Improving Demand Forecasting: The Challenge of Forecasting Studies Comparability and a Novel Approach to Hierarchical Time Series Forecasting

Zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften (Dr. rer. pol.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Markus Bauer, M.Sc.

Tag der mündlichen Prüfung: 17. Juli 2023

Referent:	Prof. Dr. Clemens van Dinther
Korreferent:	Dr. Michael Färber

Kurzfassung

Bedarfsprognosen sind in der Wirtschaft unerlässlich. Anhand des erwarteten Kundenbedarfs bestimmen Firmen beispielsweise welche Produkte sie entwickeln, wie viele Fabriken sie bauen, wie viel Personal eingestellt wird oder wie viel Rohmaterial geordert werden muss. Fehleinschätzungen bei Bedarfsprognosen können schwerwiegende Auswirkungen haben, zu Fehlentscheidungen führen, und im schlimmsten Fall den Bankrott einer Firma herbeiführen.

Doch in vielen Fällen ist es komplex, den tatsächlichen Bedarf in der Zukunft zu antizipieren. Die Einflussfaktoren können vielfältig sein, beispielsweise makroökonomische Entwicklung, das Verhalten von Wettbewerbern oder technologische Entwicklungen. Selbst wenn alle Einflussfaktoren bekannt sind, sind die Zusammenhänge und Wechselwirkungen häufig nur schwer zu quantifizieren.

Diese Dissertation trägt dazu bei, die Genauigkeit von Bedarfsprognosen zu verbessern.

Im ersten Teil der Arbeit wird im Rahmen einer überfassenden Übersicht über das gesamte Spektrum der Anwendungsfelder von Bedarfsprognosen ein neuartiger Ansatz eingeführt, wie Studien zu Bedarfsprognosen systematisch verglichen werden können und am Beispiel von 116 aktuellen Studien angewandt. Die Vergleichbarkeit von Studien zu verbessern ist ein wesentlicher Beitrag zur aktuellen Forschung. Denn anders als bspw. in der Medizinforschung, gibt es für Bedarfsprognosen keine wesentlichen vergleichenden quantitativen Meta-Studien. Der Grund dafür ist, dass empirische Studien für Bedarfsprognosen keine vereinheitlichte Beschreibung nutzen, um ihre Daten, Verfahren und Ergebnisse zu beschreiben. Wenn Studien hingegen durch systematische Beschreibung direkt miteinander verglichen werden können, ermöglicht das anderen Forschern besser zu analysieren, wie sich Variationen in Ansätzen auf die Prognosegüte auswirken – ohne die aufwändige Notwendigkeit, empirische Experimente erneut durchzuführen, die bereits in Studien beschrieben wurden. Diese Arbeit führt erstmals eine solche Systematik zur Beschreibung ein.

Der weitere Teil dieser Arbeit behandelt Prognoseverfahren für intermittierende Zeitreihen, also Zeitreihen mit wesentlichem Anteil von Bedarfen gleich Null. Diese Art der Zeitreihen erfüllen die Anforderungen an Stetigkeit der meisten Prognoseverfahren nicht, weshalb gängige Verfahren häufig ungenügende Prognosegüte erreichen. Gleichwohl ist die Relevanz intermittierender Zeitreihen hoch – insbesondere Ersatzteile weisen dieses Bedarfsmuster typischerweise auf. Zunächst zeigt diese Arbeit in drei Studien auf, dass auch die getesteten Stand-der-Technik Machine Learning Ansätze bei einigen bekannten Datensätzen keine generelle Verbesserung herbeiführen. Als wesentlichen Beitrag zur Forschung zeigt diese Arbeit im Weiteren ein neuartiges Verfahren auf: Der Similarity-based Time Series Forecasting (STSF) Ansatz nutzt ein Aggregation-Disaggregationsverfahren basierend auf einer selbst erzeugten Hierarchie statistischer Eigenschaften der Zeitreihen. In Zusammenhang mit dem STSF Ansatz können alle verfügbaren Prognosealgorithmen eingesetzt werden – durch die Aggregation wird die Stetigkeitsbedingung erfüllt. In Experimenten an insgesamt sieben öffentlich bekannten Datensätzen und einem proprietären Datensatz zeigt die Arbeit auf, dass die Prognosegüte (gemessen anhand des Root Mean Square Error

RMSE) statistisch signifikant um 1-5% im Schnitt gegenüber dem gleichen Verfahren ohne Einsatz von STSF verbessert werden kann. Somit führt das Verfahren eine wesentliche Verbesserung der Prognosegüte herbei.

Zusammengefasst trägt diese Dissertation zum aktuellen Stand der Forschung durch die zuvor genannten Verfahren wesentlich bei. Das vorgeschlagene Verfahren zur Standardisierung empirischer Studien beschleunigt den Fortschritt der Forschung, da sie vergleichende Studien ermöglicht. Und mit dem STSF Verfahren steht ein Ansatz bereit, der zuverlässig die Prognosegüte verbessert, und dabei flexibel mit verschiedenen Arten von Prognosealgorithmen einsetzbar ist. Nach dem Erkenntnisstand der umfassenden Literaturrecherche sind keine vergleichbaren Ansätze bislang beschrieben worden.

Abstract

Demand forecasting is essential in business. Based on expected customer demand, companies determine, for example, which products to develop, how many factories to build, how much staff to hire or how much raw material to order. High forecasting errors in demand forecasts can have serious consequences, lead to wrong decisions and, in the worst case, cause the bankruptcy of a company.

But in many cases, it is complex to anticipate the actual demand in the future. The influencing factors can be manifold, e.g., macroeconomic development, the behavior of competitors, technological developments, etc. And even if all influencing factors are known, the interrelationships and interactions are often difficult to quantify.

This dissertation contributes to improving the accuracy of demand forecasts.

In the first part of the thesis, as part of a comprehensive overview of the entire spectrum of application fields of demand forecasts, a novel approach is introduced on how to systematically compare studies on demand forecasts and applied to the example of 116 recent studies. Improving the comparability of studies is a major contribution to current research. Unlike, for example, in medical research, there are no significant comparative quantitative meta-studies for demand forecasts. This is because empirical studies in demand forecasting do not use a standardized way to describe their data, procedures, and results. If, on the other hand, studies can be directly compared to each other through systematic description, this will allow other researchers to better analyze how variations in approaches affect forecast performance-without the burdensome need to re-run empirical experiments that have already been described in studies. This work is the first to introduce such a systematic approach to description.

The remaining part of this work deals with forecasting methods for intermittent time series, i.e., time series with a substantial share of demands equal to zero. This type of time series does not meet the requirements for continuity of most forecasting methods, which is why common methods often achieve insufficient forecasting quality. Nevertheless, the relevance of intermittent time series is high–especially spare parts typically exhibit this demand pattern. First, this thesis shows in three studies that even the tested state of the art machine learning approaches do not bring about a general improvement for some known data sets. As a major contribution to research, this work further demonstrates a novel method: The Similarity-based Time Series Forecasting (STSF) approach uses an aggregation-disaggregation procedure based on a self-generated hierarchy of statistical properties of the time series. In connection with the STSF approach, all available forecasting algorithms can be used–due to the aggregation, the continuity condition is fulfilled. In experiments on a total of seven publicly known datasets and one proprietary dataset, the work shows the forecast quality (measured by the root mean square error RMSE) can be statistically significantly improved by 1-5% on average compared to the same procedure without the use of STSF. Thus, the method leads to a significant improvement of the forecast quality.

In summary, this dissertation contributes significantly to the current state of research through the previously mentioned approaches. The proposed procedure for standardizing empirical studies accelerates the progress of research by enabling comparative studies. And the STSF framework provides an approach that reliably improves forecasting quality, while being flexible to use with different types of forecasting algorithms.

Table of Contents

List of Figures

List of Tables

Preface

First and foremost, I would like to thank my supervisor for this dissertation, Prof. Dr. Clemens van Dinther for his excellent support, great expertise and stimulating exchanges. His advice, guidance and direction on scientific work and technical issues were a major factor in the success of this dissertation.

Furthermore, I would like to thank my colleagues at the chair of Reutlingen University for the excellent cooperation in joint publications and the inspiring exchange during the development and testing of the STSF procedure.

I would also like to thank Dr. Marc Schleyer, who made this dissertation possible for me not only through professional exchange, but also through valuable contacts in the scientific world and the support of my doctoral exemption.

Finally, I am very grateful to my family, who have always motivated and encouraged me on the way to my doctorate.

(Markus Bauer)

Karlsruhe, July 18, 2023

1 Introduction

Being able to estimate future requirements for raw materials, goods or services in advance is of great relevance in all parts of society, regardless of whether they are commercial enterprises, the public sector or other organizations. The clearer the assessment of the actors is about the type and amount of future demands, the more effectively available resources (raw materials, machines, working time, etc.) can be used to meet the demands. According to a report by Grand View Research, companies worldwide spent US$ 3.9 billion on demand planning software solutions alone in 2022 (Grand View Research 2022).

Demand forecasting models should help to estimate future demands more accurately. In this context, the scientific community has developed corresponding demand forecasting models that map the influences on demands, can be adjusted on the basis of historical data and, as a result, calculate forecast values for the expected demands.

However, the factors influencing demand are as diverse as the applications of demand forecasts. Complex relationships between influencing factors, unknown influences and random variables make accurate forecasting difficult and, in the worst case, lead to significant deviations from actual demand. Two examples from the field of electricity demand forecasting show how different the requirements can be in just one application area. In the long-term range (1-5 years, e.g., for planning new power plant construction), forecast models can be based on macroeconomic variables in particular. In contrast, for short-term forecasts (1 minute to a few hours), macroeconomic variables have no relevant influence on power consumption. Here, daily conditions (e.g. weather) are much more relevant. Based on the different influencing factors alone, it is clear that the requirements and the choice of approaches used differ fundamentally.

For this reason, scientists have continuously developed the models over the past decades and adapted new methods, for example from the field of machine learning. In particular, they have developed highly specialized methods for particular applications, producing an enormous variety of approaches. At the same time, there is no uniformly accepted and applied standard to describe scientific work on the application of prediction methods, making comparability difficult in many cases. While there are some reviews within the specialized application areas, there are no comparative works that systematically map the topic area of demand forecasting as a whole and compare the methods used. Research thus misses the opportunity to specifically compare results and adopt procedures from other sub-disciplines.

The first part of this dissertation starts exactly at this point. The thesis not only gives a clear overview of the state of the art: most relevant application areas of demand forecasting in the scientific literature as well as the methods used and current issues in research. More important, the work presents an approach for researchers to systematically describe their approaches in order to be able to consistently compare them with other methods.

Within the broad spectrum of demand forecasting, one class presents a particular challenge: Intermittent time series. These are patterns of demand where demand equals zero in many time periods. Probably the most frequently cited example of intermittent time series is the spare parts business. Here, a particular part is often not in demand for days, weeks or even months. Nevertheless, the business is very relevant for companies because of the high margins. According to a survey by Jefferies & Company, sales of spare parts in the automotive industry alone amounted to over 400 billion US dollars in 2021 (Jefferies & Company 2022). However, the intermittent characteristic contradicts the common assumptions of forecasting models, which assume a steady demand pattern. Therefore, methods that usually deliver good results for continuous distributions deliver unexpectedly low forecast quality for intermittent time series and are therefore often unsuitable.

To improve prediction results for intermittent time series, some studies use hierarchies that exist between predicted objects. For example, a well-known study here is the work of Petropoulos and Kourentzes, which uses hierarchies in Royal Air Force spare parts for forecasting (Petropoulos and Kourentzes 2015). The use of relationships among predicted objects seems to be intuitive. But for one thing, such hierarchies are regularly unavailable in practice (because they simply have not been created due to the amount of work required for often tens of thousands of parts). Nor are the hierarchies tested for true correlation of demand patterns–rather, they are mostly hierarchies of similar usage or similar visual characteristics. Thus, an important open question is whether hierarchies can be formed that can be specifically based on the intermittent demand patterns, and whether this can be used to create improved demand forecasts.

The second part of this thesis takes up both: the use of statistically meaningful similarity relationships to form a hierarchy, and the use of aggregation to ensure the continuity conditions of common forecasting methods. The Similarity-based Time Series Forecasting (STSF) framework developed from this work shows an improvement in forecast quality of up to 50% on 8 sample data sets in the empirical studies conducted.

1.1 Research questions

The focus of this dissertation is to improve the forecasting quality of demand forecasts in general. For this purpose, the thesis starts at two points. First, it aims not only to provide a comprehensive overview of the current state of research, but moreover to contribute to improving the comparability of demand forecasting studies. Second, it aims to improve the forecasting quality of demand forecasts, especially for data sets with intermittent time series that are particularly difficult to predict. From these overarching goals, the following research questions are derived to provide a structure for this work.

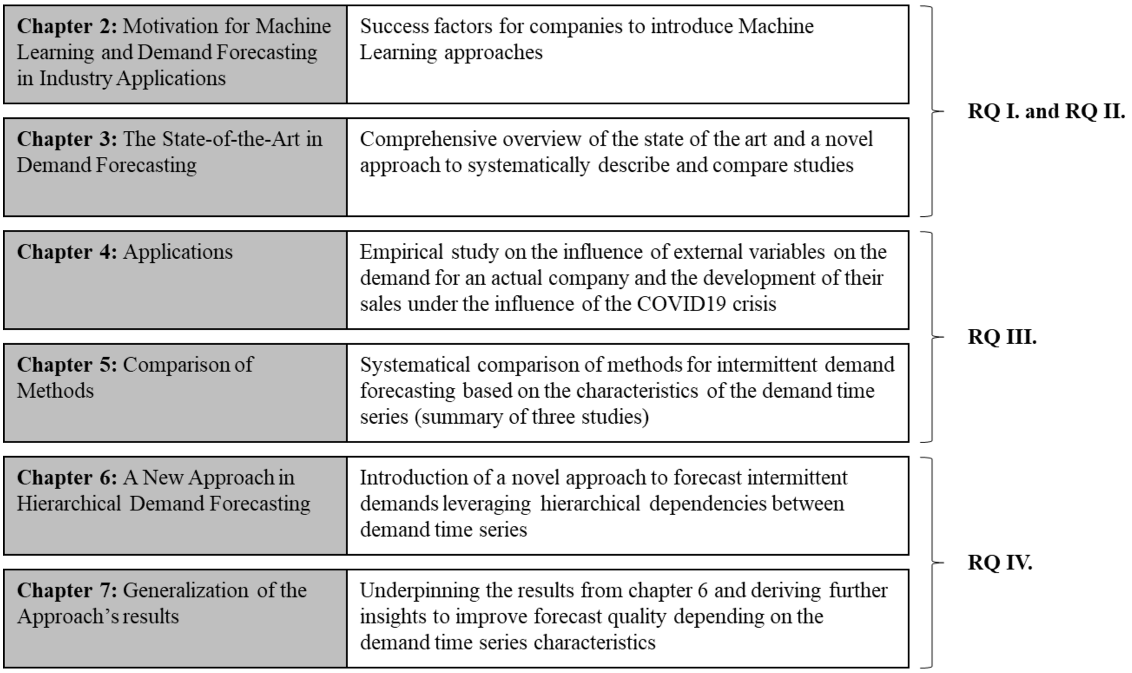
RQ I. What is the current state of demand forecasting in research, what are open issues discussed in literature and how far has adoption of the current state of the art proceeded in companies?

