïve partitioning of the range of observed values into 4 fuzzy sets is illustrated in Fig. 9b.

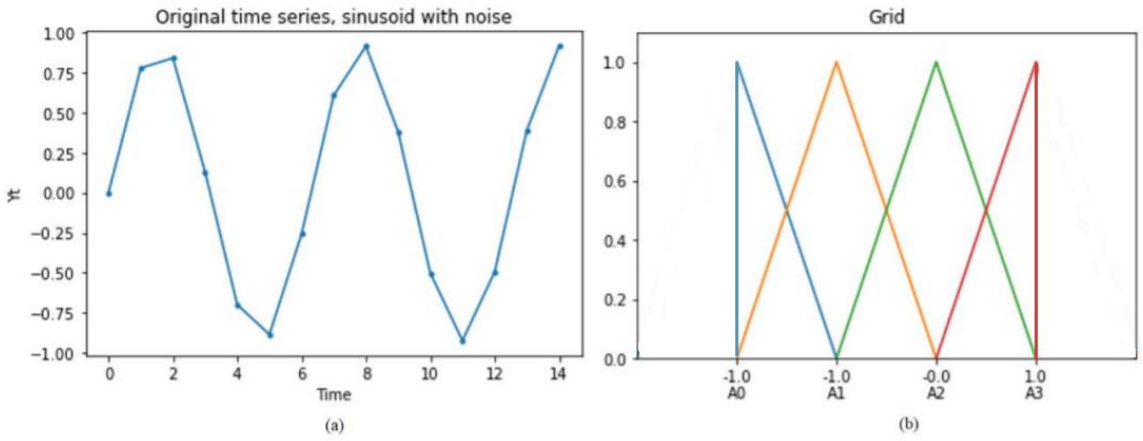

Figure 9. Partitioning of universe of discourse: (a) Simulated sinusoid series with noise; (b) Partitioning of the value range.

Transforming each value of the original time series by maximum of its membership degrees to the gridded fuzzy sets (standard procedure in non-probabilistic FTS approach), the continuous time series from example above take shape of FTS $[A_2, A_3, A_3, A_2, A_1, A_0, A_1, A_3, A_3, A_3, A_1, A_0, A_1, A_3, A_3]$.

The fuzzy time series forecasting models rely on the notion of fuzzy logical relationships (FLR). The causal relationship $R(t-1,t)$ such that $F(t) = F(t-1) \circ R(t-1,t)$ where $\circ$ is an arithmetic operator that can be denoted by $F(t-1) \rightarrow F(t)$. Since both $F(t)$ and $F(t-i)$ are represented as fuzzy numbers A_i and A_j, the logical relationship can be expressed with notation $A_i \rightarrow A_j$ (FTS model of order 1) which should be read as "if current value is A_i, the next value will be A_j" or $[A_i, A_k] \rightarrow A_j$ (high-order FTS model with 2 lags), which reads "if sequence of A_i and A_k, then the next value will be A_j". In the examples above, A_i and $[A_i, A_k]$ are called left-hand side (LHS) of an FLR, while A_j is its right-hand side (RHS).

The FLRs observed from historical data are further clustered into fuzzy logical relationship groups (FLRGs) by distinct LHS, defining the knowledge, or rule base. That rule base serves as the reference point when inferring forecasts of future observations. In the example above, the rule base consists of

$$\begin{matrix} A_0 \rightarrow A_1 \\ A_1 \rightarrow A_0, A_3 \\ A_2 \rightarrow A_1, A_3 \\ A_3 \rightarrow A_1, A_2, A_3 \end{matrix} \tag{22}$$

With conventional FTS, the forecasting procedure manages different scenarios w.r.t. the rule base in the following manner: let $F(t) = A_i$,

- if there is no relevant FLRG in the base, i.e. $A_i \rightarrow \emptyset$, then $F(t+1) = A_i$ and the defuzzified forecast $Y(t+1)$ is the midpoint of A_i;
- if the LHS A_i is uniquely represented by an FLR $A_i \rightarrow A_j$, then $F(t+1) = A_j$, $Y(t+1)$ being the midpoint of A_j;
- if for LHS A_i there are multiple FLRs $A_i \rightarrow A_{j_1}, A_{j_2}, \dots, A_{j_k}$, there is no single fuzzy representation of $F(t+1)$, but the defuzzified value is derived directly as the arithmetic average of the midpoints of $A_{j_1}, A_{j_2}, \dots, A_{j_k}$.

Weighted FTS (WFTS) implies more accurate consideration of the scenario in which $A_i \rightarrow A_{j_1}, A_{j_2}, \dots, A_{j_k}$. Designed to fix the drawback of constant importance of all RHS elements, it alters the defuzzification step in a way that

$$Y(t+1) = \sum_{j \in RHS} w_j * c_j \tag{23}$$

with

$$w_j = \frac{\#A_j}{\#RHS} \; \forall A_j \in RHS \tag{24}$$

where $\#A_j$ is the number of occurrences of A_j in FLRs with the same precedent LHS and $\#RHS$ is the total number of temporal patterns within that FLRG (Ortiz-Arroyo & Poulsen, 2018).

Probabilistic Weighted FTS (PWFTS) take a step forward to incorporate information about membership degrees of precedents, i.e. LHS of the FLRs. The knowledge base for PWFTS is given as

$$\begin{matrix} \pi_1 * A_1 \rightarrow w_{11} * A_1, \dots, w_{1k} * A_k \\ \dots \\ \pi_k * A_k \rightarrow w_{k1} * A_1, \dots, w_{kk} * A_k \end{matrix} \tag{25}$$

where each weight π_i is the normalized sum of all LHS values of membership functions where the LHS is fuzzy set A_i (Silva, 2019). Thus, π_i can be interpreted as the empirical a priori probability of having A_i as an LHS. Weight w_{ij} is the normalized sum of all RHS memberships where LHS is A_i and RHS is A_j, which can be understood as a conditional probability $P(F(t+1) = A_j | F(t) = A_i)$.

The forecasting procedure in PWFTS starts with the computation of probability distribution

$$P\big(Y(t)\big|Y(t-1)\big) = \sum_{A_j \in \tilde{A}} \frac{P\big(Y(t)\big|A_j\big) * \sum_{i=1}^{k} P\big(Y(t+1)\big|A_i, A_j\big)}{\sum_{i=1}^{k} P\big(Y(t)\big|A_i\big)} =$$

$$\sum_{A_j \in \tilde{A}} \frac{\pi_j \frac{\mu_{A_j}(Y(t))}{Z_{A_j}} * \sum_{i=1}^{k} w_{ij} \frac{\mu_{A_i}(Y(t+1))}{Z_{A_i}}}{\sum_{i=1}^{k} \pi_i \frac{\mu_{A_i}(Y(t))}{Z_{A_i}}} \tag{26}$$

where, in addition to previous notations, $\mu_A(Y)$ is degree of membership of continuous value Y to a fuzzy set A, and Z_A is the total area under membership function of A. The point forecast is then produced by

$$Y(t+1) = \sum_{A_j \in \tilde{A}} \frac{P\big(Y(t)\big|A_j\big) * E[A_j]}{\sum_{A_j \in \tilde{A}} P(Y(t)|A_j)} \tag{27}$$

where $E\big[A_j\big] = \sum_{i \in A_j^{RHS}} w_{ij} * mp_i$, mp denoting a midpoint of a fuzzy set.

FTS as a computer intelligence framework is presenting a real alternative to the traditional econometrics methods. Among other things, fuzzification of original time series makes redundant the requirement for stationarity. Reducing the allowed value domain to a finite number of fuzzy sets serves as a self-aided normalization technique that intensifies pattern recognition processes that follow.

4 Data

This chapter contains the description of the data that is taken as the basis for quantitative part of the research. The data is extracted from Sievo database, from the accounts which granted their permission to utilize anonymized historical data in the research to validate the quality of different algorithms for spend forecasting. Overall, three independent datasets are analyzed, originating from companies that operate in different industries on a global scale, hereafter referred to as companies *A*, *B* and *C*. The diversity of industry profiles enables us to compare the performance of the shortlisted forecasting methods between each other to draw conclusions with regards to potential difference in applicability of the methods to the reported cases.

As presented in Fig. 10, the first stage of data transformation and filtering takes place on transactional level in SQL Server, managed in export queries to ensure extraction of the minimum required volume of data. Further data transformation, starting with cross-sectional and temporal aggregation, is performed in iPython notebook environment which provides the flexibility of data exploration and visualization methods. The outcome of data cleansing processes described in this chapter is a collection of quantity and price datasets qualified for testing time series forecasting models.