RQ II. How can demand forecasting studies be structured to increase the comparability between studies?

RQ III. What is the forecasting performance of machine learning methods compared to classical approaches for intermittent time series and how can approaches be selected depending on the time series characteristics?

RQ IV. How can intermittent and hierarchical demand forecasting be improved?

The research questions represent the basic structure for the following chapters (see Figure 1).



Chapter 2: Motivation for Machine Learning and Demand Forecasting in Industry Applications	Success factors for companies to introduce Machine Learning approaches	**RQ I. and RQ II.**
Chapter 3: The State-of-the-Art in Demand Forecasting	Comprehensive overview of the state of the art and a novel approach to systematically describe and compare studies	
Chapter 4: Applications	Empirical study on the influence of external variables on the demand for an actual company and the development of their sales under the influence of the COVID19 crisis	**RQ III.**
Chapter 5: Comparison of Methods	Systematical comparison of methods for intermittent demand forecasting based on the characteristics of the demand time series (summary of three studies)	
Chapter 6: A New Approach in Hierarchical Demand Forecasting	Introduction of a novel approach to forecast intermittent demands leveraging hierarchical dependencies between demand time series	**RQ IV.**
Chapter 7: Generalization of the Approach's results	Underpinning the results from chapter 6 and deriving further insights to improve forecast quality depending on the demand time series characteristics	

Figure 1: Structure of the chapters of this document

1.2 Structure and contents of this thesis

In the introduction of this dissertation, it was already stated that demand forecasts are of great relevance for demand planning, especially for companies. However, in 2019, only about 10% of companies reported using machine learning or other advanced techniques for demand forecasting, despite their widespread use in academia. Thus, a significant gap exists between the state of research and its use in practice. In order to accelerate the adoption of advanced methods in practice, it is necessary to understand what the prerequisites and success factors are for their use in companies.

Chapter 2 (*On the Industry Need for Machine Learning and Demand Forecasting*), presents the results of the survey of 18 companies. The chapter contributes to the state of the research by providing a system of success factors for the successful implementation of procedures derived from the survey. It also deduces which success factors depend on the size of the company and which depend on the maturity level of the company with regard to technology use. In addition, the study contributes to research and application by providing definitive approaches and measures to close the gap between research and application by type (size) of company.

Chapter 3 (*How the Demand Forecasting Literature and Applications can Benefit from Better Comparability*) presents a comprehensive overview of the state of the art across all application areas in the field of demand forecasting, comprising 116 studies. Despite the fact, that there is currently no comprehensive and comparative overview of the research area at the current state of knowledge, the chapter contributes to research in particular through a novel approach. This approach describes how demand forecasting studies can be described in a structured way so that they can be systematically compared with other studies–and thus addresses research question RQ II.

Chapter 4 (*Developing an Understanding of External Factors Influencing Demand Forecasting Models using a Case Example*) introduces the use of machine learning for practical applications. Using the sales forecast of a medium-sized company as an example, the influence of external factors on the forecast of extreme events (COVID19 crisis) is shown in particular.

Chapter 5 (*How Time Series Characteristics Affect the Forecast Quality in State-of-the-Art Algorithms for Intermittent Demands*) addresses research question RQ III. from the introduction. Based on three studies, this chapter compares machine learning and classical approaches for forecasting intermittent demands. Through the comparison, the chapter expands the understanding of research on the relationship between time series characteristics and forecasting techniques. The first study systematically compares state of the art approaches. The second study derives characteristics of data sets that can be used to recommend the choice of forecasting method. The third study demonstrates a novel approach for combining Coston's method and deep learning algorithms.

Chapter 6 (*A New Approach in Hierarchical Demand Forecasting*) introduces the new STSF (Similarity-based Time Series Forecasting) framework. It contributes to the state of the art by showing how hierarchies between time series can be used in conjunction with an aggregation-disaggregation procedure to improve forecasting results, and thus also provides part of the answer to research question RQ IV. The chapter introduces the approach and demonstrates the improvement empirically using two data sets.

Chapter 7 *(Generalization of the Approach's results)* extends the empirical basis of the novel approach proposed in Chapter 6. In addition to the two data sets studied in Chapter 6, six more data sets are incorporated to the study. The study contributes to the research by putting the insights from Chapter 6 on a broad empirical basis and provides an improved design for the main pipeline approach, reducing the complexity and computation efforts. The chapter represents the second part of the answer to research question RQ IV.

1.3 References

Grand View Research (2022): Demand Planning Solutions Market Size, Share & Trends Analysis Report By Component, By Deployment (On-premises, Cloud-based), By Enterprise Size, By Industry, And By Region, And Segment Forecasts, 2022 - 2030. Available online at https://www.grandviewresearch.com/industry-analysis/demand-planning-solutions-market-report, checked on 5/18/2023.

Jefferies & Company (2022): Total sales of the global automotive aftermarket from 2016 to 2021 (in billion U.S. dollars). Statista. Available online at https://www.statista.com/statistics/581758/size-of-global-automotive-parts-aftermarket/.

Petropoulos, Fotios; Kourentzes, Nikolaos (2015): Forecast combinations for intermittent demand. In *Journal of the Operational Research Society* 66 (6), pp. 914–924. DOI: 10.1057/jors.2014.62.

2 On the Industry Need for Machine Learning and Demand Forecasting

[*This chapter consists of a study published in 2020:* Bauer, Markus; van Dinther, Clemens; Kiefer, Daniel (2020): Machine learning in SME: an empirical study on enablers and success factors. In Association for Information Systems (Ed.): AMCIS 2020 proceedings: Association for Information Systems, pp. 1–10. – *further referred to as* Bauer et al. 2020]

Machine Learning in SME: An Empirical Study on Enablers and Success Factors

Machine learning (ML) techniques are rapidly evolving, both in academia and practice. However, enterprises show different maturity levels in successfully implementing ML techniques. Thus, we review the state of adoption of ML in enterprises. We find that ML technologies are being increasingly adopted in enterprises, but that small and medium-size enterprises (SME) are struggling with the introduction in comparison to larger enterprises. In order to identify enablers and success factors we conduct a qualitative empirical study with 18 companies in different industries. The results show that especially SME fail to apply ML technologies due to insufficient ML knowhow. However, partners and appropriate tools can compensate this lack of resources. We discuss approaches to bridge the gap for SME.

2.1 Introduction

Since the first appearance of Machine Learning (ML) in the 1950s, the field of ML has rapidly evolved: Numerous applications have been studied in research and practice, frameworks have been developed and implemented as well as fast hardware for computation is available and affordable (OECD 2015). Therefore, the adoption of ML applications in enterprises has significantly increased. Whereas in the year 2015, only 10% of companies reported the utilization of ML in every-day operations, recent studies find about one third of the companies relying on ML (Howard and Rowsell-Jones 2019).

However, studies also report a significant difference regarding the size of the companies. A study from 2019 indicates that companies with less than 500 employees are four times less likely to have ML applied than companies with more employees (Spiceworks 2020). This finding is in line with more general studies on the flexibility of SME to adopt new technologies compared to larger companies: Larger companies overall have a higher adoption rate of information and communication technologies (ICT) than SME (OECD 2004).

We want to find out where these differences derive from and pose the following research questions:

RQ1: What is the gap of the adoption of ML in small- and medium-sized enterprises (SME) compared to the state-of-the-art and best practice?

RQ2: What are challenges in the process of implementation of ML specific to SME and what success factors enable companies to mitigate these challenges? What conditions facilitate the utilization of ML in SME?

Our approach is a synthesis of a meta-analysis of literature and surveys in combination with a qualitative empirical approach, conducted in 2020 with a focus on businesses in the industries of manufacturing and production, retailing and logistics. We subsume all ML applications for internal processes and products.

2.2 Adoption of ML in Research and Enterprises

For this study, we follow the definition of Mitchell for ML: “A computer program is set to learn from an experience *E* with respect to some task *T* and some performance measure *P* if its performance on *T* as measured by *P* improves with experience *E*.” (Mitchell 1997). We are aware that practitioners oftentimes also use the term artificial intelligence (AI) interchangeably, even though the terms are not identical.

Various studies on case specific applications of ML for business use cases have been published in academic literature. In this context we only provide an overview of sample applications to demonstrate the fact that the theoretical foundation for most ML applications in enterprises is available. For each application, we conducted a systematic literature review following Webster and Watson and Levy and Ellis and each chose one article with highest citation count (Levy and Ellis 2006; Webster and Watson 2002 – see Table 1).

We conclude that there is a high number of articles that provide both a theoretical foundation for ML technologies as well as application use cases as basic requirement of RQ1. Most relevant use cases for business applications are intensely studied and practical solutions were demonstrated by academia.

Finding I: *Research provides a strong foundation of ML basic techniques but also of ML applications for business applications. ML technologies are ripe for implementation in enterprises.*

Forecasting	Classification	Optimization	NLP & IR
6,500 papers*	8,200 papers*	7,500 papers*	7,700 papers*
• Supply Chain (Carbonneau et al. 2008) • Fashion (Ren et al. 2017) • Spare parts (Hua and Zhang 2006) • Smart grids (Muralitharan et al. 2018) • …	• Predictive maintenance (Susto et al. 2015) • Credit risk assessment (Twala 2010) • Intrusion detection (Tsai et al. 2009) • Recommender systems (Zhang et al. 2019) • …	• Robotics (Levine et al. 2016) • Advertising (Jin et al. 2018) • Plant control (Lazic et al. 2018) • Job scheduling (Priore et al. 2006) • Chemistry R&D (Zhou et al. 2019) • …	• Chat bots (Xu et al. 2017) • Warehouse inventory control (Xu et al. 2018) • Legal document analysis (Ashley and Walker 2013) • …

Table 1: Exemplary excerpts of research in ML applications for specific business use cases (Natural language processing: NLP, Image recognition: IR). The number of papers* is calculated by the number of peer-reviewed articles that apply to the query terms of the header, "ML"/"AI" and "enterprise application" in the semantic scholar database.

In a next step, we studied the current state of actual adoption of ML technologies in companies. For this purpose, we reviewed studies of the recent years that provide an overview of the implementation of ML technologies (opposed to Table , where we show the existence of a theoretical background). Our literature review shows that no considerable peer-review literature is available that answers the question to what extent these ML technologies are actually employed in companies. This emphasizes the need for further research in this field.

In order to also incorporate grey literature studies, we imposed the following restrictions on our search in order to filter considerable quality studies only: A. The study was conducted from 2018 to January 2020, B. The data foundation is documented: Number of respondents by company size, industry and ML implementation maturity level, C. The study is published by a renowned organization or company.

We will first summarize the findings of studies that incorporate both SMEs and larger companies (Set I). In the second part of this section, we narrow down the focus on studies that differentiate between companies of different sizes (Set II).

Set I comprises ten studies with an average number of respondents of 2,600 each (minimum 200; maximum 11,400) mostly from the Americas, Europe and Asia. The studies do not differentiate between company size.

About 20% of the companies interviewed in Set I confirm that they use an ML technology implementation in their planning, control or operational processes. The overall maturity level of ML technology implementations is low – about 25% are in early stages and are either gathering first experiences in ML technologies or are about to implement technologies. On average, the studies indicate that about 30% of the companies that have not yet implemented ML technologies are intending to do so soon. The remaining companies do not have specific plans yet to implement ML technologies or are investigating potential use cases. This intention is confirmed by the finding that about 30% of annual IT budgets were dedicated to the implementation of use cases where ML technologies were supposed to be deployed.

The studies of Set I also examine major challenges for companies when investigating potential use cases for machine learning and during implementation. The most frequent challenges are:

1. the lack of sufficient employees with ML/AI know how (by far most frequently),
2. limited budget and other (non-ML) projects competing for funds,
3. difficulties to identify positive business cases,
4. too little acceptance for ML on a managerial level.

(Chui and Malhotra 2018; Lorica and Paco 2019, 2018; Loucks et al. 2018; Ransbotham et al. 2017; Ransbotham et al. 2018; Stancombe et al. 2017; Howard and Rowsell-Jones 2019; Teradata 2017; Algorithmia 2018)

We conclude that the overall prevalence of ML technologies is still medium, however great interest in the technologies exists and applications are being evaluated by companies. Companies with a higher ML maturity level apply more advanced ML techniques than others – as assumed in RQ1.

> **Finding II***: ML technologies are already established in business applications and interest in the technologies is high. Yet, the prevalence of the technologies is medium and the technologies applied are mostly of medium complexity also – however some companies already employ very advanced techniques.*

Set II, in contrast to Set I, comprises seven studies with an average number of respondents of 900 each (minimum 190; maximum 3,100), in part with a focus on North America and Europe. Using Set II, we highlight the differences of the studies' insights with respect to company size.

Compared to Set I, where about 20% of companies interviewed currently employ ML technologies and 25% are in the process of evaluation or implementation, Set II reveals a more differentiated picture. The studies in Set II confirm the data from Set I for companies with 500 or more employees. However, companies with 500 employees maximum exhibit significantly lower maturity: Only 8% of the companies have already deployed ML technologies and only 20% are evaluating ML technologies for business applications. However, other studies that also incorporate businesses that sell ML technologies as distinct products (ML consultancy, tools, …). Here, small companies and especially start-ups demonstrate their competencies and exhibit a ML maturity level equal to larger companies. Therefore, we consider companies that do not primarily apply ML to optimize internal processes or to enhance traditional products as a different company type of “tech start-ups”.

The studies in Set II also report different challenges for companies concerning the implementation of ML technologies:

1. too little acceptance for ML amongst users and operatives,
2. data privacy concerns (e.g. violation of GDPR regulations),
3. the lack of sufficient employees with ML/AI know how,
4. too little acceptance for ML on a managerial level.

(Spiceworks 2020; Böttcher et al. 2018; Abel-Koch et al. 2019; Reder 2018, 2019; Algorithmia 2018, 2019)

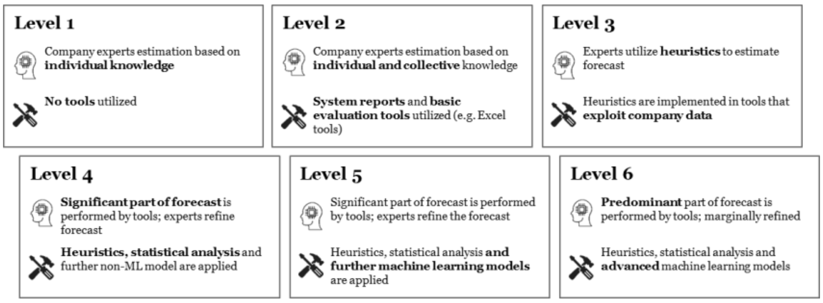
We conclude that, referring to RQ1, the prevalence of ML technology applications in SME is significantly lower than in the overall industry. Moreover, SME state elementary different challenges than larger companies: They struggle more with entry-barriers whereas larger companies typically rather have difficulties to scale their ambitions to apply ML-technologies to the resources available.

> **Finding III***: Small and medium businesses are significantly less likely to have ML technologies deployed yet. Their present challenges differ from larger businesses and reflect their lower ML maturity: little acceptance for ML both amongst users and operatives as well as managers and limited ML know how.*

2.3 Survey Methodology and Insights

For our own survey, we interviewed 18 CXOs and Managing Directors of small, medium and large-sized companies. The interview involved (i) a self-assessment of the maturity level of the state of implementation of ML in the company (see Figure 2) and (ii) the respondents' assessment of challenges and success factors encountered in previous ML implementation projects and anticipated for future ML implementation projects (see Figure 3).

Figure 4 and Figure 5 give an overview of the interviewees of the survey. Interviewees are employed at companies of all sizes like tech-startups (less than 35 employees), small businesses (less than 500 employees), medium businesses (less than 1,000 employees) and large business (10,000 to 30,000 employees). The number of interviewees is relatively evenly distributed over all company sizes (see Figure 4). From the companies interviewed, 39% have reached a maturity level where ML is actually implemented in day-to-day operations or their products (levels 5 and 6). 39% of the companies do not use ML technologies yet but employ heuristics or statistical analysis to leverage internal processes or enhance their products (levels 3 and 4). The remaining 22% state that neither ML technologies nor heuristics or statistical analysis are employed at their company (levels 1 and 2). However, all companies report that they are actively evaluating use cases for ML applications at the moment of the survey (see Figure 5).



Level 1

Company experts estimation based on **individual knowledge**

No tools utilized

Level 2

Company experts estimation based on **individual and collective** knowledge

System reports and **basic evaluation tools** utilized (e.g. Excel tools)

Level 3

Experts utilize **heuristics** to estimate forecast

Heuristics are implemented in tools that **exploit company data**

Level 4

Significant part of forecast is performed by tools; experts refine forecast

Heuristics, statistical analysis and further non-ML model are applied

Level 5

Significant part of forecast is performed by tools; experts refine the forecast

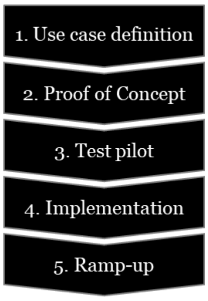
Heuristics, statistical analysis **and further machine learning models** are applied

Level 6

Predominant part of forecast is performed by tools; marginally refined

Heuristics, statistical analysis and **advanced** machine learning models

Figure 2: Six levels of ML implementation maturity for the respondents' self-assessment



1. Use case definition

2. Proof of Concept

3. Test pilot

4. Implementation

5. Ramp-up

Figure 3: Five exemplary project phases during ML implementation

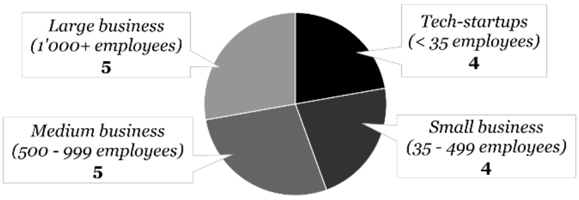


Figure 4: Respondents by company size

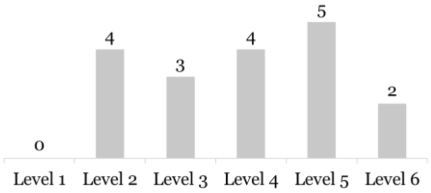


Figure 5: Respondents by ML-maturity level

Following the approach of Mayring, we conducted a qualitative analysis of the interviewees' replies to our survey. The answers were categorized and correlated to the company's characteristics: such as company size, ownership structure of the company and ML maturity level. In addition, we clustered the replies by their relationship to company size and maturity by qualitative assessment (Mayring 2000).

From the analysis, we summarize the following main statements:

A. There is a clear dependency between the ML maturity level and the company size. SME are much less matured than large business. The only exception to this finding are the tech startups that are extremely small but see ML technologies as their main product – they are the most ML mature companies of the survey.
B. Small businesses (SB) are currently still struggling to identify use cases for ML applications (process stage 1)
   - *ML know how*: SB possess too little knowledge and experience to assess ML suitable use cases.
   - *Personnel capacities*: Too few personnel capacities to systematically advance use cases from the use case identification to a first proof of concept. SB strongly rely on the personal initiative of single employees to investigate ML use cases.
   - *Data availability*: SB report to lack the quantity of records required to train ML algorithms.
   - *Acceptance of ML technologies*: Limited knowledge of ML results in concerns about ML technologies amongst operatives and management. However, SB can benefit from flat hierarchy and from a determined management that encourages employees to advance in ML technologies.
   - *Interdisciplinarity*: ML initiatives lay in the hands of individual persons in the company which combine data science, domain and IT expertise in one person with no interfaces.

  - o *External partners*: Dissent amongst SB whether to consult external partners to advance in ML applications. However, those open to external partners, all prefer software providers that specialized in domain specific solutions.

C. Medium businesses (MB) have typically passed the use case definition phase.
  - o *ML know how*: MB asses their ML know how to be sufficient for the definition of ML use cases. However, MB are not able to independently implement the technology with internal knowledge.
  - o *Personnel capacities*: Rely on personal initiative of employees to advance ML use cases. Regard use cases as pilots to prove the worthiness of technology and are hesitant to dedicate resources to it.
  - o *Data availability*: Have accrued sufficient data records but typically not in a standardized form – they are not used to data driven approaches in their usual operations.
  - o *Acceptance of ML technologies*: Successful, where ML strategy is installed by their management. The management identifies ML as an important impact to their business. The concerns towards ML are more distinguished than in small businesses: Lacking "explainability" and transparency of ML algorithms as well as the thread of the substitution of labor by algorithms are the main concerns of operatives. However, there is also a positive perspective that ML can help to focus on more valuable work by having simple and repetitive tasks executed by ML algorithms.
  - o *Interdisciplinarity*: MB face the problem of interdisciplinarity more than SB. Data science, domain knowledge and IT are represented by separate persons and departments. However, in contrast to LB, interdisciplinary collaboration is less supported by frameworks and standardized processes and rely more on the individual experience and ability for collaboration.
  - o *External partners*: Open to external cooperation. Research projects with universities and business schools preferred as aim to build up ML know how in the course of the cooperation.

D. Large businesses (LB) have experience in the field of ML applications and implemented use cases.
  - o *ML know how*: LB have specialized data science departments that are able to implement ML technologies fully internally.
  - o *Personnel capacities*: Due to the AI strategies employed, resources are actively dedicated to the identification and implementation of ML technologies. However, LB find it challenging to employ enough data science experts to realize the identified use cases. They exhibit that the demand for data science experts exceeds the market supply of free experts at this point of time.
  - o *Data availability*: LB systematically gather and record business data and have identified data as important advantage.
  - o *Acceptance of ML technologies*: LB consistently state that they consider ML applications as a vital part of their business strategy. ML is widely accepted due to positive experience in practice.
  - o *Interdisciplinarity*: Strong division of labor and specialization complicates the interdisciplinary cooperation within LB. LB face the challenge that the specialists "do not speak the same language". However, LB also report that the outcome of ML implementation projects heavily depends on a successful exchange between data science, process owners and IT department.
  - o *External partners*: External partners do not play a significant role to LB as they already possess the necessary ML know how. Moreover, third party tools are often restricted by governance policies and difficult to integrate into existing systems.

E. Tech start-ups (TS) exhibit the highest maturity levels of all companies of the survey and a high specialization on distinct industries.

- *ML know how*: TS excel in ML technologies and apply advanced techniques.
- *Personnel capacities*: TS consider ML know how as a primary asset. Therefore, they concentrate to establish attractive work environment for ML experts.
- *Data availability*: With specialization, TS develop interfaces optimized for their customers industry and requirements, such that the available data can be efficiently used.
- *Acceptance of ML technologies*: TS mainly generate acceptance by show casing successful implementations from previous projects.
- *Interdisciplinarity*: Experts in small teams combine ML, domain and IT know how.
- *External partners*: External partners do not play a relevant role.

In general, the survey results are in line with the findings from the previous section: First, ML is already established in enterprises, however only to a medium degree of prevalence and maturity (see Finding II). Second, SME lag behind larger enterprises with respect of prevalence and maturity of the application of ML and exhibit challenges that are specific to SME (see Finding III).

In addition, we summarize the following main challenges and success factors as addressed in RQ2 for small and medium businesses:

> **Finding IV***: Company size and ML maturity are strongly dependent: Larger businesses are more mature than smaller businesses.*
>
> **Finding V***: Primary challenges to small businesses are basic understanding of ML capabilities for use case definition and the availability of data. Primary success factors are flat hierarchies and a determined management which supports and encourages committed employees as well as external partners with the appropriate domain knowledge.*
>
> **Finding VI***: Primary challenges to medium businesses are ML implementation know how and the increasing issues of interdisciplinary collaboration. Primary success factors are an external research cooperation and a pronounced ML strategy by the management that provides the necessary support for committed and volunteering employees.*

Finding IV raises the question whether the relationship between maturity and company size is causally determined – which we address in the following section.

2.4 Size and Maturity Related Challenges and Success Factors

In the previous section, we observed the correlation between company size and ML maturity (see Finding IV). One could therefore conclude that there is one path of development that all companies follow – and larger companies might have just already progressed further than smaller companies. However, this would be a misconception, as we will show based on the survey. According to the respondents' replies, challenges are allocated to their relationship to company size and maturity (see Table 2).

	Maturity related	Size related
Challenges	• Data availability and quality (-) • Lack of ML know how (-) • Insufficient ML results (-) • Governance policies (+) • Acceptance of ML (-)	• Lack of personnel capacities (-) • Dedicated ML experts (-) • Insufficient input data (-) • Division of labor and specialization (+) • Computation power in-house (-)
Success factors	• Existing ML know how (+) • Standardized data interfaces (+) • External partners (-) • Small steps and early success stories (-)	• Existing business intelligence / data science team (+) • Commitment of individual employees (-) • Good interdisciplinary collaboration (+) • Fast decision-making (+)

Table 2: Overview of maturity and size related challenges and success factors. (+/-) denotes a positive or negative correlation: Importance increases / decreases with increasing maturity or size respectively.

Table 2 shows several relevant challenges that are mainly correlated to company size and cannot or only marginally be compensated by increasing ML experience from ML maturity. Especially determining is the issue of sufficiently trained personnel: SB typically do not have the economies of scale to afford to employ data scientists dedicated to the implementation of ML technologies. Therefore, they cannot go beyond the first stages of the process of ML implementation (cf. Finding V and Finding VI) accounting to RQ2.

> **Finding VII**: *Challenges can be company size related and not or only marginally influenced by ML maturity. Such challenges prevent companies of different size from undergoing the same maturity development process. Therefore, SME should use different approaches in ML projects.*

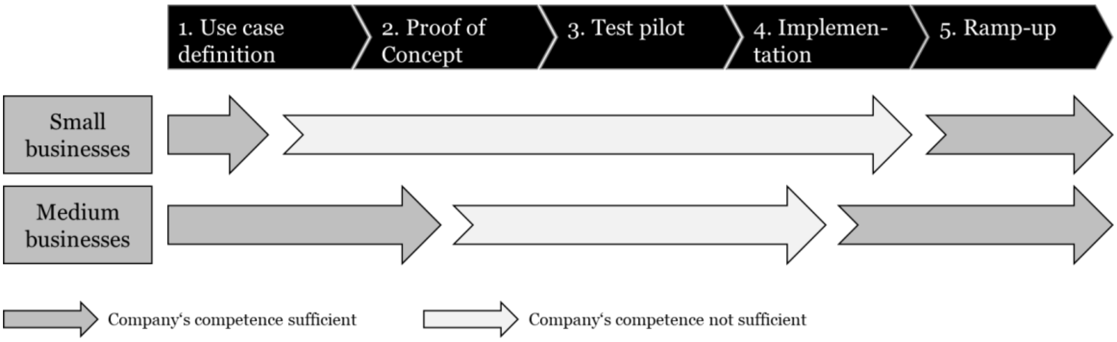


Figure 6: Company size related challenges lead to differences in ML implementation competences.

2.5 Approaches to Close the Gap for SME

From Finding VII we conclude that SME can benefit from specific approaches in ML projects. As shown before, challenges for SB start already during the use case definition phase and maintain during all phases that require deeper ML know how. MB typically have more ML know how, but still face challenges during proof of concept, testing and implementation (see Finding VI). In the following, we use the previous findings to briefly discuss general measures that facilitate the access of SME to ML technologies. We then discuss a particular framework for demand forecasting in SME.