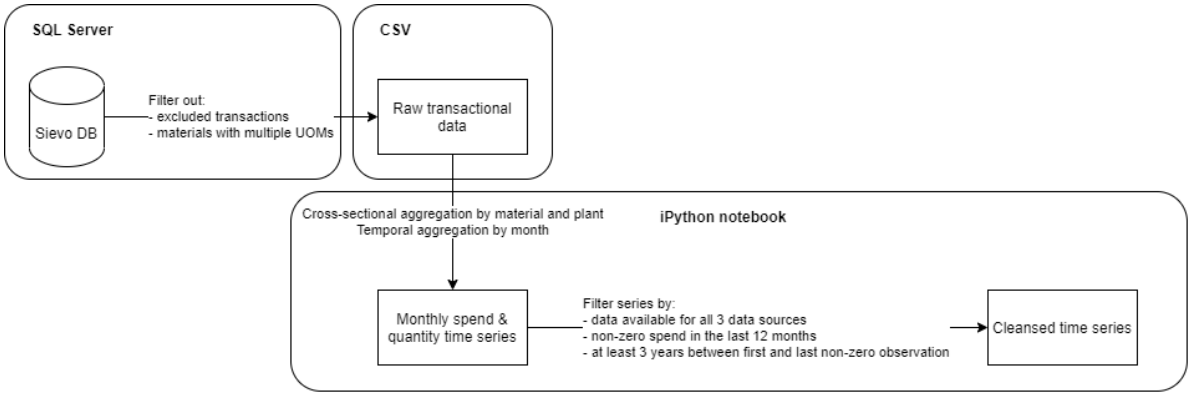

Figure 10. Overview of data filtering and cleansing processes

4.1 Source data structure

Sievo in-house developed data model is leveraged to ensure correct linkages of the data points. Most commonly, Sievo obtains the transactional and master data via automated monthly extraction from one or multiple ERP systems that are part of the customers' IT landscape. When data is extracted from modern versions of ERP systems, a so-called three-way match is performed to normalize the data (Fig. 11).

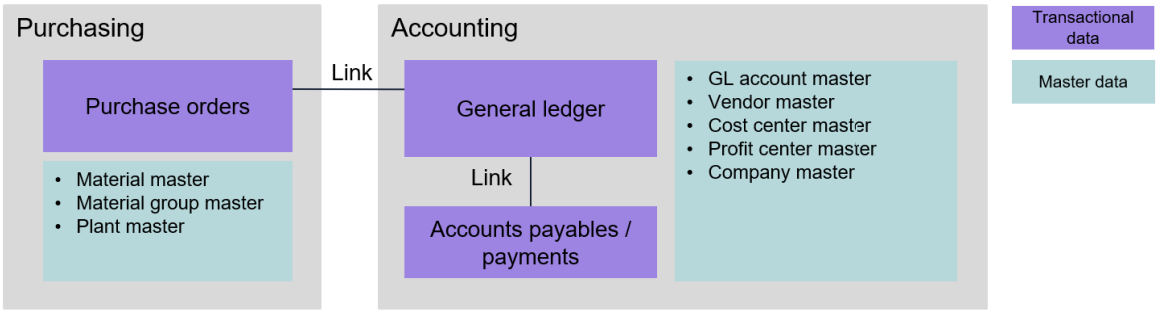

Figure 11. Sievo data model

By taking the numeric fields from accounting documents and enriching them with information from associated purchase orders and payments, Sievo secures such data points as GL account, material group, cost and profit center, etc., thus ensuring maximal identifiability of the purchased goods or services while preserving accurate accounting detail.

One of the early stages of procurement analytics process is the definition of addressable spend. Transactions are excluded from focused analysis based on combinations of one or more data points, most commonly to remove the effect of intercompany transfers, payroll, VAT and other taxes. To ensure business relevance of the results from present research, we follow the established guidelines and do not export transactions marked as excluded in Sievo.

Another filter that is applied at the early stage of data export from Sievo database is related to units of measurement. While there exist such things as “service materials”, i.e. SKUs tracked in ERP systems that represent certain services rather than products, most of the material items tend to be tangible goods, measured discretely or continuously, most commonly – by pieces, units of mass, volume or length. When aggregating the data into time series format, it is important to account for different units of measurement (UOMs) in the source. Exploratory research in the source database reveals that material purchases are mostly measured in single UOM, with small number of materials measured in 2 UOMs. To avoid parsing distinct UOMs and their conversion for each material and considering low representation of such materials in the base (0-2% of addressable spend), we simply disregard purchases of materials with more than 1 UOM. The excluded materials with more than 1 UOM are not characterized by any common parameter, and thus do not distort the results.

As shown in Table 2, the applied filters result in the scope of research narrowed down to 21-46% of total addressable spend figures, and 71-96% of direct addressable spend. At the same time, the theoretical maximum of number of time series to analyze (both in cases of material and material + plant cross-sectional aggregation) are reduced by no more than 12%, which is a justified limitation to enable smooth transition to aggregation and further cleansing steps.

Table 2. Transactional data cleansing funnel

		All →	Addressable →	With material →	With material and plant →	With single UOM
		All metrics are presented in % against the "Addressable" column				
Company A	**Spend**	**157**	**100**	**23**	**21**	**21**
	# transactions	121	100	15	15	13
	# materials	118	100	100	100	100
	# materials & plants	117	100	100	100	100
Company B	**Spend**	**278**	**100**	**48**	**48**	**46**
	# transactions	195	100	35	35	35
	# materials	129	100	100	97	97
	# materials & plants	132	100	100	97	97
Company C	**Spend**	- **89**	**100**	**41**	**29**	**29**
	# transactions	167	100	27	24	23
	# materials	137	100	100	88	88
	# materials & plants	194	100	100	88	88

The ultimate dataset used in the present analysis comprises transactional data covering addressable spend extracted from Sievo database that includes accounting document line detail on invoiced amount and quantity as well as unique identifiers of the associated SKU and the location of the purchase. Additional export is run over materials and the corresponding material groups of their belonging, to see if the quantity series of materials behave similarly within their groupings. If the source data indicate that certain material relates to more than one material group, the decision is made in a spend-weighted manner, i.e. we pick the material group that includes highest proportion of spend associated with that material.

4.2 Time period selection and cross-sectional aggregation

Time period selection is an essential step of a data extraction process. When exploring the space of feasible time selections, the three main criteria that we consider are

- availability of material-covered transactional data for all 3 data sources to enable representative comparison, i.e. report the results from the same time period;
- potential to reveal annual seasonality, i.e. at least 3 full years of data – 2 periods to capture seasonality and 1 to test the performance;
- relevance for the business, i.e. the most recent data available.

According to these conditions, a period of January 2016 – November 2020 has been selected as the most recent representation of direct material purchases of the three partner companies.

Multidimensional dataset is transformed into time series format through cross-sectional and temporal aggregation. Amount values of transactions (onwards – spend) and quantities for each material (or combination of material with another relevant data point) are aggregated through addition within predefined time intervals, thus representing total spend and quantity of purchases of each material during the periods of selection. According to industry best-practice, implemented in Sievo MF solution, the period is set to calendar month, while cross-sectional aggregation is performed on a combination of material and plant, i.e. each series represents amount of monthly purchases of different materials by separate operational units of the business. Monthly aggregation is a compromise that allows to capture temporal patterns both on quarterly and annual level, aligned with financial reporting standards, and is reasonable for the users of the final solution to introduce expert corrections as part of the monthly process (as opposed to e.g. daily granularity).

4.3 Data filtering

Before moving forward with parametrized data cleansing steps, we refer to Pareto principle, also known as "80-20" rule, to further narrow down the focus of quantitative research. As shown on Fig. 12, the dominating share of spend originates from relatively low share of total number of SKUs.

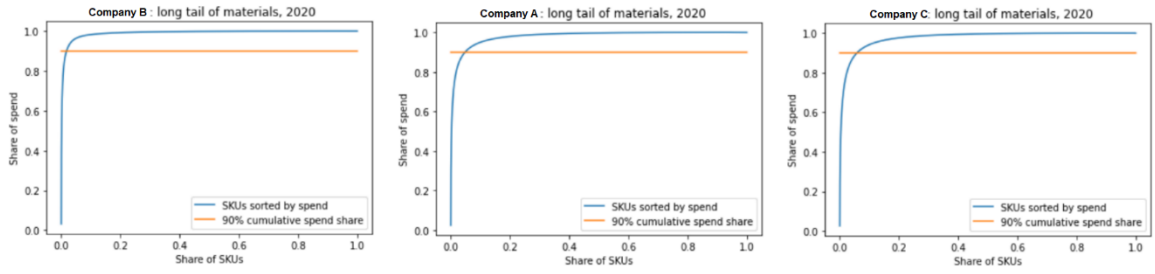

Figure 12. Long tail of SKUs and 90% spend threshold

In all 3 partner cases, the total count of unique SKUs land in the order of dozens and hundreds of thousands, with the number getting bigger when we consider unique combinations of SKUs and purchasing business unit. Such a high number of time series is not feasible to analyse with the computing power at hand, as calculations will take unreasonably long. We therefore include such combinations of SKUs and purchasing business units that contain materials from the subset of those largest ones that correspond

with 90% of spend amount. From all variety of SKUs and their combinations with purchasing business units, from 1.34% to 3.70% of series add up to 90% of cumulative spend (Table 3).