2.5.1 General measures to facilitate ML in SME

In the following, we will address facilitating conditions as mentioned in RQ2. According to Finding V and Finding VI, especially SB but also MB face their initial challenge during the use case definition phase: Assessment of the applicability of ML in particular use cases. Interviewees report three ways to pass this hurdle: The exchange with other companies that already passed this step, support by consultancies or software service providers as well as research cooperations. The interviewees agree that a crucial factor is the domain or industry specific knowledge combined with ML experience. We conclude that SB can benefit most if suitable products are available that match their use case requirements. In this case, the product can fulfill their needs without the need to deeply understand ML technologies. MB can benefit most if they cooperate in a research project with an external partner with the appropriate ML know how. This way, the partner can elaborate a use case specific solution and the company can build up ML know how internally during the project. Both approaches also mitigate the challenge of interdisciplinary collaboration between data science, which arises already in MB.

> **Finding VIII**: *Small and medium businesses can mitigate the lack of ML know how in ML use case definition and implementation through external partners. The survey results suggest software providers with market-ready products for small and research cooperations for medium businesses.*

Based on this finding, we encourage universities and research faculties to enter cooperation with MB to develop further ML applications together. We also suggest that politics and governments should support the funding of such research cooperations.

In Finding V and Finding VI we also showed that success in ML implementation projects strongly depends on the personal initiative and interest of employees as well as short decision-making processes. We propose that the companies' management should actively encourage employees by creating favorable conditions: Allow for advanced training of employees, establishment of ML labs with the necessary hardware equipment and interdisciplinary workshops for interested employees for use case definition.

> **Finding IX**: *Personal initiative of employees is found to be crucial for the success of ML projects in SME and should be fostered by the management of companies by favorable conditions. Trainings, equipment, interdisciplinary work and short decision-making processes are proposed.*

2.5.2 A suitable framework as entry point to ML applications

The measures described above address the challenges of SME on an entrepreneur and governmental level. In addition to this, we also consider the contribution of research and the IS community to the issue. The survey shows that SME require ML technologies of a confined complexity that can be implemented with the limited experience and knowledge of SMEs. ML frameworks with auto-hyperparameter tuning (AutoML) exist from various research projects (e.g. Auto-WEKA, Thornton et al. 2013; Auto-sklearn, Feurer et al. 2019; TPOT, Olson and Moore 2019;

Auto-keras, Jin et al. 2019) and vendors (e.g. AzureML, Uber Ludwig; Google Cloud AutoML). The specific advantages and disadvantages have been investigated in several studies and competitions, showing that AutoML frameworks can achieve good results compared to instance specific implementations and are significantly easier to manage (He et al. 2019; Truong et al. 2019; Guyon et al. 2019).

> **Finding X**: *AutoML frameworks encapsulate and automate major parts of the ML implementation and optimization process. Companies can use these frameworks to speed up ML projects, if they possess limited ML implementation know how and if their problem instance can be solved using generic optimization strategies.*

Apart from these general auto-tuning frameworks, our literature research shows that no relevant studies exist that systematically address the issues and requirements of SME in the application of ML technologies. In this context, the survey shows two areas where SME could benefit most from research: Comprehendible use cases and suitable ML applications and SME specific frameworks that can be applied with limited ML know how.

2.6 Conclusion and Outlook

In this work, we raised two questions: 1. What is the gap of SME in ML adoption compared to the state of the art and 2. what challenges and success factors are typical for SME in the ML adoption process. We find that research provides a strong theoretical foundation, but practice is yet in the process of the adoption of ML technologies and that SME significantly stay behind larger companies. We observe that larger companies are generally more mature in the adoption of ML, and that size-specific factors prevent SME from taking the same path of ML knowledge development as larger businesses.

We also identified the major challenges of SME in the adoption of ML: Insufficient ML know how in SME for the identification of use cases and implementations, poor data quality in small businesses and obstacles in interdisciplinary work in medium businesses. We find that external cooperations were observed as major success factors to overcome the challenges, as well as personal initiative of employees. We propose three concrete measures to facilitate ML in SME.

This study reflects the current situation of companies interviewed in our survey. The situation may change in the next years. However, while large businesses are systematically progressing in ML applications, SME risk to fall behind. Research can contribute to further facilitate the access of SME to ML technologies by appropriate frameworks that reduce the need for technical knowledge and that are adopted to the requirements of SME. As shown in this study, the prevalence of such frameworks is too low, yet.

Also, we are aware that the survey was conducted over a relatively small set of companies. Therefore, we can only deduct qualitative statements. However, the findings are in line with surveys that involve a higher number of respondents and logically sound. A larger survey could show the statistical significance of the statements.

2.7 References

Abel-Koch, Jennifer; Al Obaidi, Leath; El Kasmi, Sabrina; Acevedo, Miguel Fernández; Morin, Letitia; Topczewska, Anna (2019): GOING DIGITAL The Challenges Facing European SMEs. European SME Survey 2019. KfW Bankengruppe (KfW).

Algorithmia (2018): The State of Enterprise Machine Learning. Algorithmia research.

Algorithmia (2019): 2020 State of Enterprise Machine Learning. Algorithmia research.

Ashley, Kevin D.; Walker, Vern R. (2013): Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In Bart Verheij (Ed.): Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. the Fourteenth International Conference. Rome, Italy, 6/10/2013 - 6/14/2013. New York, NY: ACM (ACM Digital Library), p. 176.

Bauer, Markus; van Dinther, Clemens; Kiefer, Daniel (2020): Machine learning in SME: an empirical study on enablers and success factors. In Association for Information Systems (Ed.): AMCIS 2020 proceedings: Association for Information Systems, pp. 1–10.

Böttcher, Björn; Velten, Carlo; Schwalm, Anna-Lena (2018): Machine Learning in deutschen Unternehmen. Eine empirische Studie zu Betrieb und Anwendung von Künstlicher Intelligenz. Crisp Research AG. Kassel.

Carbonneau, Real; Laframboise, Kevin; Vahidov, Rustam (2008): Application of machine learning techniques for supply chain demand forecasting. In *European Journal of Operational Research* 184 (3), pp. 1140–1154. DOI: 10.1016/j.ejor.2006.12.004.

Chui, Michael; Malhotra, Sankalp (2018): AI adoption advances, but foundational barriers remain. McKinsey Global Institute. San Francisco. Available online at https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain.

Feurer, Matthias; Klein, Aaron; Eggensperger, Katharina; Springenberg, Jost Tobias; Blum, Manuel; Hutter, Frank (2019): Auto-sklearn: Efficient and Robust Automated Machine Learning. In Frank Hutter, Lars Kotthoff, Joaquin Vanschoren (Eds.): Automated Machine Learning, vol. 45. Cham: Springer International Publishing (The Springer Series on Challenges in Machine Learning), pp. 113–134.

Guyon, Isabelle; Sun-Hosoya, Lisheng; Boullé, Marc; Escalante, Hugo Jair; Escalera, Sergio; Liu, Zhengying et al. (2019): Analysis of the AutoML Challenge Series 2015–2018. In Frank Hutter, Lars Kotthoff, Joaquin Vanschoren (Eds.): Automated Machine Learning, vol. 35. Cham: Springer International Publishing (The Springer Series on Challenges in Machine Learning), pp. 177–219.

He, Xin; Zhao, Kaiyong; Chu, Xiaowen (2019): AutoML: A Survey of the State-of-the-Art. Available online at http://arxiv.org/pdf/1908.00709v3.

Howard, Chris; Rowsell-Jones, Andy (2019): 2019 CIO Survey: CIOs Have Awoken to the Importance of AI. Gartner. Available online at https://www.gartner.com/doc/reprints?id=1-1W8H2L3G&ct=191017&st=sb.

Hua, Zhongsheng; Zhang, Bin (2006): A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. In *Applied Mathematics and Computation* 181 (2), pp. 1035–1048. DOI: 10.1016/j.amc.2006.01.064.

Jin, Haifeng; Song, Qingquan; Hu, Xia (2019): Auto-Keras: An Efficient Neural Architecture Search System. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, George Karypis (Eds.): KDD2019. Anchorage, Alaska, USA. the 25th ACM SIGKDD International Conference. Anchorage, AK, USA, 8/4/2019 - 8/8/2019. New York, NY: Association for Computing Machinery, pp. 1946–1956.

Jin, Junqi; Song, Chengru; Li, Han; Gai, Kun; Wang, Jun; Zhang, Weinan (2018): Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising. In Alfredo Cuzzocrea (Ed.): Proceedings of the 27th ACM International Conference on Information and Knowledge Management. the 27th ACM International Conference. Torino, Italy, 10/22/2018 - 10/26/2018. [Place of publication not identified]: ACM, pp. 2193–2201.

Lazic, Nevena; Boutilier, Craig; Lu, Tyler; Wong, Eehern; Roy, Binz; Ryu, M. K.; Imwalle, Greg (2018): Data center cooling using model-predictive control. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.): Advances in Neural Information Processing Systems 31: Curran Associates, Inc, pp. 3814–3823. Available online at http://papers.nips.cc/paper/7638-data-center-cooling-using-model-predictive-control.pdf.

Levine, Sergey; Finn, Chelsea; Darrell, Trevor; Abbeel, Pieter (2016): End-to-End Training of Deep Visuomotor Policies. In *J. Mach. Learn. Res.* 17 (1), pp. 1334–1373.

Levy, Yair; Ellis, Timothy J. (2006): A systems approach to conduct an effective literature review in support of information systems research. In *Informing Science* 9.

Lorica, Ben; Paco, Nathan (2018): The state of machine learning adoption in the enterprise. First edition. Sebastopol, CA: O'Reilly Media. Available online at https://www.bastagroup.nl/wp-content/uploads/2019/01/the-state-of-machine-learning-adoption-in-the-enterprise.pdf.

Lorica, Ben; Paco, Nathan (2019): AI Adoption in the Enterprise. How Companies Are Planning and Prioritizing AI Projects in Practice: O'Reilly Media.

Loucks, Jeff; Davenport, Tom; Schatsky, David (2018): State of AI in the Enterprise, 2nd Edition. Early adopters combine bullish enthusiasm with strategic investments. Deloitte Insights. Available online at https://www2.deloitte.com/content/dam/insights/us/articles/4780_State-of-AI-in-the-enterprise/DI_State-of-AI-in-the-enterprise-2nd-ed.pdf.

Mayring, Philipp (2000): Qualitative Content Analysis. Forum Qualitative Sozialforschung. In *Forum: Qualitative Social Research* Vol 1, No 2. DOI: 10.17169/FQS-1.2.1089.

Mitchell, Tom Michael (1997): Machine learning. International ed. New York, NY: McGraw-Hill (McGraw-Hill series in computer science).

Muralitharan, K.; Sakthivel, R.; Vishnuvarthan, R. (2018): Neural network based optimization approach for energy demand prediction in smart grid. In *Neurocomputing* 273, pp. 199–208. DOI: 10.1016/j.neucom.2017.08.017.

OECD (2004): ICT, E-Business and Small and Medium Enterprises. OECD Publishing. Paris (OECD Digital Economy Papers, 86).

OECD (2015): Data-Driven Innovation: Big Data for Growth and Well-Being. Paris: OECD.

Olson, Randal S.; Moore, Jason H. (2019): TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In Frank Hutter, Lars Kotthoff, Joaquin Vanschoren (Eds.): Automated Machine Learning, vol. 17. Cham: Springer International Publishing (The Springer Series on Challenges in Machine Learning), pp. 151–160.

Priore, Paolo; La Fuente, David de; Puente, Javier; Parreño, José (2006): A comparison of machine-learning algorithms for dynamic scheduling of flexible manufacturing systems. In *Engineering Applications of Artificial Intelligence* 19 (3), pp. 247–255. DOI: 10.1016/j.engappai.2005.09.009.

Ransbotham, Sam; Kiron, David; Gerbert, Philipp; Reeves, Martin (2017): Reshaping Business With Artificial Intelligence. Closing the Gap Between Ambition and Action. MITSloan Management Review and The Boston Consulting Group. Available online at http://image-src.bcg.com/Images/Reshaping%20Business%20with%20Artificial%20Intelligence_tcm9-177882.pdf.

Ransbotham, Sam; Reeves, Martin; Gerbert, Philipp; Kiron, David; Spira, Michael (2018): Artificial Intelligence in Business Gets Real. Pioneering Companies Aim for AI at Scale. MITSloan Management Review and The Boston Consulting Group.

Reder, Bernd (2018): Studie Machine Learning / Deep Learning 2018. IDG Research Services. München.

Reder, Bernd (2019): Studie Machine Learning / Deep Learning 2019. IDG Research Services. München.

Ren, Shuyun; Chan, Hau-Ling; Ram, Pratibha (2017): A Comparative Study on Fashion Demand Forecasting Models with Multiple Sources of Uncertainty. In *Ann Oper Res* 257 (1-2), pp. 335–355. DOI: 10.1007/s10479-016-2204-6.

Spiceworks (2020): The 2020 State of IT. The annual report on IT budgets and tech trends. Available online at https://www.spiceworks.com/marketing/state-of-it/report/.

Stancombe, Christopher; Tolido, Ron; Thieullent, Anne-Laure; Buvat, Jerome; KVJ, Subrahmanyam; Khadikar, Amol; Chandna, Apporva (2017): Turning AI into concrete value: the successful implementers' toolkit. The Digital Transformation Institute.

Susto, Gian Antonio; Schirru, Andrea; Pampuri, Simone; McLoone, Sean; Beghi, Alessandro (2015): Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. In *IEEE Trans. Ind. Inf.* 11 (3), pp. 812–820. DOI: 10.1109/TII.2014.2349359.

Teradata (2017): State of Artificial Intelligence for Enterprises. Teradata.

Thornton, Chris; Hutter, Frank; Hoos, Holger H.; Leyton-Brown, Kevin (2013): Auto-WEKA. In Robert L. Grossman, Ramasamy Uthurusamy, Inderjit Dhillon, Yehuda Koren (Eds.): KDD '13. The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining : August 11-14, 2013, Chicago, Illinois, USA. the 19th ACM SIGKDD international conference. Chicago, Illinois, USA, 8/11/2013 - 8/14/2013. New York: ACM, p. 847.

Truong, Anh; Walters, Austin; Goodsitt, Jeremy; Hines, Keegan; Bruss, C. Bayan; Farivar, Reza (2019): Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. Available online at http://arxiv.org/pdf/1908.05557v2.

Tsai, Chih-Fong; Hsu, Yu-Feng; Lin, Chia-Ying; Lin, Wei-Yang (2009): Intrusion detection by machine learning: A review. In *Expert Systems with Applications* 36 (10), pp. 11994–12000. DOI: 10.1016/j.eswa.2009.05.029.

Twala, Bhekisipho (2010): Multiple classifier application to credit risk assessment. In *Expert Systems with Applications* 37 (4), pp. 3326–3336. DOI: 10.1016/j.eswa.2009.10.018.

Webster, Jane; Watson, Richard T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In *MIS quarterly* 26.

Xu, Anbang; Liu, Zhe; Guo, Yufan; Sinha, Vibha; Akkiraju, Rama (2017): A New Chatbot for Customer Service on Social Media. In Gloria Mark, Susan Fussell, Cliff Lampe, m.c. schraefel, Juan Pablo Hourcade, Caroline Appert, Daniel Wigdor (Eds.): CHI'17. Proceedings of the 2017 ACM SIGCHI Conference on Human Factors in Computing Systems, May 6-11, 2017, Denver, CO, USA. the 2017 CHI Conference. Denver, Colorado, USA, 5/6/2017 - 5/11/2017. CHI 2017; Annual CHI Conference on Human Factors in Computing Systems. New York, NY: ACM, pp. 3506–3510.

Xu, Lichao; Kamat, Vineet R.; Menassa, Carol C. (2018): Automatic extraction of 1D barcodes from video scans for drone-assisted inventory management in warehousing applications. In *International Journal of Logistics Research and Applications* 21 (3), pp. 243–258. DOI: 10.1080/13675567.2017.1393505.

Zhang, Shuai; Yao, Lina; Sun, Aixin; Tay, Yi (2019): Deep Learning Based Recommender System. In *ACM Comput. Surv.* 52 (1), pp. 1–38. DOI: 10.1145/3285029.

Zhou, Zhenpeng; Kearnes, Steven; Li, Li; Zare, Richard N.; Riley, Patrick (2019): Optimization of Molecules via Deep Reinforcement Learning. In *Scientific reports* 9 (1), p. 10752. DOI: 10.1038/s41598-019-47148-x.

3 How the Demand Forecasting Literature and Applications can Benefit from Better Comparability

[*This chapter corresponds to an article submitted to the International Journal of Information Technology & Decision Making (IJITDM). It is currently in second revision status in the acceptance process at IJITDM:* Bauer, Markus; van Dinther, Clemens (2023): Quantitative Methods in Demand Forecasting: A Survey of the State-of-the-Art. – *further referred to as* Bauer and van Dinther 2023]

Quantitative Methods in Demand Forecasting: A Survey of the State-of-the-Art

Demand forecasting is an essential element of business planning and management, and an accurate estimate of future demand helps companies improve their resource allocation. In recent years, researchers have developed and evaluated numerous quantitative models. The advent and development of machine learning methods have driven innovation in demand forecasting. However, as the number of studies increases, it becomes increasingly difficult to provide an overview of the state of the art and recent developments in demand forecasting. This paper helps to provide an overview of the broad research field of demand forecasting. Based on 116 recent studies, this article structures application fields and a general process for demand forecasting. Systematically, this article compares methods and procedures used and identifies open questions in research. As the most significant contribution, this article also presents a novel systematic for describing empirical studies in a standardized way to establish comparability to other studies. This improved comparability promotes future meta-studies: without the effort of tedious empirical experiments, approaches of different studies can thus be compared quantitatively.

3.1 Introduction and Problem Description

Demand forecasting has a long tradition in academic research. In 1922, Prescott already described how to predict the development of the demand for automobiles, railroad services, and tobacco in a comparative quantitative study (Prescott 1922). Since then, the need for precise forecasting of demand has rapidly increased throughout all industries.

Academia has identified this need and produced thousands of studies in which researchers have developed specialized approaches for numerous distinct use cases. The research was fueled by advancements in data processing techniques and forecasting algorithms from the full range of academia and further boosted by the advent of machine learning in the 2000s.

While this evolution provides valuable contributions to the research community and practitioners, it makes it difficult for both novice and experienced researchers to keep track of the body of knowledge and recent developments.

Moreover, no unified and fully comprehensive descriptive approach to empirical studies for demand forecasting has yet evolved. Authors focus on the major innovations of their approaches, but for the most part do not describe the full process, data used, and results in a way that allows other researchers to fully compare the approaches with other studies. In this context, however, a high degree of comparability helps ensure that comparative studies can build on existing ones without having to replicate the empirical experiments themselves. The reduced effort and increased speed provided by such quantitative comparative ones could accelerate progress in research.

To address these challenges, we derive the following research questions to structure the research:

RQ 1. What are the main applications of demand forecasting in the research and how do they contribute to the development of the discipline?
RQ 2. How can the process of demand forecasting in the literature be abstracted and structured to yield a general framework for the comparison of different approaches?
RQ 3. What are the current state-of-the-art methods per process step in the different demand forecasting disciplines?
RQ 4. Where are the gaps between the state-of-the-art methods applied when comparing the demand forecasting disciplines?

This study contributes to the discussion in the demand forecasting community following the principles of Webster and Watson for systematic literature reviewing (Webster and Watson 2002) whereby the survey:

- *structures* the available *research*, making it easier for both newcomers and senior scholars to review the body of knowledge and classify methods and studies.
- *summarizes* the *state-of-the-art* methods and approaches of *demand forecasting* as well as related disciplines of forecasting and clusters them into *concepts*.
- highlights the *gaps* between the state-of-the-art in the application of demand forecasting and other forecasting disciplines and summarizes as well as highlights *open questions* for future research.
- *develops and introduces a systematic* by which future studies can describe data, objectives, methods and procedures used, and results in a standardized manner–to increase comparability to other studies.

The remainder of this paper is structured as follows. In Section 3.2, we introduce the applied research methodology. Section 3.3 introduces the fundamentals of demand forecasting and provides an initial overview of the literature of demand forecasting and key surveys. In Sections 3.5-

3.8, the study systematically reviews the different approaches per step in the process of demand forecasting, comparing different demand forecasting disciplines. In Section 3.9 we summarize and discuss open issues identified in the field of demand forecasting before we conclude our findings in Section 3.10. The appendix in Section 3.11 provides a comprehensive overview of all reviewed studies.

3.2 Research Methodology

This section describes our approach to structuring the literature on demand forecasting for this article whereby Figure 7 provides a schematic visualization of the process.

Initially, we defined our research questions as outlined in Section 3.1 (Step I in Figure 7).

Throughout the survey, we employed a comprehensive database of studies that relate to the field of demand forecasting. The initial database was set up using the *Microsoft Academic* search engine. After the discontinuation of the service in December 2021, we consolidated the existing database with additional query results from the search engines *Semantic Scholar* and *CrossRef*, while taking care to omit any duplicates. We applied the same query term “demand AND forecasting” to all search engines. To export the search engine results into the database, we used the tool *Publish or Perish* (Harzing 2007).

To narrow down the scope, we only considered studies with more than five citations on average per year and studies that were published in the period from 2018 to 2022. This initial database of studies included 1,199 studies (Step II in Figure 7). We deliberately deviated from the suggestion of Levy and Ellis to concentrate on ranked Information Systems (IS) journal and conference papers only, as we find numerous studies to be relevant that were published in other journals (Levy and Ellis 2006). Instead, we prioritized and classified the studies in the database as explained in more detail later on.

After the setup of the comprehensive database, we screened all studies by title and abstract for their actual relevance and excluded non-relevant articles. We consider a study to be relevant when (1) its focus is to (2) apply quantitative methods to predict (3) measurable demands of (4) products or services based (5) directly or indirectly on customer behavior for (6) future (i.e., unknown and uncertain) time periods and (7) evaluates the deviation of the prediction from the realized value. We also classified the articles as surveys (if their main purpose was the comparison of other studies) or actual studies that proposed novel approaches. In addition, we classified each study according to its field of application (Step III in Figure 7). After all the studies were classified, we consolidated the field of applications into five main domains of applications, which will be introduced in Section 3.3.2 (Step IV in Figure 7).

Based on the existing surveys of demand forecasting that were identified, we followed the road map of Webster and Watson and evaluated the surveys to identify the main contributions and recently identified open issues (Webster and Watson 2002) whereby Table 3 summarizes the existing surveys.

Following the recommendations of Webster and Watson, we applied a backward and forward search approach using the citations from the existing literature reviews to add relevant studies to the database. Again, the selection of relevant studies was based on the same criteria as described above. After the forward and backward search, the database comprised 116 studies in total (Step V in Figure 7).

To further structure the survey, we analyzed the process of demand forecasting as described by the studies. As only some studies explicitly describe the process, we also examined forecasting processes in other fields of research and synthesized a general abstract process as described in Section 3.4 (Steps VI and VII in Figure 7).

We subsequently reviewed all studies in the database in detail and classified the approaches that were applied in each process step (Step VIII in Figure 7). To compare the observed approaches to the respective state-of-the-art in other fields of research, we also reviewed surveys in other forecasting disciplines (Step IX in Figure 7).

Table 17 in Section 3.11 „Appendix” provides the comprehensive list of reviewed studies. It shows the details of each study by process step, namely the target variable and inputs, preprocessing and feature engineering, forecasting algorithms and hyper-parameter optimization methods, and evaluation metrics and cross-validation schemes—all organized by the field of application.

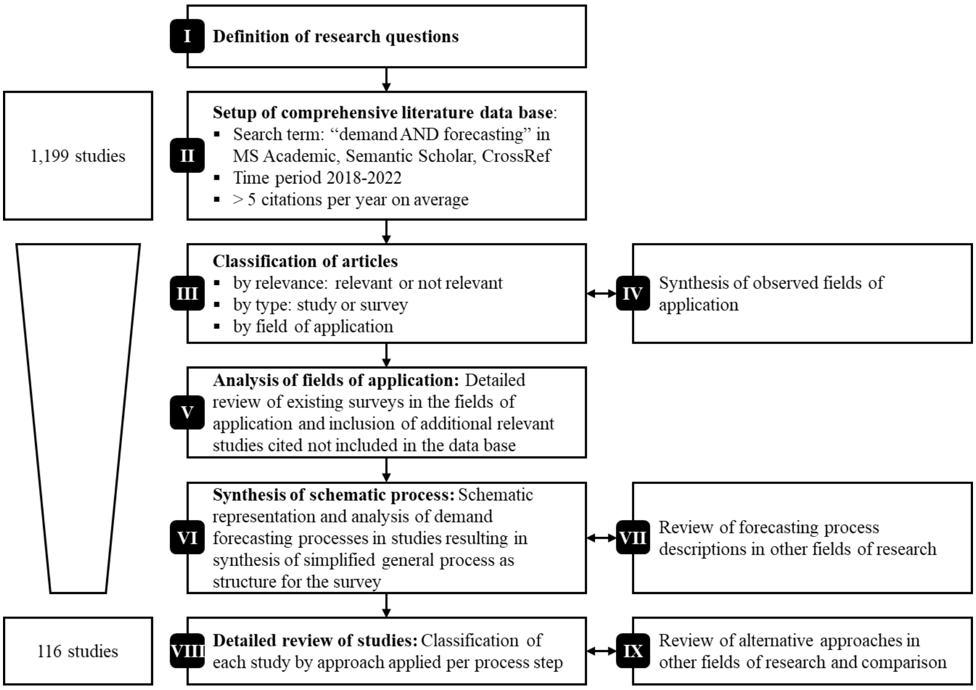

I
Definition of research questions
1,199 studies
II
Setup of comprehensive literature data base:
▪ Search term: "demand AND forecasting" in MS Academic, Semantic Scholar, CrossRef
▪ Time period 2018-2022
▪ > 5 citations per year on average
III
Classification of articles
▪ by relevance: relevant or not relevant
▪ by type: study or survey
▪ by field of application
IV
Synthesis of observed fields of application
V
Analysis of fields of application: Detailed review of existing surveys in the fields of application and inclusion of additional relevant studies cited not included in the data base
VI
Synthesis of schematic process: Schematic representation and analysis of demand forecasting processes in studies resulting in synthesis of simplified general process as structure for the survey
VII
Review of forecasting process descriptions in other fields of research
116 studies
VIII
Detailed review of studies: Classification of each study by approach applied per process step
IX
Review of alternative approaches in other fields of research and comparison

Figure 7: Literature research methodology scheme.

3.3 First Steps Toward Understanding the Demand Forecasting Literature and Research

Section 3.3 introduces the basic concepts of demand forecasting and elementary mathematical notations after which the survey develops the structure as a framework for the remainder of the article in Sections 3.3.2 and 3.4.

3.3.1 Defining the Scope

In this review, we concentrate on quantitative demand forecasting whereby demand forecasting is defined as a process of predicting future and yet unknown demands. We consider all kinds of demands: customer demands in business to customer, supply demands in business to business, demands for services (e.g., telecommunication), or demands for capacities (e.g., production or health care). In all of these cases, there can be a target variable $y_t = f(X_t, Y_{t-1}, \mathrm{E}_t)$ defined for each time step $t \in \{0,1,\dots,T\}$ that is determined by past and actual values of external factors $X_t = \{x_0, x_1, \dots, x_t\}$ (x_t may be a scalar or vector), a certain stochastic influence $\mathrm{E}_t = \{\epsilon_0, \epsilon_1, \dots, \epsilon_t\}$ and potentially also historical values of the target variable $Y_{t-1} = \{y_0, y_1, \dots, y_{t-1}\}$. The values of x_t and ϵ_t can also potentially depend on their past values and additional stochastic processes. In the demand forecasting process, the goal is to determine the relevant external influences X, to find a model for the stochastic process E, and approximate the function $f(X, Y, \mathrm{E})$ from the known history, so that future values of the target variable $Y_F = \{y_t, y_{t+1}, y_{t+2}, \dots\}$ can be estimated $\hat{Y}_F = \{\hat{y}_t, \hat{y}_{t+1}, \hat{y}_{t+2}, \dots\}$. Most forecasting approaches reduce the complexity by omitting parts of the time series history or the entire history. Depending on the use case, this can also be in line with the nature of $f(.)$.

Generally, we observe two kinds of forecasts in the reviewed literature:

- Point forecasts predict one discrete value in each forecast (for each target variable, time step, and object).
- Probability forecasts predict a confidence interval in which the future value of the target variable is assumed at a certain probability.