Table 3. Share of SKUs and SKU-business unit combinations adding up to 90% of spend

		Top 90% spend	All spend
Company B	SKUs	0.65%	100%
	SKU + Plant	1.34%	100%
Company A	SKUs	1.37%	100%
	SKU + Plant	1.67%	100%
Company C	SKUs	3.05%	100%
	SKU + Plant	3.70%	100%

As material resource planning process tends to take place for months in advance, it is natural to expect that supply schedule is intermittent. Even with monthly aggregation of spend and quantity values, we see all of the previously qualified series to show empty observations in all datasets.

The forecasting methods in use need to be able to capture the factor of intermittency as well as the overall trend, it is therefore reasonable to leave null observations intact. However, we need to account for possible changes in the material master data: it is not uncommon that some records are replaced by new ones, or simply become redundant. In order to meet the relevance criterion and only test the proposed methods on those material records that are being used to the day, we perform the analysis on series that have non-zero values of quantity and spend in the last 12 months of recorded period. Across the datasets, 77.31-90.41% of time series fulfil this requirement.

Another criterion for exclusion is the availability of sufficient training data for the models. This condition is met by removing the series for which the time between the earliest and the most recent observations is under 3 years. Depending on the data source, only 18.30-53.96% of the time series have enough observations between first and last non-zero month.

When the subsets of qualified time series are united, we see 17.99-50.99% acceptance rate, i.e. this share of all SKU-business unit combinations qualified from previous filtering processes both have been actively used in the last 12 months, and have enough training data for the purpose (Table 4).

Table 4. Qualification of time series by business relevance and sufficiency of training data

		Spend during last 12 months	Enough training data	Qualified
Company A	Share of spend, %	99.08	35.08	34.82
	Share of series, %	90.41	18.30	17.99
Company B	Share of spend, %	91.87	86.16	82.14
	Share of series, %	77.31	53.96	50.99
Company C	Share of spend, %	98.59	55.32	55.32
	Share of series, %	89.78	31.46	31.45

Low share of time series with more than 3 calendar years of observation may be explained by modernization of material master data in source ERP systems or relative recency of client environment in Sievo solution. Multistep process of data filtering and cleansing results in thousands of time series to be used in the evaluation of forecasting models. While the introduced conditions are justified in mathematical modelling context, the business need is equally present in cases of material purchases with and without sufficient amount of historical data. The impact of these limitations will be considered in the conclusion of this research.

4.4 Outlier detection

Outlier detection is an essential step of exploratory data analysis. Outlier is an observation in the dataset that deviates significantly from the mean, median or most commonly observed value. It is important to distinguish the outliers that indicate data errors from those that correctly depict the business case.

Outliers are unlikely to appear in the spend datasets, which are based on accounting document entries, as those would create excessive financial liability. On the other hand, information on quantities of different material purchases is fetched from purchase orders, which are often filled in manually and are subject to human error. Following a common convention, we use statistical measures of median and standard deviation and call value y_t from series Y an outlier if

$$|y_t - median(Y)| \geq X * \hat{\sigma}(Y) \tag{28}$$

where $median(Y)$ is the median value of observed series of length N, and $\hat{\sigma}(Y) = \sqrt{\frac{\sum(y_t - \bar{Y})^2}{N}}$ is its standard deviation. The selection of median over mean value in the formula is explained by the motivation to exclude the bias of outlier observation, which may deviate from "normal" ranges manifold. Values $X = \{2, 3\}$ are common in the analytical practice,

representing approximately 95% and 99% of observed values drawn from normal distribution. However, we cannot expect the values of purchased quantities to bear the properties of a normal distribution, and therefore run the outlier detection procedure over values of X from 1 to X_{max}, where X_{max} is the first natural number resulting in zero share of time series with non-zero number of outlier observations. The results of the procedure are presented in Fig. 13.

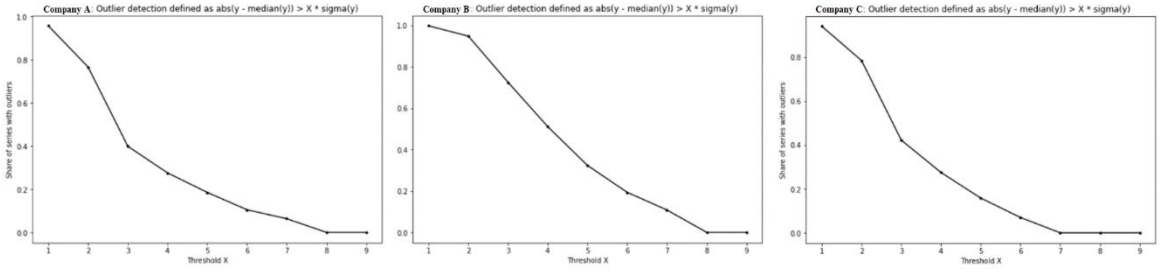

Figure 13. Share of series with identified outliers

Up until values of $X \geq 7$, we see more than 10% share of the time series that contain outliers. Without having more grounds to distinguish between data errors and actual sourcing patterns, we leave all values as observed in the original extract. Potential presence of outliers is considered separately in the evaluation of forecasting methods.

4.5 Data normalization

Data normalization is a transformation of original numerical values, commonly linear, into a new representation with controlled statistical properties such as mean, standard deviation, minimal and maximal values. Quantity series are normalized using min-max normalization technique (eq. 29), resulting in series scaled to range [0, 1] to enable comparison of similarity and forecasting accuracy metrics.

$$y_{norm} = \frac{y_t - \min(Y)}{\max(Y) - \min(Y)} \tag{29}$$

4.6 Master data grouping evaluation

Apart from the data points primarily used in temporal and cross-sectional aggregation, i.e. posting date, material and plant numbers, data export includes information about the hierarchical material-material group relationship. In case of large number of material-plant level time series, we utilize a number of methods described in section 2.1 of this thesis to examine the degree of similarity between material series belonging to the same material

group. High similarity of patterns within those will enable additional level of aggregation which can be utilized as the basis for the final forecast decisions.

The first stage of the evaluation relies on the Pearson correlation metrics; for each material group that contains more than 1 qualified SKU + business unit combinations, pairwise correlation coefficients are averaged and described across all relevant material groups (Table 5). The ultimate metrics is derived as weighted average of within-material group correlation coefficients, with spend under material group as weighting factor. Resulting values in range between 0.099 and 0.254 bring us to the conclusion that master data groupings are of limited value when it comes to spend forecasting modelling.

Table 5. Descriptive statistics of correlation between time series within same Material groups

Correlation	Company A	Company B	Company C
Mean	0.336	0.232	0.208
Standard deviation	0.227	0.246	0.184
Min	0.060	0.020	0.020
25%	0.174	0.077	0.078
50%	0.272	0.125	0.143
75%	0.462	0.322	0.289
Max	1.000	0.972	0.951
Weighted average (by spend)	**0.254**	**0.099**	**0.172**

4.7 Clustering

Clustering is an advanced stage of exploratory data analysis that is performed in an unsupervised manner, i.e. not relying on any prior information, with a view to introduce meaningful grouping of original time series. K-means clustering algorithm is applied to normalized quantity series to detect similar patterns within resulting groups of SKU + business unit combinations.

A key element to K-means clustering is identification of optimal number of clusters, which needs to be specified prior to running the algorithm. Elbow method, a common visual way to identify significant marginal improvement in clustering quality when increasing number of clusters, reveals that $k \approx 100$ is worth further evaluation (Fig. 14).

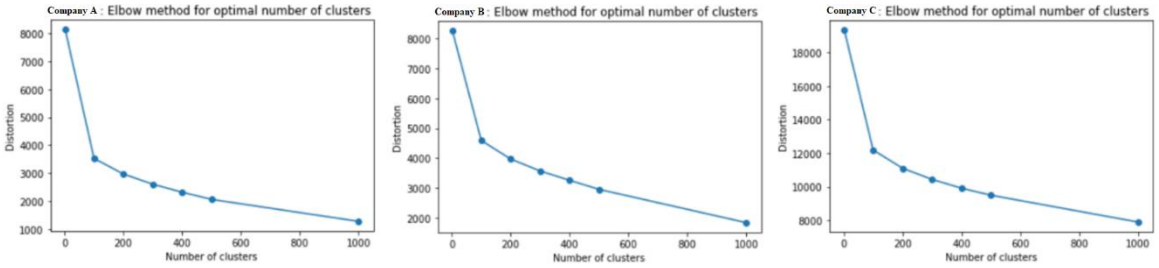

Figure 14. Elbow method for optimal number of clusters

Silhouette score is a core quantitative metrics used to evaluate the quality of clustering results. It combines measurements of how similar objects are within clusters and how different they are across those. Values range between 0 and 1, showing higher performance of clustering as the score increases.

We see low values of Silhouette score for the tested cases (Table 6) which does not confirm the feasibility of clustering with $k \approx 100$ or any other value from the tested range. Highest scores are derived with $k = 2$, which means partitioning of the dataset into 2 parts and does not add value for dimensionality reduction or intrinsic pattern recognition purposes.

Table 6. Silhouette score for clustering evaluation

	Silhouette score		
Number of clusters	**Company A**	**Company B**	**Company C**
2	0.331	0.271	0.194
3	0.176	0.165	0.123
4	0.074	0.160	0.122
5	0.081	0.133	0.087
10	0.104	0.083	0.047
50	0.190	0.056	0.064
100	0.204	0.079	0.071
200	0.214	0.087	0.079
500	0.225	0.089	0.090
1000	0.243	0.108	0.095

The above results indicate that there is no significant grouping to be sought in the datasets.

5 Design of experiments

This chapter contains a description of experiments that are conducted to evaluate the quality of different time series forecasting methods, described in chapter 2, based on the real direct procurement data from industry partners. In part 5.1, we go through the core principles of splitting time series data into parts for training and testing purposes. In parts 5.3 – 5.5, we describe the experiments and sets of hyperparameters included for relevant comparison in cases of exponential smoothing, SARIMA and fuzzy time series methods.

5.1 Performance measurement

Selection of an appropriate measurement of the algorithm performance is essential to a quantitative research in time series forecasting domain. Once the model is fitted based on training data, its predictive capacity is evaluated on a test data sample. Forecast errors are defined as difference between forecasted and original values. The errors are aggregated in one or more of the alternative metrics, including

- Mean Absolute Error (MAE), calculated as $MAE = \frac{\sum_{i=1}^{n} |y_{pred} - y_i|}{n}$;
- Mean Squared Error (MSE), calculated as $MSE = \frac{\sum_{i=1}^{n} (y_{pred} - y_i)^2}{n}$;
- Root Mean Squared Error (RMSE), calculated as $RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{pred} - y_i)^2}{n}}$;

where y_{pred} is the series of forecasted values, y_t is the series of the corresponding original values. Commonly, RMSE is a good representation of an average error term, which accounts for both positive and negative deviations and is scaled back to the original power.

Using normalized values of purchase quantities – scaled to range [0, 1] – allows us to utilize RMSE metrics across datasets and forecasting windows as a comparable performance measure.

5.2 Datasets for training and testing

It is essential for any quantitative study to have appropriate representation of data to fit the model parameters and test their performance. Time series data characteristics present additional challenges related to separation of observations for training and testing purposes. Random selection of objects may ruin the experiment because of sequential nature of observations, i.e. the order of values in data feed cannot be disregarded. It is therefore

common that the split of time series data is performed over temporal indices, thus ensuring original order of values.

The first parameter that needs to be specified with regards to the selection of testing window is its width, or forecast horizon, i.e. number of future observations that we want to generate as a model output. Business context needs to be accounted for in the selection. In this research, we deal with specifics of financial departments and the process of budgetary revisions, which tend to take place with quarterly frequency. With monthly granularity of original observations, we see it reasonable to specify the forecast horizon to 3 months (1 quarter) to support the budgetary revision process.

Dealing with anonymized real data from existing companies implies potential bias related to seasonality or simple coincidence related to externalities. One way to overcome the bias is to include multiple testing windows in the analysis, as per availability and volume of original data. This means a multitude of divisions of a single time series into training and testing periods. Rolling and expanding windows are two related approaches to achieve better representation of data characteristics in testing periods. A traditional rolling window means that, once picked an initial separation, we gain alternative split by shifting the indices by a step of 1 (or any other custom number of observations). With that approach and a step = 1, all data is used in either training and testing capacity, and we have every data point included in the testing period in at least one splitting. The expanding window, in comparison, means that we gradually increase the number of observations in the training dataset, shifting the index of the testing period in a similar fashion. This provides additional dimension to the analysis of results by revealing the sensitivity of algorithms to the amount of training data, thus being a preferred approach in present research.

Finally, there is a decision to be made on the step of index shifting for expanding window approach. Having the length of the series between 36 and 58 (representing 3 to ~5 years of data with monthly frequency), a step of 1 observation would inflate the number of models to be fitted by fold of the length of the series, which is not feasible with the thousands of series that we have at hand. However, there is a clear research need to address the mentioned bias. A compromise would be achieved by specifying an appropriate shifting step, thus controlling the total factor of how many models are fitted on each series.

One important factor that needs to be considered when fixing the approach to the splitting logic is the potential misalignment of time indices for relevant observations across the time series. Data filtering steps guarantee that there is no time series qualified for analysis that would have fewer than 3 full years of relevant observations; however, start and end months of those series may differ within the datasets. If we were to introduce additional constraint and only keep the series that contain business-relevant values over same periods of time, it would further reduce the resulting number of series for analysis, which is not justified. Choosing larger count of qualified time series over aligned time periods, we lose comparability of aggregated results against time dimension (e.g. average performance across all time series based on a fixed testing period). Alternatively, we address this analytical need by introducing a linguistic attribute that specifies the amount of data available for fitting. With allowed length of the series between 36 and 58 values, we call the amount of training data in series under 4 full years of data as "Low" and the rest of them as "High" which allows for some degree of horizontal aggregation.

All things considered, the experiment for each series is implemented in the following algorithm:

1. Identify the first and last period with non-zero normalized quantity values and remove leading and lagging null observations;
2. Split the resulting series into $n_{windows}$ expanding windows, starting with the first $33 + 3 = 36$ months of data (33 observations for training and 3 – for testing purposes) and incrementing the index of last observation included in the sample by $\left[\frac{i_{max}-36}{n_{windows}}\right]$ where i_{max} is the largest integer index of the series (starting with 1, equal to number of observations) and $n_{windows}$ is the target number of windows per series;
3. Run all configurations of each model family (Exponential Smoothing, SARIMA or FTS) on each of the windows and store the results in such a format that it would include
    a. unique identifier of an experiment;
    b. identifier of the dataset;
    c. identifier of the series;
    d. number of observations in training dataset as per the expanding window approach;
    e. attribute of "High" or "Low" amount of training data;

f. values of tested hyperparameters;
g. RMSE measure calculated on forecasted normalized values over the testing window of specified width (3 months).

Having this level of granularity enables multidimensional representation of final results using pivot table and charts functionalities.

5.3 Holt-Winters hyperparameters

In the following sections of the chapter we follow the same notations as those introduced in the respective parts of chapter "Methods".

In triple exponential smoothing, also known as Holt-Winters model, loss optimization methods are leveraged to determine overall, seasonal and trend smoothing factors α, β and γ. Their actual values do not bring additional information to the analysis, and are therefore omitted from the results summary.

The hyperparameters relevant for Holt-Winters model configuration include

- Trend type: *Additive* or *Multiplicative*;
- Seasonality: True or False depending on whether seasonal component is enabled.

In case of enabled seasonality, the following parameters are added to the list:

- Seasonal trend type: *Additive* or *Multiplicative*;
- Length of a seasonal period: [4, 12] describing quarterly and annual seasonality respectively.

In total, there are 10 configuration types (2 non-seasonal and 8 seasonal) to be tested per each time series entity.

5.4 SARIMA hyperparameters

SARIMA model requires prior specification of all non-seasonal and seasonal lags, as well as the length of the season, i.e. the complete list of hyperparameters comprises

- Number of AR lags $p \in \{0, 1, 2, 3\}$;
- Number of MA lags $q \in \{0, 1, 2, 3\}$;
- Number of seasonal AR lags $P \in \{0, 1, 2\}$;
- Number of seasonal MA lags $Q \in \{0, 1, 2\}$;

- Length of the season $L \in \{4, 12\}$;

The range of tested values for p and q relates to the statistical property of time series to be in their majority represented by one of the resulting specifications. Maximal values of P and Q are limited to 2 given that this is the highest lag we can guarantee for annual seasonality with minimal length of the training dataset set to 36 values, i.e. 3 full years of data.

The total number of model specifications tested on each time series is 144. For each of the experiments, AIC criterion is stored alongside with the hyperparameter values.

5.5 Fuzzy Time Series hyperparameters

Similarly, in the domain of fuzzy time series analysis there is configuration setup that needs to be established prior to fitting the models. According to the description of shortlisted FTS methods (section 2.5.4), we test all three main types of models: simple High Order FTS (HOFTS), Weighted High Order FTS (WHOFTS) and Probabilistic Weighted FTS (PWHOFTS).

Specific to fuzzification mechanism, we introduce additional parameter for number of linguistic variables to divide the universe of discourse n_{part}. Without better means to make the decision prior to fitting the model, we specify this parameter discretely in a number of non-evenly distributed steps between extreme values of 5, representing low number of parts, and 50, representing excessively high number.