The advantage of probability forecasts over point forecasts is that the width of the confidence interval can indicate the certainty of the assumed forecast. In practice, decisions taken based on forecasts can strongly depend on the predicted certainty and while most reviewed studies address point forecasts, we emphasize the application of probability forecast approaches throughout the survey.

Given the structure of the problem of demand forecasting, the two major approaches to demand forecasting that we observe in practice become evident(Ivanov et al. 2021):

- Time series forecasting assumes that the most relevant information about future demands Y_F is already contained in past demands Y_{t-1}. Hereby, the target variable history simultaneously serves as input for the demand forecast. This is also intuitive, as demands can either have a constant trend over time in the long run or exhibit a seasonality influence. For example, in German grocery retailing, the first and the last day of a work week have higher sales than the other days because the customers anticipate that the shops will be closed on the weekend.

- Causal approaches emphasize the correlation of the target variable y_t with the external factors X_t. As is common in demand forecasting, we interchangeably call these external factors "*features*" (practitioners' term), "*additional inputs*" (general term), or "*exogenous variables*" (econometric term). Econometric models make use of exogenous variables such as overall economic growth, employment rates, or exchange rates. However, other external factors such as the weather, and social or other aspects can also influence demand behavior.

In practice, we usually observe mixed models which incorporate both the target variable history and exogenous variables.

The forecasting model plays a significant role in demand forecasting and two types of models can be differentiated by the underlying model assumptions:

- Classical approaches assume an underlying a-priori model. Dependencies between variables or time periods that are not considered by the model therefore cannot be mapped by the model. Additionally, the model design determines all mathematical functions considered by the model in advance. For example, a linear model can only represent linear and not non-linear functions.
- Machine learning (ML) approaches in contrast do not assume an a-priori model. An early definition of ML by Mitchell is: "A computer program [that] is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997). The general advantages of artificial neural networks (ANN) as a special type of ML reported by Zhang et al. can be summarized by: (1) The ability to deduct insights through a data-driven approach rather than theoretical assumptions in an a priori model; (2) The ability to generalize these insights gathered from the data; (3) The ability to approximate data and continuous or non-continuous relations, as well as linear and non-linear relations (Zhang et al. 1998). Cybenko showed that ANN can represent arbitrary mathematical functions (Cybenko 1989) whereas Leshno et al. indicated that this not only holds true for sigmoid activation functions but for all non-polynomial activation functions with a single hidden layer and a finite number of nodes (Leshno et al. 1993).

3.3.2 A First Overview of the Literature on Demand Forecasting

To our knowledge, no survey comparable to the current work has been published that systematically reviews and compares demand forecasting overarching the fields of applications in each step of the process from the definition of goals to the systematized evaluation and validation of results. Instead, all surveys we reviewed concentrate on a particular aspect of demand forecasting research, which is typically defined by the application. In this section, we introduce five areas of applications that are most common in the demand forecasting literature and subsequently present recent and influential surveys for each area of application.

Demand forecasting for *electricity demand* (EC) accounts for most studies. The number of studies increased with the boom of renewable energy sources (RES) after the year 2000, because of a paradigm shift in electricity grids. Decentralized energy production through rooftop photovoltaics on people's houses poses new challenges to energy providers as imbalanced electricity grids lead to failures of the grid. Therefore, supply and demand must be carefully predicted and balanced. The topic of *travel & transportation* (TR) falls under the research area with the second-largest number of articles in our literature review. This field exhibits a strong seasonality, wherefore time series models are frequently applied in this domain. The third category of use cases is *logistics & inventory* (LI), where especially intermittent and lumpy demands are forecasted. Another frequently addressed industry is that of *water & energy* (WE) demand. Due to their relatively low frequency of appearance in the studies that were reviewed, several industries, such as telecommunication, cloud computing, health, food, education, estates, construction, fashion, and a few others are subsumed under the term "*others*". Table 3 summarizes the main fields of application.

	Industry	Recent and influential literature reviews
EC	Electricity	(Aslam et al. 2021; Ahmad et al. 2020; Raza and Khosravi 2015)
TR	Travel & transportation	(Ghalehkhondabi et al. 2019; Sison et al. 2021; Song et al. 2019)
WE	Water & energy (excl. EC)	(Ghalehkhondabi et al. 2017a; Ghalehkhondabi et al. 2017b; Suganthi and Samuel 2012)
LI	Logistics & inventory	(Mediavilla et al. 2022; Pinçe et al. 2021; Seyedan and Mafakheri 2020)
OT	Other (health, telecom, food, finance, farming, construction & estate, education)	(Zhu et al. 2019; Masdari and Khoshnevis 2020; Meade and Islam 2015)

Table 3: Main fields of application with important reviews.

The forecasting competitions span the different applications of demand forecasting. Here, forecasting approaches competed for the highest forecasting performance on predefined data sets. Both Fildes and Ord and Hyndman summarized the most remarkable competitions (Hyndman 2020; Fildes and Ord 2007). Among these are the Makridakis "M Competitions" (1.: 1982 (Makridakis et al. 1982); 2.: 1993 (Makridakis et al. 1993); 3.: 2000 (Makridakis and Hibon 2000); 4.: 2020 (Makridakis et al. 2020); 5.: 2021 (Makridakis et al. 2022)) and the NN3 spin-off (Crone et al. 2011), the Global Energy Forecasting Competitions (1.: 2012 (Hong et al. 2014); 2.: 2014 (Hong et al. 2016); 3.: 2017 (Hong et al. 2019)) and several smaller competitions on the online platform Kaggle (e.g., tourism forecasting (Athanasopoulos et al. 2011)). In this context, it is important to note that the competition format poses an exception within demand forecasting, as it ensures a high level of comparability between the results since the competitions provide the same basic conditions (i.e., same input data) to all contestants and also standardize the evaluation criteria. In this article, we will show how these preconditions result in highly comparable study results.

Table 4 presents a summary of the information provided by the most recent and influential surveys for each field of application according to the field of application and subdisciplines, periods reviewed, the structure of the reviews, methods reviewed, and open issues raised. We chose the three most highly cited surveys in the period from 2018 to 2022. In some cases, we decided to consider additional surveys that did not apply to this scope if we found that they provide an especially helpful overview of the topics or a significant subdiscipline.

Survey	Survey focus	Main statements
Electricity		
Aslam et al., 2021 (Aslam et al. 2021)	**Subdisciplines:** Solar and wind energy production, electricity consumption **Methods:** Deep learning **Time period under review:** 2015-2020 **Content:** Comparison of approaches by data sets, forecasting algorithm, results, and metrics	**Reviewed:** Organized by subdiscipline, 11 general surveys, 18 wind energy studies, 13 solar irradiance studies, 12 electricity consumption studies compared by MAE, MAPE and RMSE **Open issues:** Availability and processing cost of big data required for deep learning model training, forecasting of new energy sources
Ahmad et al., 2020 (Ahmad et al. 2020)	**Subdisciplines:** Solar and wind energy production, electricity consumption, short, medium, and long term **Methods:** Machine learning, neural networks, and ensemble approaches **Time period under review:** 2009-2019 **Content:** Comparison of approaches by forecasting algorithm, results, and metrics	**Reviewed:** Organized by subdiscipline, 50 short term solar irradiance studies, 51 short term wind energy studies, 5 geothermal studies, 20 short term, 15 medium term and 10 long term electricity consumption studies **Open issues:** Consideration of regional influences and particularities in short term forecasting, medium and long term forecasts need to incorporate future climate changes, consideration of temporal and spatial differences
Raza and Khosravi, 2015 (Raza and Khosravi 2015)	**Subdisciplines:** Electricity consumption, short, medium, and long term **Methods:** Statistical and neural network approaches **Time period under review:** up to 2014 **Content:** Comparison of approaches by forecasting algorithm, results, and metrics	**Reviewed:** Organized by subdiscipline, 7 smart grid studies, 32 highlight studies of neural networks and comparison approaches **Open issues:** Consideration of meteorological factors, hybridization of approaches, incorporation of price influences, smart grid and smart buildings incorporation into model, active demand management instead of mere forecasting
Travel & transportation		
Sison et al., 2021 (Sison et al. 2021)	**Subdisciplines:** Travel demand (airline and hotel) **Methods:** Machine learning **Time period under review:** 2000-2020 **Content:** Overview of terms and qualitative comparison of methods	**Reviewed:** Organized by process step and approach: Target variables, 5 studies preprocessing, 13 studies ARIMA, 10 studies linear regression, 3 econometric models, 9 backpropagation ANN, 9 SVM, 8 hybrid models, model evaluation (no references) **Open issues:** Intensified application of ML algorithms for non-linear data relationships, incorporation of cost-efficiency, computing time as well as consistency over different forecasting horizons in addition to forecasting accuracy, development of hybrid approaches incorporating decomposition techniques, adoption of techniques from other fields of research as cold start forecasting or microlevel forecasting, securing data privacy, adoption of federated learning.
Ghalehkhondabi et al., 2019 (Ghalehkhondabi et al. 2019)	**Subdisciplines:** Tourism and transportation demand **Methods:** Statistical and machine learning **Time period under review:** 2005-2018 **Content:** Qualitative comparison of studies by method applied	**Reviewed:** Organized by approach, 18 statistical time series studies (ARIMA; regression), 12 ML studies **Open issues:** Development of hybrid methods, determination of qualitative variables impact on forecast accuracy, incorporation of multi-level seasonality and extreme events, data quality assurance methods
Song et al., 2019 (Song et al. 2019)	**Subdisciplines:** Tourism demand **Methods:** Statistical, econometric, judgmental and machine learning **Time period under review:** 1968-2018 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** Organized by approach, 211 studies, best performing algorithms: 72 statistical time series, 82 econometric model, 6 judgmental, 33 artificial intelligence, 18 hybrid approach **Open issues:** Development of ML approaches to exploit big data, hybridization of time series and ML approaches, especially forecasting combination weighting schemes, consideration of complex inputs and dependencies as cross-border employment, cultural differences, globalization, etc., on micro-level,

Song and Li, 2008 (Song and Li 2008)	**Subdisciplines:** Tourism demand **Methods:** Statistical, econometric and machine learning **Time period under review:** 2000-2007 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** Organized by approach, 121 studies, best performing algorithms: 52 statistical time series, 40 econometric model, 17 artificial intelligence, 1 hybrid approach, 11 other (mainly panel analysis) **Open issues:** More advanced incorporation of seasonality, research on benefits of disaggregation approaches, analysis of the impact of crisis on demand.
Water & energy		
Ghalehkhondabi et al., 2017a (Ghalehkhondabi et al. 2017a)	**Subdisciplines:** Energy demand: Electricity, gas, heat and building energy **Methods:** Statistical and machine learning **Time period under review:** 2005-2015 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** Organized by approach, tabular comparison of studies: 7 ANNs, 5 fuzzy logic, 5 time series approaches, 5 grey prediction, 5 moving average, 5 regression, 5 support vector machines, 3 genetic algorithms, 9 econometric and system dynamics **Open issues:** Intensification of peak energy demand forecast research, consideration of saturation factors that come into play when energy demand stops rising in the future, incorporation of weather/climate influences other than temperature, improvement of hybrid forecasting approaches, improvement of error metrics, difficulties of neural network approaches: overfitting, overall implementation complexity, amount of data required for training.
Ghalehkhondabi et al., 2017b (Ghalehkhondabi et al. 2017b)	**Subdisciplines:** Water demand **Methods:** Soft computing, i.e. machine learing, not statistical methods **Time period under review:** 2005-2015 **Content:** Qualitative comparison of studies by method applied	**Reviewed:** Organized by approach: 25 ANN studies, 9 fuzzy logic studies, 9 support vector machine studies, 10 metaheuristic or hybrid approach studies, 4 system dynamics studies. **Open issues:** Research on more sophisticated neural network approaches (e.g., recurrent networks), adoption of unsupervised learning methods, research of further state-of-the-art metaheuristics (ant colony simulation, simulated annealing, etc.) and ensemble approaches, study of input factor effectiveness, development of strategies for non-stationary time series neural network approaches.
Suganthi and Samuel, 2012 (Suganthi and Samuel 2012)	**Subdisciplines:** Energy demand **Methods:** Statistical, econometric and machine learning **Time period under review:** up to 2011 **Content:** Qualitative comparison of studies by method applied	**Reviewed:** Organized by approach: 23 time series studies, 20 regression studies, 51 econometric studies, 10 decomposition model studies, 46 unit root test and cointegration model studies, 9 ARIMA studies, 81 ANN studies, 12 grey prediction studies, 7 input-output studies, 22 fuzzy or genetic algorithm studies, 4 support vector machine studies, 8 optimization heuristics studies, 41 buttom-up model studies. **Open issues:** Application of the available methods in practical applications and derivation of policies for energy demand planners and politicians, adoption of the state-of-the-art methods by researchers.
Logistics & inventory		
Mediavilla et al., 2022 (Mediavilla et al. 2022)	**Subdisciplines:** Supply chain demand **Methods:** Statistical and machine learning **Time period under review:** 2017-2021 **Content:** Qualitative comparison of studies by method applied	**Reviewed:** 33 studies, classified by: industry and position in supply chain, empirical data set, algorithms applied, metric used and implementation tool. **Open issues:** Combining demand forecasts along the supply chain to prevent bull-whip effect while most recent literature concentrates on the retailer (last step in supply chain), trend towards complex ML methods observed in recent years, intensification of integration of additional external inputs and integration of feature engineering methods in pipelines.
(Pinçe et al. 2021)	**Subdisciplines:** Spare parts demand **Methods:** Statistical, contextual and machine learing **Time period under review:** up to 2020 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** 56 studies, organized by forecasting approach, additional overview of performance measure metrics.