Finally, order of an FTS model reflects the number of lags included in the fitting process. Thus, hyperparameters for FTS models can be summarized in the following list:

- Model type: HOFTS, WHOFTS, PWHOFTS;
- Number of fuzzy sets in partitioning $n_{part} \in \{5, 10, 20, 50\}$;
- Order of the model ~ number of lags $order \in \{1, 2\}$;

resulting in a total of 24 models per time series.

6 Results and discussion

This concluding chapter of the thesis reveals the outcomes of multiple predictions run over prepared datasets. For each of the methods described in section 3.5, we present a comparative analysis across datasets, evaluate different model specification options to see which parameters stand behind significant differences in the performance metrics, and compare the latter against those of the benchmark approach.

6.1 Holt-Winters performance analysis

The comprehensive results report containing RMSE measurement for every prediction, i.e. every combination of parameters and testing windows of the time series, allows for granular analysis of the model performance across datasets.

6.1.1 Coverage and outliers

It is worth mentioning that not all model specifications resulted in successful fitting. Specifically, over 68% of the predictions failed across the datasets, if we consider the total number of distinct time series – testing windows – hyperparameter values combinations. (Table 7). Further investigation reveals that the cause for failures is that log-likelihood maximization does not converge in a number of cases that are characterized with specified multiplicative type of a general or seasonal trend. It should not raise concern as we omit the exploratory phase of analysis for each individual case and thus do not have full information on feasibility of different specifications. In other words, for some time series it is so inappropriate to fit a model with multiplicative trend, that there is no feasible solution to the error minimization problem, i.e. the parameters cannot be found. It is important to note that every testing window under every series receives at least one feasible prediction (fitting failures in cases of additive trend types are non-existent). This observation proves feasibility of using Holt-Winters model as a method for material forecasting despite unsuccessful fittings of certain model specifications.

Table 7. HW successful and failed experiments

	Success	Count of predictions
Company B	FALSE	33362
	TRUE	21478
Company C	FALSE	32816
	TRUE	17604
Company A	FALSE	18095
	TRUE	9785

While certain specifications cause failure in the convergence of log-likelihood maximization, others yield feasible outcomes which however show prediction values outside of plausible range, resulting in RMSE values approaching infinity. These situations are believed to be caused by multiplicative type of general or seasonal trends too, when fitting such models would not be appropriate even though numerically feasible.

We ensure comparability of results by predicting normalized time series, transformed with a min-max technique to a range of allowed values between 0 and 1. Thus, RMSE below 0.5 can be intuitively perceived as an error of a reasonably good prediction. An overview of predictions labelled as outliers depending on the threshold of RMSE (Fig. 15) shows that most of the cases (47863) are within 0-0.5 range of the error term; a small yet significant number of experiments (4073) show RMSE between 0.5 and 1.0, with another 649 corresponding with the 1.0-2.0 range of the value. The scarcity of the remaining cases, represented by RMSE > 2.0, justify the decision of dropping those experiments from further analysis to avoid distortion in aggregations. After removal of outliers, we continue seeing at least one feasible outcome for each combination of series with all associated testing windows, thus proving that in a business scenario there exists a better alternative – in terms of RMSE – for each of the cases affected by outlier detection procedure.

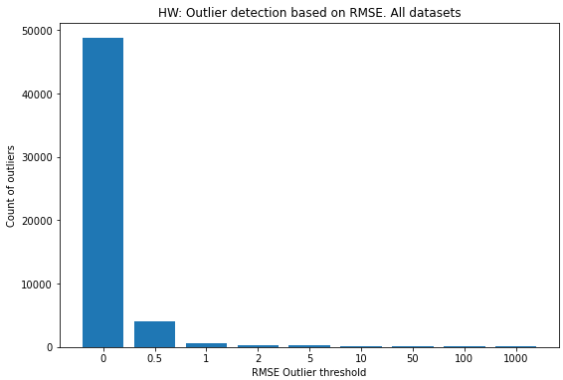

Figure 15. HW results outlier detection

6.1.2 Performance against benchmark

For the sake of benchmarking general performance of forecasting methods against the naïve approach, the lowest RMSE for each time series is selected from the variety of predictions representing different model specifications. This brings us closer to the original business scenario in which the best-performing model would be utilized based on out-of-sample

validation. Thus, aggregation of results is performed on the individual time series level, applying mean and standard deviation calculation to both RMSE and benchmark RMSE measurements (Table 8).

Table 8. HW performance against benchmark

	HW RMSE		**Naive RMSE**	
	Mean	Std.	Mean	Std.
Company B	0.165	0.084	0.219	0.123
Company C	0.210	0.107	0.295	0.153
Company A	0.163	0.075	0.199	0.101

Overall, we see significant improvement in forecasting accuracy (18.1-28.9% reduction in average RMSE compared to the naïve approach) coupled with 25.7-32.1% decrease in its standard deviation, i.e. showing better results in a less volatile way.

In Fig. 16 and in similar visualization describing further methods, we show overlapping distributions of RMSE measurement across predictions made with the tested and benchmark approaches. In the overlapping region of the histograms, a third color is visible to make the overlap distinguishable. In comparison of HW model against the naïve method (Fig. 16), we see a more narrow distribution of observed errors in case of triple exponential smoothing which illustrates the decreased volatility of predictions.

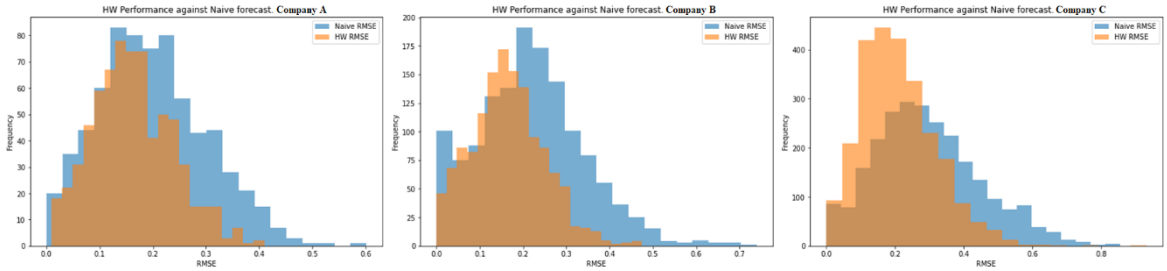

Figure 16. HW performance against benchmark across datasets

6.1.3 Model specifications ranking

Having the complete data at hand, we turn to an overview of model specifications and number of predictions in which those specifications end up being the best-performing alternative for a particular combination of dataset, series, and testing window. The overview (Table 9) indicates that simple additive trend without any seasonal component resulted in lowest RMSE in the highest number of cases (4961), followed by combinations of additive trend with annual (3802) and quarterly (3735) seasonality. Variations with multiplicative

trend type represent minority of predictions, however those can be expected to yield significantly better results on occasions when such specification would be appropriate.

Table 9. Overview of HW model specifications

Trend type	Seasonal trend type	Length of a season	Best performance cases
Additive	None	0	4961
Additive	Additive	12	3802
Additive	Additive	4	3735
Multiplicative	None	0	201
Multiplicative	Additive	12	121
Additive	Multiplicative	4	108
Multiplicative	Additive	4	103
Additive	Multiplicative	12	99
Multiplicative	Multiplicative	4	99
Multiplicative	Multiplicative	12	83

Next, we proceed with more granular analysis of results with regards to different values of configurable parameters. Starting with number of observations in a seasonal period, we have tested scenarios of 0, 4 and 12 months representing no seasonality, quarterly and annual seasonality, respectively. The mean and standard deviation of the RMSE metrics (Table 10) indicate that there is no strong domination of a single length of a season, with only minor (<10%) swing towards quarterly seasonality in 2 out of 3 datasets, backed by lower standard deviation values.

Table 10. HW Performance with different lengths of a seasonal period

		HW RMSE	
	Periods in a season	Mean	Std.
Company B	0	0.218	0.120
	4	0.204	0.093
	12	0.229	0.100
Company C	0	0.277	0.150
	4	0.264	0.120
	12	0.293	0.132
Company A	0	0.217	0.118
	4	0.212	0.094
	12	0.211	0.083

The absence of a clear pattern with seasonality is confirmed based on the overlapping distributions overview (Fig. 17).

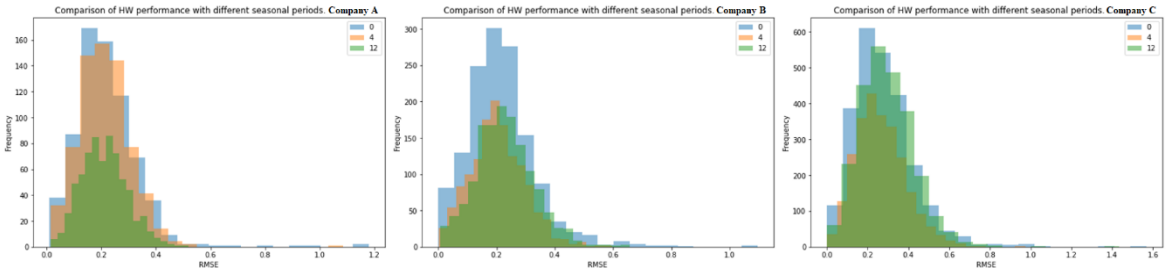

Figure 17. HW RMSE by different lengths of a seasonal window across datasets

Alternative trend types – additive and multiplicative – have been mentioned earlier in this section in the context of non-converging maximum likelihood method resulting in failed fittings. However, as we have seen earlier (Table 9), the multiplicative trend may come as a reasonable solution for certain type of series, e.g. with exponentially increasing or decreasing values. This is also visible in Table 11, where we present the mean RMSE based on utilized trend type across successful predictions, and models with multiplicative trend type show competitive error metrics, while the low number of successful predictions highlights the limited applicability of this specification.

Table 11. HW RMSE by trend type across datasets

		HW RMSE		
	Trend type	Mean	Std.	Count
Company B	Additive	0.166	0.084	1371
	Multiplicative	0.170	0.112	284
Company C	Additive	0.211	0.107	2521
	Multiplicative	0.285	0.113	219
Company A	Additive	0.164	0.075	697
	Multiplicative	0.154	0.115	80

Finally, the amount of training data, defined as *Low* for number of observations below 40, and *High* for number of observations above 40, seems to have a predictably refining impact on the performance measurement (Table 12). Small deviation of mean RMSE depending on the amount of training data can be explained by the same ballpark count of observations, i.e. even though we distinguish between *Low* and *High* amount of training data, in the big picture the line is not too clear, and one could characterize the number of observations used for fitting the models as low for all time series in scope of the analysis. However, even with a small difference in number of values in the training dataset, having more of those has a positive impact on the accuracy volatility which is visible in lower standard deviation levels. Additionally, it is worth mentioning that in one of the datasets the overall availability of

historical data leaves room for only limited analysis with no more than 38 observations used for training.

Table 12. HW RMSE by amount of training data across datasets

		HW RMSE	
	Amount of training data	Mean	Std.
Company B	High	0.164	0.092
	Low	0.166	0.134
Company C	Low	0.210	0.107
Company A	High	0.166	0.084
	Low	0.154	0.123

Based on the overview of HW applied to material forecasting, we see that this traditional time series model represents a viable alternative to the naïve approach. No clear pattern has been identified with regards to the preferred length of a seasonal period, it is therefore reasonable to perform out-of-sample validation and select the appropriate value among no-, quarterly or annual seasonality per time series. With regards to the trend type to be used in predictions, we see that additive trend is much more common in the analyzed datasets; however, if prediction accuracy is a priority, multiplicative trend models can be considered for some cases. The implementation effort for a Holt-Winters model is lowest among the shortlisted methods, which, coupled with clearly better performance as compared to the naïve approach, makes it a good candidate for consideration.

6.2 SARIMA performance analysis

6.2.1 Coverage and outliers

Maximum likelihood methodology is employed to estimate parameters of SARIMA models, too. As opposed to ordinary least squares (OLS), it does not require all regressors to be observable, which is a condition that would not be met in cases of activated moving average (q) or seasonal moving average (Q) components. Log-likelihood maximization, in its turn, makes room for unsuccessful fittings. Exploratory analysis reveals (Table 13) that only a small fraction of error minimization attempts do not converge, and the set of successful fittings is sufficient to have at least one feasible outcome for each testing window of each time series in scope.

Table 13. SARIMA successful and failed experiments

	Success	Count of experiments
Company B	FALSE	23
	TRUE	16429
Company C	FALSE	1
	TRUE	15125
Company A	FALSE	5
	TRUE	8359

Another distinction from previously analysed HW methodology is the practical non-existence of outliers among the results, as defined in earlier sections by RMSE > 2.0. Dozens of thousands of experiments land in RMSE < 0.5, with only 24 in range between 1.0 and 2.0 (Fig. 18).

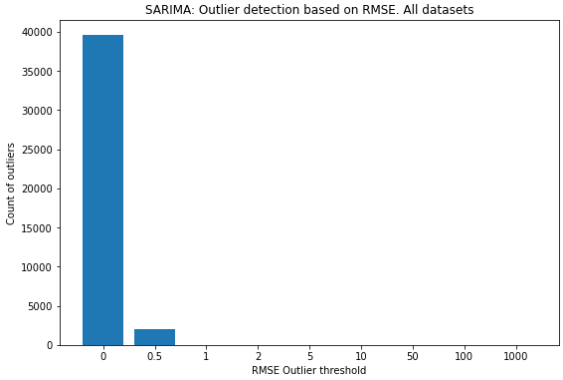

Figure 18. SARIMA results outlier detection

6.2.2 Performance against benchmark

The best-performing configuration selected based on out-of-sample validation is compared against naïve benchmark (Table 14) revealing 10.8 to 23.4% improvement in average RMSE, with 24.5 to 29.4% decrease in standard deviation of the metrics.

Table 14. SARIMA performance against benchmark

	SARIMA RMSE		Naive RMSE	
	Mean	Std.	Mean	Std.
Company B	0.179	0.087	0.219	0.123
Company C	0.226	0.109	0.295	0.153
Company A	0.178	0.076	0.199	0.101

As shown in Fig. 19, the improvement in RMSE is achieved by more narrow distribution (quantified above with standard deviation), and in 2 out of 3 datasets – shifted peak of the histogram towards lower error values.

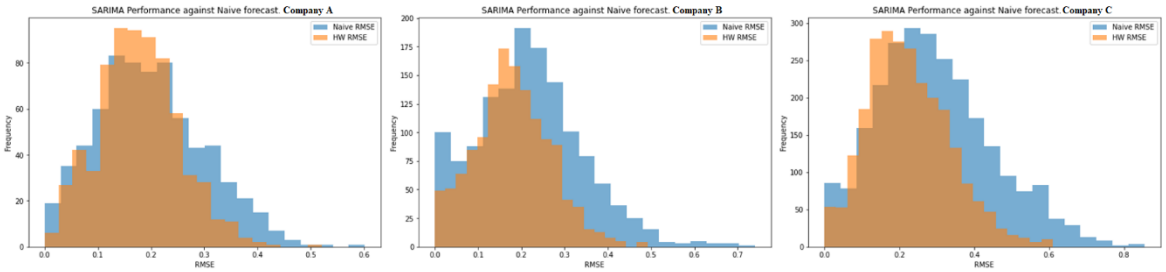

Figure 19. SARIMA performance against benchmark across datasets

6.2.3 Model specifications ranking

As ARIMA models are not designed to deal with intermittent time series, which many of the experiments in this research are based on, it is especially interesting to observe which specifications resulted in lowest error metrics, beating the benchmark by a margin. In total, 302 distinct sets of parameters showed best results on various occasions. More than 75% of the cases are covered with top 19 most frequent specifications, presented in Table 15.

Table 15. Overview of SARIMA configuration options (top 19)

p	d	q	P	D	Q	Intercept	Seasonal	Best performance cases
0	0	0	0	0	0	TRUE	FALSE	4352
0	0	1	0	0	0	TRUE	FALSE	1195
1	0	0	0	0	0	TRUE	FALSE	1063
0	1	1	0	0	0	FALSE	FALSE	825
2	0	0	0	0	0	TRUE	FALSE	486
0	0	0	0	0	1	TRUE	TRUE	403
0	0	0	1	0	0	TRUE	TRUE	251
3	0	0	0	0	0	TRUE	FALSE	225
0	1	0	0	0	0	FALSE	FALSE	204
0	1	2	0	0	0	FALSE	FALSE	175
0	0	0	0	0	0	FALSE	FALSE	167
1	0	1	0	0	0	TRUE	FALSE	151
0	1	1	0	0	0	TRUE	FALSE	148
1	0	1	0	0	0	FALSE	FALSE	130
1	1	1	0	0	0	FALSE	FALSE	126
1	1	0	0	0	0	FALSE	FALSE	120
2	1	0	0	0	0	FALSE	FALSE	118
0	0	2	0	0	0	TRUE	FALSE	107
0	0	1	0	0	1	TRUE	TRUE	105

It is notable that in the 32.7% of all cases the best performing configuration appeared to be what is technically not even an ARIMA model – all zero-valued parameters except for an intercept, or a constant term, which essentially means a simple average of observed historical values. Top 3 is composed of MA(1) and AR(1) models, followed by ARIMA(0, 1, 1) and AR(2). Unlike HW results, seasonal patterns are more heavily underrepresented, with only three options on the top 20 list containing seasonal components, and in total accounting for 15.8% of all cases.