		Open issues: Unification of performance measures as well as more detailed descriptions of the used data sets for better comparability of approaches – ideally resulting in a standardized study reporting scheme, intensification of application of inventory performance measures over forecast accuracy measures to improve practical relevance, intensification of development of big data capable approaches to exploit for example installed base information, development of judgmental approaches to better incorporate human knowledge, development of supply chain overarching forecasting methods and collaboration approaches, improvement of demand classification (e.g. anomalies through maintenance works) and preprocessing of such anomalies
(Aamer et al. 2020)	**Subdisciplines:** Supply chain demand **Methods:** Machine learning **Time period under review:** 2010-2019 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** 79 studies, organized by approach: neural networks and variants 61%, SVM 27%, Decision trees and boosted/bagged variants 7%, other 5% **Open issues:** Intensification of research in fields which have yet not received sufficient attention with respect to supply aspects: agriculture, healthcare and transportation of goods, integration of demand forecasting in collaborative supply chain planning processes.
(Seyedan and Mafakheri 2020)	**Subdisciplines:** Supply chain demand **Methods:** Machine learning **Time period under review:** 2005-2019 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** 49 studies, organized by approach and application: ANN, Clustering, k-nearest neighbors, regression, SVM, statistical time series forecasting **Open issues:** Combination of supply chain demand forecasting with integrated cost optimization approaches, inclusion of reverse and closed loop logistics in research.
Other		
(Zhu et al. 2019)	**Subdisciplines:** Healthcare and emergency demand **Methods:** Statistical and machine learning **Time period under review:** 1980-2018 **Content:** Qualitative comparison of studies by method applied	**Reviewed:** 1,230 studies in keyword search, of which 32 described in detail, organized by approach **Open issues:** Intensification of research and standardization of accepted forecasting models, adoption of methods to exploit big data considering the real-life limitation of data availability and data privacy restrictions, diversification of target variables and performance metrics
(Masdari and Khoshnevis 2020)	**Subdisciplines:** Cloud computing demand **Methods:** Statistical and machine learning **Time period under review:** Up to 2020 **Content:** Qualitative and quantitative comparison of studies by method applied	**Reviewed:** 54 studies, organized by approach and classified by data set, implementation, evaluation metrics and predicted factors **Open issues:** Deeper research of application of ML approaches, improvement of schemes to adopt to real life situations, improved discipline metrics, coverage of longer horizons with seasonal variations, improvement of resource allocation algorithms, integration of intrusion detection, development of lightweight models (e.g., for IoT or mobile devices), autoscaling schemes for DDoS detection and prevention.
(Meade and Islam 2015)	**Subdisciplines:** Telecommunication demand **Methods:** Statistical and machine learning **Time period under review:** up to 2014 **Content:** Qualitative and quantitative comparison of studies by topic	**Reviewed:** Organized by topics: 13 single country ICT diffusion model studies, 12 multi country ICT diffusion model studies, 21 call center time series model studies, 7 univariate telecommunications time series studies, 9 internet usage and provision model studies, 14 internet technologies studies. **Open issues:** Low level of innovation in single country diffusion modelling compared to multi country studies, intensification of comparing studies in technology choice models, availability of adequate data for research.

Table 4: Overview of the reviewed surveys in different applications of demand forecasting.

3.4 The Process of Demand Forecasting

In the first step, we review the process of demand forecasting to provide a framework for the approaches that we discuss in detail in the subsequent sections. The details of the literature research process are provided in Section 3.3.2 from which it becomes apparent that only a few studies explicitly formulate the process of demand forecasting, and most papers focus on specific aspects of the process to contribute to their field of research. This section gives an overview of the process inferred from three types of studies: A. studies that mention the process as part of their approach to a certain use case or methodology, B. surveys of forecasting techniques, and C. studies on AutoML.

Type A studies implicitly use models of the demand forecasting process. For example, Raza and Khosravi (Raza and Khosravi 2015), Herrera et al. (Herrera et al. 2010), and Hua (Hua 1996) describe slightly different versions of a five-step approach: 1. selection of inputs, 2. different techniques of preprocessing, 3. model setup and training, 4. post-processing of the results, and 5. evaluation and interpretation of the results. Other recent studies apply comparable approaches, for example, Abbasimehr et al., Xie et al., and several others (Xie et al. 2021; Cheng et al. 2017; Ryu et al. 2017; Cai et al. 2021; Abbasimehr et al. 2020; Iftikhar and Khan 2020; Bandara et al. 2019; Fu et al. 2018; Tsao et al. 2022; Venkatesh et al. 2014; Bega et al. 2019; Kulshrestha et al. 2020; Li et al. 2020a; Raza et al. 2022; Sison et al. 2021).

Studies by de Gooijer and Hyndman (de Gooijer and Hyndman 2006) and Montgomery et al. (Montgomery et al. 2008) describe the process of demand forecasting from a general perspective as part of their surveys on the history of demand forecasting (Type B). AutoML studies (Type C) provide examples of the demand forecasting process. AutoML is a discipline in ML that studies concepts to automate the entire ML process, especially concerning the tuning of model parameters. This discipline is predestined as the automation of the process requires researchers to thoroughly describe and understand the process. We find numerous appropriate examples that describe quantitative forecasting processes: Feurer et al. (Feurer et al. 2015), Hall et al. (Hall et al. 2009), He et al. (He et al. 2019), Molino et al. (Molino et al. 2019), Truong et al. (Truong et al. 2019), Yan (Yan 2012), Yao et al. (Yao et al. 2018b) whereby all these examples are in line with the findings from the field of demand forecasting.

For the remainder of the review on the state-of-the-art, we adopt the forecasting process approaches from the literature and conclude a unified process (see Figure 8). It considers forecasting goal definition and input selection in the first step, various preprocessing techniques in the second step, feature engineering as a separate third step (as proposed in the AutoML context), and model design, hyper-parameter optimization, and model evaluation in the fourth step.

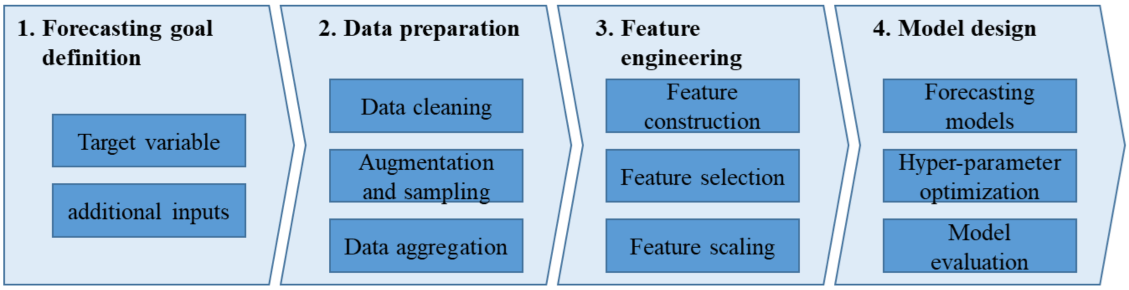


Figure 8: A schematic process of demand forecsting

- Section 3.5 provides an overview of characteristic target variables in the different industries that apply demand forecasting, as well as additional external inputs.
- Section 3.6 provides an overview of data preprocessing techniques applied in demand forecasting and a broader view of general techniques from other domains.
- Section 3.7 summarizes the feature engineering techniques used to select and generate optimal feature sets from demand forecasting and other domains.
- Section 3.8 surveys the vast number of available and applied forecasting models and techniques to evaluate model predictions.

As some of the techniques described in this review are also applied in other ML disciplines and not only in demand forecasting, a brief introduction to these techniques is provided for those who are new to the field of demand forecasting. More importantly, the survey focuses on the particularities of the applications of the techniques specific to demand forecasting.

The following sections provide brief summaries and conclusions of the respective sections and we pick up and unify the section-wise conclusions in Section 3.9 (Summary of the Findings and Open Issues) to draw a comprehensive conclusion.

3.5 The Starting Point of Demand Forecasting: Forecasting Goal Definition

The standard approach in the literature is to start the demand forecasting process with the definition of the target variable, and appropriate inputs for the model and the target variable are worth examining more closely in all studies. Thus, in this section, we investigate target variable definition and review how different industries define their target variables and what additional inputs are typically used. Table 5 provides a comprehensive overview of these points, while Table 17 provides the target variables and additional inputs of each reviewed study.

The target variable is the first – and one of the most determining factors – that studies define in the first step. Song and Li published a comprehensive study showing the multitude of target variables in travel demand forecasting (Song and Li 2008). However, even though it has a strong impact on the entire demand forecasting process, this step is often underestimated. For example, we consider a researcher who estimates the water demand of a municipality. The target variable has numerous manifestations including, among others, a time component (e.g., daily, weekly, monthly, yearly demand), a spatial component (e.g., per household, per area, per city), a metric component (e.g., liters per capita, liters absolute), a social component (e.g., per income group, per age), and a point, distribution, or profile forecast (e.g., mean, median, peak, actual distribution). Even though the task to define a target variable seems straightforward at first, we can tell from this example that it influences all subsequent steps. For example, the yearly water demand is likely to correlate with economic indicators, whereas the daily demand will not necessarily do so. The daily water demand exhibits strong intra-day patterns and fluctuations – in contrast to the annual demand. Therefore, we would need to choose a model that can represent intra-day patterns and fluctuations – other than a model appropriate for annual demand.

The field of electricity demand forecasting is one of the most mature fields and hence, studies relating to it exhibit very differentiated target variables including different time horizons (from very short to long term), single value predictions or load profiles, or price forecasting. Due to their use case similarity, water and energy exhibit similar target variables as electricity. In contrast to these deeply evolved target variables, travel demand forecasting addresses the broad aspects of use cases by a variety of target variables. While transportation and hospitality providers require traveler numbers as planning input, businesses are more interested in travelers' expenditures on goods and services.

Looking at the example of logistics & inventory demand forecasting, we can show how the target variable is not only decisive from a business standpoint but also has important implications for the mathematical characteristics of the demand forecasting process. For example, spare part demands for inventory planning exhibit particularly intermittent demands where demand events are followed by periods with zero demands. Such demand behavior is opposed to the usual assumptions in time series forecasting, where the basic assumption is that while demand may be fluctuating, it however is mostly greater than zero. As we will see in the following sections, this affects all further steps and results in unusual models that are applied in spare parts demand forecasting – compared to other disciplines.

We first examine the additional inputs through the product (i.e., goods) demand forecasting in which products exhibit pronounced lifecycles. Companies develop new products, set up production and logistics capacities, introduce the product to the market, and then typically observe increasing sales, a plateau of sales, and subsequently a constant decrease. Before the market launch, no historical sales data is available to determine the mechanisms of the demand behavior for a given product. This is opposed to other industries that are characterized by long or infinite product lifespans. The fashion industry illustrates the challenges of the issue of product life cycles as the demand for a new collection is difficult to predict before its launch because no historical data is available. Since companies decide on all relevant details (such as the number of produced items) before the launch, the industry mainly applies two techniques: preview sales (Mostard et al. 2011) and demand forecasting based on product features (Ferreira et al. 2016). Preview sales are used to evaluate the customers' reactions based on a small and representative sample of customers. Product feature-based forecasting relies on reducing the product demands to the products' features and the re-composition of the expected demand for new products based on their feature combination.

Further examples show the relevance of additional model inputs to enhance demand forecasting. Spare parts demand forecasting is improved through the knowledge of the use of the products that require spare parts (known as Active Installed Base, AIB) (Kim et al. 2017). Point of sale (POS) data can improve supply chain forecasting by two means. First, if the POS information is available to upstream players in the supply chain, the information can travel the supply chain faster. Second, direct correlations between customer characteristics and product demand behavior can be recorded and deducted (Williams and Waller 2010).

Apart from these problem-instance specific model inputs, we find that the obvious indicators are applied as often as expected and, among others, the date information such as day-of-week and day-of-month enable models to identify seasonality while public holidays and public events help to explain deviations from regular patterns. Even though weather and climate information is automatically incorporated into the models in many articles, the correlation between weather and demand is particularly strong in electricity and energy demand. Economic indicators, such as gross domestic product (GDP) or purchasing power influence purchasing behavior and therefore also correlate with the demand for electricity, energy, travel, and leisure.

Despite the large number of studies that incorporate additional variables into their models, it is debatable under which conditions they can actually improve forecasting. In their tourism forecasting competition, Athanasopoulos et al. (Athanasopoulos et al. 2011) conclude that pure time series models achieve the same and better results than models that include additional econometric inputs. However, the majority of studies include exogenous variables and exhibit an improved forecasting quality.

Industry	Target variable	Additional inputs
Electricity	Very short (Shang 2013), short (Ryu et al. 2017; Taylor 2010), medium (Chang et al. 2011) and long term (Hyndman and Fan 2010) forecast; one load forecast (most common) or load profile forecast (Amini et al. 2016) or price forecast (Chan et al. 2012)	Target variable history, date information (day-of-week, public holidays, …), climate and/or weather, economic inputs
Travel & transport	Travel: Tourist arrivals, expenditures (by product category), tourism employment (Song et al. 2010) Transportation: Passenger rides (Ke et al. 2017)	Target variable history (Du Preez and Witt 2003), economic inputs, e.g. income and exchange rate (Song et al. 2003)
Water & energy (excl. electricity)	Short (day, peak), medium, long: demand and revenue (Billings and Jones 2008)	Target variable history (Herrera et al. 2010), climate and/or weather (Izadyar et al. 2015; Adamowski et al. 2012), economic inputs (e.g. purchasing power) (Nasr et al. 2002), demographics, date information (Rockaway et al. 2011; Donkor et al. 2014)
Logistics & inventory	Materials demand (production inputs (Toktay and Wein 2001) and spare parts (Willemain et al. 2004)), logistics capacities (Carbonneau et al. 2008)	(active) installed base (Kim et al. 2017), point of sale data (Williams and Waller 2010), economic inputs (Hou and Zhang 2005)
Other	Fashion: Products/product groups, product life cycle (by regions) (Mostard et al. 2011; Nenni et al. 2013) Health: Professional labor capacities (Landry et al. 2016; Maier and Afentakis 2013), materials & drugs demand (Drackley et al. 2012) Finance: Cash demand (Venkatesh et al. 2014), inflation (Dreger and Wolters 2014)	Fashion: Preview sales (Mostard et al. 2011), product features (Ferreira et al. 2016), events (Ferreira et al. 2016) Health: Demographics (Drackley et al. 2012), economic indicators (Maier and Afentakis 2013), social media (Kim et al. 2015) Finance: Target variable history, date information (Venkatesh et al. 2014), economic indicators (Dreger and Wolters 2014)

Table 5: Industries and their typical forecasting target variables and model inputs. The citations highlight examples of the application of the named variable.

Even though the literature regularly considers historical sales data as a proxy for historical demand, it should be noted that these two factors are not equivalent and historical sales frequently do not reflect lost sales, i.e., those demands that could not be fulfilled due to insufficient stock, production, service capacities, or similar aspects. In some cases, companies keep records of indicators for lost sales, e.g., order book backlog and lost tender processes or rejected orders. In these cases, the real demand can be deduced (Gilliland and others 2010).