In terms of amount of training data, one would expect ARIMA-type of models to be most sensitive to this limitation as compared to alternatives. However, it is important to remember that in the big picture the amount of training data available for each time series in scope of this research can be characterized as "Low", and the distinction that we are making in this dimension (calling series of <40 observations as "Low" and >40 observations as "High") is quite situational, given that the maximum available number of values for fitting the model is 59. We nevertheless see (Table 16) similar patterns as in the HW case – significantly lower standard deviation metrics for experiments with more training data, even though the tendency is not captured in the mean RMSE.

Table 16. SARIMA RMSE by amount of training data across datasets

		SARIMA RMSE	
	Amount of training data	Mean	Std.
Company B	High	0.180	0.096
	Low	0.176	0.133
Company C	Low	0.226	0.109
Company A	High	0.182	0.084
	Low	0.165	0.126

Automatic approach to SARIMA model specification selection has yielded interesting results – the most successful combination of parameter specifies a simple arithmetic average rather than a complex univariate equation. This fact, alongside with higher mean error term as compared to the HW results, makes SARIMA a less favorable option for implementation in material forecasting information system. Two main reasons for suboptimal performance are named to be intermittent nature of the target series and low amount of training data.

In the course of results interpretation, we propose to think of SARIMA method, which in many cases transformed into a simple arithmetic average, as another benchmark solution. In

cross-method comparison the best-performing specification is picked for each time series, which means that the performance measurement of a SARIMA model will be no-worse than the performance of a simple arithmetic average forecast; thus, if we see evidence of other methods beating SARIMA as of this research, it proves that these methods are also capable of outperforming the simple average as a benchmark method.

6.3 Fuzzy Time Series performance analysis

Fuzzy Time Series modelling, better suited to deal with outliers and intermittency in the original observations, have not been tested before as a method for supply chain forecasting tasks, and thus represents the most interesting and innovative piece of this research.

6.3.1 Coverage and outliers

Overview of successful and failed predictions (Table 17) indicates very low count of series that resulted in unsuccessful fitting due to divergence of error optimization techniques. The same view highlights the special focus on FTS models through high number of predictions conducted with different values of hyperparameters, which allows for granular investigation of optimal specification options.

Table 17. FTS results outlier detection

	Success	Count of experiments
Company B	FALSE	18
	TRUE	131598
Company C	FALSE	0
	TRUE	121008
Company A	FALSE	15
	TRUE	66897

Similar to SARIMA model, FTS does not yield any outlying results, with <0.1% of predictions showing RMSE between 1.0 and 2.0, and none – beyond RMSE = 2.0 threshold.

6.3.2 Performance against benchmark

Even before direct comparison of aggregated RMSE to that from other model types, we see absolute dominance of FTS in terms of its predictive power (Table 18, Fig. 20). Benchmarked against naïve forecasts, we see 39.4 to 47.2% increase in accuracy, and ~40% decrease in standard deviation measuring the variability of error terms across different series and associated testing windows.

Table 18. FTS performance against benchmark

	FTS RMSE		Naive RMSE	
	Mean	Std.	Mean	Std.
Company B	0.125	0.069	0.219	0.123
Company C	0.156	0.088	0.295	0.153
Company A	0.121	0.055	0.199	0.101

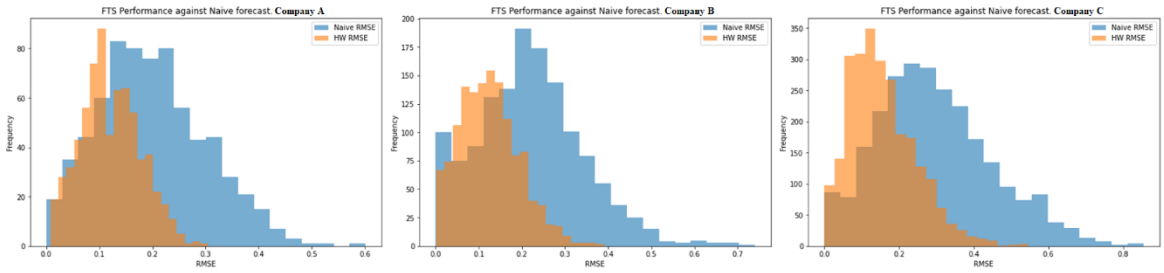

Figure 20. FTS performance against benchmark across datasets

6.3.3 Model specifications ranking

Looking at an overview of model specifications that resulted in best performance over different time series and testing windows combinations (Table 19), we do not observe any clear dominance in any of the dimensions. However, we see that probabilistic weighted fuzzy time series seem to have outperformed weighted high order fuzzy time series in most cases where complex dependencies need to be captured, alongside with simple high order models in cases where the patterns are allegedly more linear. Order values 1 and 2 are evenly present in the report.

Table 19. FTS configuration options

Model type	Number of fuzzy sets	Order	Best performance cases
PWFTS	50	2	1104
PWFTS	50	1	838
HOFTS	50	2	769
HOFTS	5	1	740
PWFTS	5	1	705
PWFTS	20	2	698
HOFTS	50	1	658
HOFTS	10	2	601
HOFTS	5	2	593
HOFTS	20	2	592
PWFTS	5	2	591
PWFTS	10	2	585
HOFTS	20	1	541
PWFTS	20	1	525
WHOFTS	5	2	499
HOFTS	10	1	482
WHOFTS	10	2	410
WHOFTS	50	1	400
WHOFTS	5	1	387
WHOFTS	20	1	369
WHOFTS	10	1	340
PWFTS	10	1	303
WHOFTS	50	2	293
WHOFTS	20	2	291

Arguably the clearest pattern with regards to different values of hyperparameters is visible in the case of number of fuzzy sets into which the universe of discourse is partitioned. Having tested options of [5, 10, 20, 50] characterized as ranging between extremes of insufficiency and excess, we see a downward trend (Fig. 21) in RMSE against the value of that parameter. At the same time, it is visible that the trend does not extend to higher values of fuzzy sets count, taking a slight lift in the rightmost part of the range, implying that the optimal value of the parameter resides within the proposed subset.

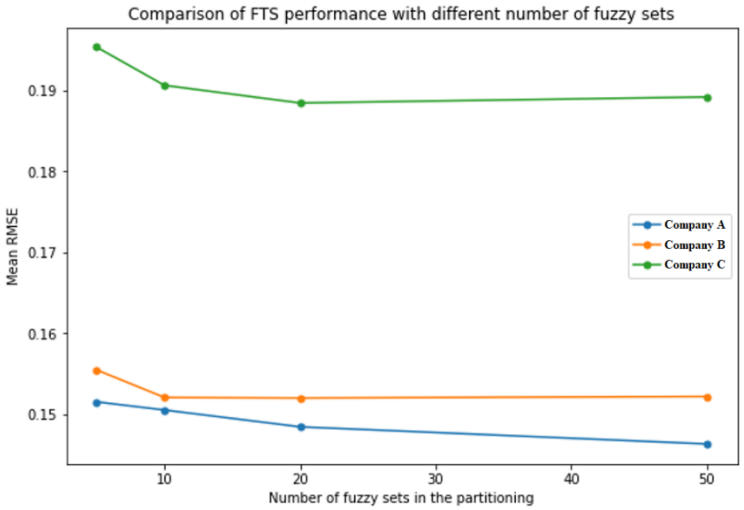

Figure 21. FTS RMSE by number of fuzzy sets in partitioning across datasets

6.4 Comparison of performance across methods

Combining information on mean RMSE as a measurement of performance of different methods (Table 20), we see confirmation of FTS yielding best results across all datasets in terms of both average error term and its standard deviation. All three methods beat the benchmark by a margin, which has been mentioned in earlier sections of this chapter.

Table 20. Comparison of performance of different methods

	HW RMSE		**SARIMA RMSE**		**FTS RMSE**		**Naive RMSE**	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Company B	0.165	0.084	0.179	0.087	0.125	0.069	0.219	0.123
Company C	0.210	0.107	0.226	0.109	0.156	0.088	0.295	0.153
Company A	0.163	0.075	0.178	0.076	0.121	0.055	0.199	0.101

For additional perspective, let us consider in how many cases each model type showed best performance (Table 21). Properly configured FTS would be the best choice in almost 60% of the cases, followed by HW and Naïve benchmark. Stemming from limitations related to amount of training data and intermittency of the values, SARIMA outperforms other models in only about 8% of the experiments, in this respect showing worst results among the alternatives incl. the benchmark.

Table 21. Number of experiments by best performing method

Model	Best performance in … cases
FTS	7947
HW	2642
Naive	1624
SARIMA	1099

Referring to an earlier remark on SARIMA models transforming into simple arithmetic average in a large number of cases, the above results may also claim that Fuzzy Time Series and Holt-Winters models outperform both naïve and simple average benchmarks, thus proving that more sophisticated time series forecasting techniques have the potential to reveal hidden patterns in the real-world supply chain data.

Fuzzy approach shows particularly good prediction accuracy compared to other methods in scope of this research. The main reason for this is its ability to handle the intermittency situation by fuzzification of original series, as opposed to the methods that operate on a continuous scale. During that part of data processing, zero values alternating with non-zero ones are translated into a discrete number of fuzzy sets, which reduces the noise in identifying sequential patterns.

7 Conclusion

In this research, we have outlined the theoretical background for supply chain forecasting problem, shortlisted and tested three types of univariate time series models using real direct materials purchasing data of the industry partners. Next, we summarize the outcomes of the research and implications in business context.

Supply chains of modern enterprises are complex and involve multiple parties. Previous research papers unanimously claim that most accurate and interpretable demand forecasts require collaborative mechanisms over the length of the supply chain, which, for a variety of business reasons, tend to fail. It is therefore important to understand the univariate models for direct material forecasting. Holt-Winters exponential smoothing, SARIMA and Fuzzy Time Series have been shortlisted for detailed analysis as part of this research; the latter being a novel approach to the domain.

In the quantitative part of the research, we have tested the shortlisted models on three independent datasets containing historical direct material purchasing data of industry partners. Overall, the results of the research indicate that there is potential to reveal hidden intrinsic and seasonal patterns and increase the accuracy of the currently used naïve approach by a margin of up to 47% depending on the dataset and the method. We have also shown substantial improvement compared to simple statistical forecasts, such as arithmetic average.

Fuzzy Time Series models have shown the best performance across all datasets, arguably due to their ability to reduce the noise caused by intermittency of the original series. Holt-Winters represents another viable alternative to benchmark methods, showing stable improvement to the error metrics. We do not recommend exploring SARIMA in Sievo business context, because the amount of training data is insufficient for this model type, resulting in its underperformance compared to the alternatives.

When it comes to method selection for business application, there is a tradeoff between forecast accuracy and complexity of implementation. We recommend that those are considered for Fuzzy Time Series and Holt-Winters models. The selected method may be used to generate automatic forecasts of direct material demand quantity for the series with sufficient amount of historical data – those would cover 35-86% of spend in the analyzed datasets; benchmark methods, such as simple average or last observed value, can be used for others.

Bibliography

Ali, M. M., & Boylan, J. E. 2012. On the effect of non-optimal forecasting methods on supply chain downstream demand. IMA Journal of Management Mathematics, 23, 81–99

Boylan, J. E., & Syntetos, A. A. 2015. Supply chain forecasting: the customer dimension. Conference Proceedings. 27th European Conference on Operational Research EURO 2015, July 12-15, 2015.

Carbonneau, Réal & Laframboise, Kevin & Vahidov, Rustam. 2008. Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research. 184. 1140-1154.

Chandra, C., Grabis, J. 2005. Application of multi-steps forecasting for restraining the bullwhip effect and improving inventory performance under autoregressive demand. European Journal of Operational Research 166 (2), 337–350.

Chen, H., & Boylan, J. E. 2007. Use of individual and group seasonal indexes in sub-aggregate demand Forecasting. Journal of the Operational Research Society, 58, 1660–1671.

Chopra, S., & Meindl, P. 2012. Supply chain management: strategy, planning and operation.5th edition. USA New Jersey: Pearson.

Dalhart, G. 1974. Class seasonality – a new approach. Conference proceedings. American production and inventory control society.

Daniel Ortiz-Arroyo and Jens Runi Poulsen, 2018. A Weighted Fuzzy Time Series Forecasting Model, Indian Journal of Science and Technology, Vol. 39, pp. 1-11

Davis, E.W., Spekman, R. 2004. Extended Enterprise. USA, New Jersey, Upper Saddle River: Prentice-Hall.

Dejonckheere, J., Disney, S.M., Lambrecht, M.R., Towill, D.R., 2003. Measuring and avoiding the bullwhip effect: A control theoretic approach. European Journal of Operational Research 147 (3), 567–590.

Draper, N. R.; Smith, H. 1998. Applied Regression Analysis. Wiley-Interscience. ISBN 978-0-471-17082-2.

Efendigil, Tuğba & Önüt, Semih & Kahraman, Cengiz. 2009. A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. Expert Systems with Applications. 36. 6697-6707.

Efendigil, Tuğba & Önüt, Semih. 2012. An integration methodology based on fuzzy inference systems and neural approaches for multi-stage supply-chains. Computers & Industrial Engineering. 62. 554-569.

Fildes, R., & Goodwin, P. 2007. Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces, 37, 570–576.

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

Fisher,R. A. 1936. The use of multiple measurements in taxonomic problems. Annual Eugenics, 7, Part II, 179-188

Gardner, E. S., Jr. 2011. Forecasting for operations. Conference proceedings. 31st international symposium on forecasting, Prague, June 27-29, 2011

Heikkila¨, J., 2002. From supply to demand chain management: Efficiency and customer satisfaction. Journal of Operations Management 20 (6), 747–767.

Hyndman, R. J. & Kostenko, A. V. 2007. Minimum sample size requirements for seasonal forecasting models. Foresight, 6, 12–15.

Hyndman, R.J. & Athanasopoulos, G., 2018. Forecasting: principles and practice, OTexts: Melbourne, Australia, 2nd Edition

Kaiser, H. F., 1974. An index of factorial simplicity, Psychometrika, Vol. 39, pp. 31-36

Kurrotul Ayun, Agus Maman Abadi, Fitriana Yuli Saptaningtyas, 2015. Application of Weighted Fuzzy Time Series Model to Forecast Trans Jogja's Passengers, International Journal of Applied Physics and Mathematics, Vol. 5, Issue 2, pp. 76-85

Lee, H. L., Padmanabhan, V., & Whang, S. 1997. Information distortion in a supply chain: the bullwhip effect. Management Science, 43, 546–558.

Luiz, K. H., Pedro, A. M., & Pedro, L. V. P. 1992. The effect of overlapping aggregation on time series models: an application to the unemployment rate in Brazil. Brazilian Review of Econometrics, 12, 223–241.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967), 281-297.

Marzban, C., R. Illian, D. Morison, P. D. Mourad, 2013. Within-group and between-group correlation: Illustration on noninvasive estimation of intracranial pressure, IEEE Journal of Biomedical and Health Informatics

Mohammadipour, M., Boylan, J. E., & Syntetos, A. A. 2012. The application of product-group seasonal indexes to individual products. Foresight, 26, 18–24.

Nikolopoulos, K., Syntetos, A. A., Boylan, J., Petropoulos, F., & Assimakopoulos, V. 2011. An aggregate-disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. Journal of the Operational Research Society, 62, 544–554.

Ouwehand, P., van Donselaar, K. H., & de Kok, A. G. 2005. The Impact of the forecasting horizon when forecasting with group seasonal indexes. The Netherlands: Eindhoven University of Technology Working paper 162.

Pearson, K. 1895. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. 58: 240–242.

Premkumar, G. 2000. Interorganization Systems and Supply Chain Management: An Information Processing Perspective. IS Management. 17. 1-14.

Raghunathan, Srinivasan. 2007. Interorganizational Collaborative Forecasting and Replenishment Systems and Supply Chain Implications. Decision Sciences. 30. 1053 - 1071.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., & Ducq, Y. 2013. Demand forecasting by temporal aggregation. Naval Research Logistics, 60, 479–498.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. 1986. Learning internal representations by error propagation. In: Rumelhart, J.L., McCleland, J.L. (Eds.), Parallel distributed processing, vol. 1. USA, MA, Cambridge: MIT Press, pp. 318–362.

Silva, Petrônio, 2019. Scalable Models for Probabilistic Forecasting with Fuzzy Time Series, Thesis for: Ph.D.

Spithourakis, G. P., Petropoulos, F., Babai, M. Z., Nikolopoulos, K., & Assimakopoulos, V. 2011. Improving the performance of popular supply chain forecasting techniques: an empirical investigation. Supply Chain Forum: An international Journal, 12, 16–25.

Syntetos, A. A. 2014. Forecasting by temporal aggregation. Foresight, 34, 6–11.

Syntetos, A. A., Kholidasari, I., & Naim, M. 2015. The effects of integrating management judgement into OUT levels: in or out of context? European Journal of Operational Research.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. 2009. The effects of integrating management judgment into intermittent demand forecasts. International Journal of Production Economics, 118, 72–81.

Thonemann, U.W. 2002. Improving supply-chain performance by sharing advance demand information. European Journal of Operational Research 142-1, 81–107.

Vakharia, A.J. 2002. E-business and supply chain management. Decision Sciences 33 (4), 495–505.

Weiss, A. A. 1984. Systematic sampling and temporal aggregation in time series models. Journal of Econometrics, 26, 271–281.

Werbos, P.J. 1990. Backpropagation through time: What it does and how to do it. Proceedings of IEEE 78 (10), 1550–1560.

Withycombe, R. 1989. Forecasting with combined seasonal indexes. International Journal of Forecasting, 5, 547–552.

Yusuf, Y.Y., Gunasekaran, A., Adeleye, E.O., Sivayoganathan, K., 2004. Agile supply chain capabilities: Determinants of competitive objectives. European Journal of Operational Research 159 (2), 379–392.