A detailed description of the characteristics of the data used for empirical research is important for other researchers to compare their results to existing studies. Throughout our review, we observed that only a few studies describe the statistical properties of their data set in such detail that a direct comparison is possible. Some studies, however, use publicly available data (such as competition data) to make their data source transparent.

3.6 Setting the Foundations: Preprocessing of Data

This section provides an overview of the techniques that researchers and practitioners usually apply for demand forecasting. Most studies mention the use of the following techniques although they do not focus on their optimization but rather apply them as a prerequisite for subsequent steps.

To a certain degree, the overall data obtained in business environments regularly contains faulty data which researchers and practitioners need to remove from their data source before building and training forecasting models. This process is usually called **data cleaning**. Part of this step is to impute missing values (*missing value treatment*), which is not the focus of this article. Furthermore, *anomaly or outlier detection* helps to identify faulty data. Apart from the manual inspection of the data and basic plausibility rules (e.g. (García Valverde et al.)), studies apply outlier detection approaches to detect anomalies that are possibly invalid data points. Shang and Vilar, et al. apply *functional principal component analysis* (FPCA) (Shang 2013; Vilar et al. 2016). Del Real et al. apply outlier detection in electricity demand forecasting similar to approaches proposed in other studies (Del Real et al. 2020; Pérez-Chacón et al. 2020; Shakarami et al. 2021). General concepts for outlier detection are inter alia: *least absolute shrinkage and selection operator* (LASSO) outlier detection (Au et al. 2010), *symbolic aggregate approximation* (SAX) (Lin et al. 2005), Isolation Forests (Liu et al. 2008), or Local Outlier Factors (Breunig et al. 2000). Some forecasting models require outlier removal due to their sensitivity to outliers, e.g. as shown by Tsay for ARMA models (Tsay 2000). However, it should be noted that outliers are not always data errors but can also be valid observations, for example, caused by unusual events. However, they do present indications that require further investigation.

Some models such as the ARMA and GARCH models incorporate stationarity assumptions (see Section 3.8.1). **Stationarity tests** are applied to test input data to comply with the stationarity assumptions. Demand forecasting studies regularly use unit root tests such as the *Dickey-Fuller, Kwiatkowski–Phillips–Schmidt–Shin* (KPSS), or *Hylleberg, Engle, Granger and Yoo* (HEGY) Tests (for example (Claveria et al. 2015; Taylor 2010)). Song and Li provide an overview of stationarity tests and remark that they are only applicable in the case of deterministic seasonality (Song and Li 2008). If stationarity is not given for the input data, *differentiation* is applied to provide a stationary input to the model.

In the context of data preprocessing, techniques to deal with small data sets or imbalanced data sets should also be mentioned. **Data augmentation** techniques help to artificially enlarge small data sets and Herrera et al. propose *sliding time windows* and Monte Carlo simulations to present different windows of the same time series to the forecasting models (Herrera et al. 2010). Additional approaches such as *time window warping* (DTW) (Rashid and Louis 2019) or *dynamic time warping barycenter averaging* (DBA) (Forestier et al. 2017) have been proposed in forecasting research but not notably adopted in demand forecasting. Although DTW can also be considered a feature extraction technique, it is widely described as an augmentation technique in time series forecasting. Liu et al. describe and compare general time series augmentation techniques that provide *noise injection* to existing time series (Liu et al. 2020a), and Shakarami et al. apply noise injection in a demand forecasting study (Shakarami et al. 2021). In their comprehensive overview of general time series augmentation methods, Wen et al. also describe advanced augmentation techniques such as the use of *generative adversarial networks* (GAN) (Wen et al. 2021b) although we do not find examples of this technique being applied in demand forecasting. In contrast to data augmentation techniques, **sampling** (or **bootstrapping**) approaches take samples from the input data (randomly or systematically, with or without replacement). This can be helpful when classes of time series are underrepresented in the data set. General research in this field proposes approaches such as *adaptive synthetic sampling* (ADASYN) (He et al. 2008) and *synthetic minority oversampling technique* (SMOTE) (Chawla et al. 2002), inter alia to balance data sets. Hyndman and Fan recommend bootstrapping in electricity demand forecasting, where they also exploit the opportunity to make density instead of point predictions (Hyndman and Fan 2010).

Especially when demands are intermittent (i.e., time series frequently containing zero demands), studies apply **aggregation-disaggregation** or **bottom-up/top-down** approaches. These approaches aggregate time series along a dimension (e.g., time, product features, or others), or predict the smoothed aggregated time series and then break down the aggregated forecast back to the actual target variable. We find examples in demand forecasting by Nikolopoulos et al. (spare parts, product feature aggregation) (Nikolopoulos et al. 2011), Rostami-Tabar et al. (spare parts, temporal aggregation) (Rostami-Tabar et al. 2013), Carson et al. (travel demand, spatial aggregation) (Carson et al. 2011), and Hyndman et al. (tourism demand, multilevel hierarchical) (Hyndman et al. 2011).

Observed methods applied in reviewed studies	**Additional state-of-the-art methods in other fields of forecasting**
Data cleaning • Missing data interpolation (Chen et al. 2017; Bandara et al. 2019; Wang et al. 2020; Ryu et al. 2017; Shakarami et al. 2021; Liu et al. 2020b; Kurek et al. 2021; Romano and Kapelan 2014; Del Real et al. 2020; Xenochristou and Kapelan 2020) • Outlier removal (Pérez-Chacón et al. 2020; Shakarami et al. 2021; Xenochristou and Kapelan 2020) Stationarity tests • (Augmented) Dickey-Fuller test (ADF)(Jiang et al. 2017; Claveria et al. 2015; Dreger and Wolters 2014; Huang et al. 2021; Dittmer et al. 2021; Kulshrestha et al. 2020; Yao and Cao 2020; Volchek et al. 2019) • Kwiatkowski–Phillips–Schmidt–Shin Test (KPSS)(Kulshrestha et al. 2020; Volchek et al. 2019; Claveria et al. 2015) • Phillips-Perron (PP)(Volchek et al. 2019) Trend removal (Williams and Short 2020; Felice et al. 2015; Volchek et al. 2019) Data augmentation • Noise injection (Shakarami et al. 2021) • Dynamic Time Warping (Zhu et al. 2021; Zhang et al. 2020; Yao et al. 2018a) Box Cox transformation (Kim and Kim 2021)	Data cleaning • Outlier removal (least absolute shrinkage and selection operator (LASSO) outlier detection (Au et al. 2010), symbolic aggregate approximation SAX (Lin et al. 2005), Isolation Forests (Liu et al. 2008) or Local Outlier Factor (Breunig et al. 2000)) Data augmentation: basic approaches (cropping, flipping, jittering time) (Wen et al. 2021b), dynamic time window warping (DTW) (Rashid and Louis 2019) or dynamic barycenter averaging (DBA) (Forestier et al. 2017), other advanced approaches (decomposition, statistical and ML generative as generative adversarial networks (GAN)) (Wen et al. 2021b) Sampling: adaptive synthetic sampling (ADASYN)(He et al. 2008), synthetic minority oversampling technique (SMOTE)(Chawla et al. 2002), etc.

Table 6: Comparison of the reviewed methods in the data preprocessing and state-of-the-art methods in other fields of forecasting.

Table 6 summarizes and compares the methods observed in demand forecasting studies and further state-of-the-art preprocessing methods. In contrast to the subsequent tables, we do not differentiate between the fields of applications in this table, as documented preprocessing methods are rare throughout all fields of applications.

Summarizing this section, we can state that demand forecasting applies several data preprocessing methods and studies are especially rigorous in applying stationarity tests and detrending when the forecasting algorithm used requires this as a precondition. Lacking data and outlier treatment are rarely described and, if they are, they are hardly ever covered in detail. In addition, we suggest that the rich range of data augmentation techniques could be adopted more in demand forecasting studies to enlarge the available data sets. Section 3.9.2 shows that the availability of data is often an issue in demand forecasting literature.

3.7 Enriching the Model Input: Feature Engineering

In Section 3.3.1, we introduced the external variable $X_t = \{x_0, x_1, \dots, x_t\}$ as a feature (if x is a scalar) or a set of features, if x is a vector. We can also consider the historical values of the target variable $Y_{t-1} = \{y_0, y_1, \dots, y_{t-1}\}$ as a feature. In many cases in forecasting, we obtain these variables from practical applications in which they are determined by the data that can be recorded and not necessarily what is required for the model design. Therefore, researchers have studied ways to create the most suitable model inputs from the raw data obtained.

In the following sections, we review the methods to build novel features from existing features (feature construction and extraction), select an optimal subset of features (feature selection), and prepare features for the forecasting models that we introduce in Section 3.8.

Section 3.7.4 summarizes and collectively discusses Sections 3.7.1 to 3.7.3.

3.7.1 Feature construction and extraction

This section addresses two conceptual approaches: feature construction and feature extraction. As the literature is not always consistent in using these terms, we chose to follow the definition by Motoda and Liu (Motoda and Liu 2002).

- Feature construction (FC): FC constructs new features by applying mathematical functions to one or more of the available features, thereby generating new and additional features from existing ones (*feature generation*).
- Feature extraction (FE): In contrast to FC, FE aims at reducing the number of features by transforming the features into a new feature space – yet at the same time without a significant loss of information (*dimensionality reduction*).

There are evident examples of why constructing new features is beneficial. For example, the body mass index (BMI) is a feature constructed from the features "body length" and "body weight". However, the BMI might be much more strongly correlated to the prevalence of diseases than body length or body weight by themselves. From this example, we see that there is a large number of re-combination possibilities even when we only consider basic mathematical operations whereby the following state-of-the-art feature construction algorithms from ML (see Sondhi (Sondhi 2009)) and time series forecasting are considered:

- Non-time series-specific FC: Interaction terms (e.g., summation, multiplication, polynomials (Sutton and Matheus 1991) of features), Cartesian product (Pazzani 1998), M-of-N (Pazzani 1998), logical conjunctions (Boolean features), decision tree approaches (FRINGE (Pagallo 1989), FICUS (Markovitch and Rosenstein 2002)), and genetic programming (Krawiec 2002).
- Time series-specific FC: Functions based on the date and time of the observation (e.g., day-of-week, month, public holidays), time-lagged features (e.g., observation of last season, last year, rolling window), and time series statistics (e.g., mean, variation, trend) (Christ et al. 2016; Fulcher et al. 2013)

Reviewing the approaches, the countless possibilities for how to generate features become evident, and, therefore, the literature proposes frameworks such as FINGE, FICUS, and others, which utilize heuristics to limit the number of newly generated features. Section 3.7.2 shows how relevant features can be selected for model input.

Both Fulcher et al. and Christ et al. propose frameworks that calculate thousands of characteristics (especially statistics) of the time series and use these as new features (Fulcher et al. 2013; Christ et al. 2016). This detailed characterization is similar to some approaches that we find among the feature extraction techniques. Here, the literature proposes transformations of time series into other representations, for example using Fourier transformation, wavelet transformation, or representation of time series through statistical time series models. Optimally, feature extraction approaches such as Fourier transformation describe a time series without a loss of information. At the same time, the information in the transformed feature space is considerably condensed.

- Non-time series-specific FE: Principal components analysis (PCA) (Pearson 1901), independent component analysis (ICA) (Jutten and Herault 1991), linear discriminant analysis (LDA) (Cohen 2013), locally linear embedding (LLE) (Roweis and Saul 2000), t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton 2008), autoencoders (AE) (Masci et al. 2011), pre-trained ANN (Setiono and Liu 1998) and more, see Lee and Verleysen (Lee and Verleysen 2007).
- Time series-specific FE: Empirical mode decomposition (EMD) (Huang et al. 1998), Fourier and wavelet transformation (Wu et al. 2000), time series forests (Deng et al. 2013), generic time series calculation from various statistics, for example, tsfresh, TSFEL, FeatuRe, and others (Christ et al. 2018; Barandas et al. 2020; Hyndman et al. 2015; Tavenard et al. 2020).

Overall, demand forecasting studies do not frequently apply feature construction and feature extraction methods in a systematic way and most of the reviewed studies did not emphasize FC and FE techniques. However, all studies implicitly optimized their model input for good forecasting results.

We observe that *electricity* demand forecasting most extensively employs FC/FE. Here, transformation approaches (Fourier and wavelet transformation) are most common, especially in the field of *water & energy* demand forecasting. *Travel & transportation* demand forecasting mostly applies techniques that can be interpreted from an econometrical perspective or that relate to trend and seasonal decomposition. PCA and EMD are techniques that we occasionally observe in all disciplines and Table 7 provides an overview of the methods applied in the fields of demand forecasting.

Wavelet transformation techniques in conjunction with ANN are intensely studied. In fact, they are used as a stand-alone forecasting approach, which is why we review them more closely in Section 3.8.1.

Feature construction and extraction methods in the demand forecasting literature	
Electricity	Principal Component Analysis (PCA) (Wang et al. 2020), Fourier Transformation (Li et al. 2020b), Wavelet Transformation (Ofori-Ntow Jnr et al. 2021), Empirical Mode Decomposition (EMD) (Hwangbo et al. 2019; AL-Musaylh et al. 2018b; Qiu et al. 2017), Piecewise/Symbolic Aggregate Approximation (PAA, SAX)(Williams and Short 2020), Convolutional ANN (Williams and Short 2020), Mixed-Data Sampling (MIDAS) (Choi et al. 2020), Taylor Approximation (Khan and Jayaweera 2020)
Travel & transportation	Empirical Mode Decomposition (EMD) (Li and Law 2020; Jiang et al. 2014), STL approach (Seasonal and Trend Decomposition using Loess) (Zhang et al. 2021), Generalized Dynamic Factor Model (GDFM) (Wen et al. 2021a), Seasonal trend decomposition (Zhang et al. 2020), Hodrick-Prescott (HP) filter (Yao and Cao 2020)
Water & energy (excl. electricity)	Wavelet Transformation (Rezaali et al. 2021; Panapakidis and Dagoumas 2017; Adamowski et al. 2012), Singular Spectrum Analysis (SSA)(Zubaidi et al. 2020), Variational Mode Decomposition (VMD) (Sun and Zhao 2020), statistics applied to variables (Kurek et al. 2021)
Inventory & logistics	Statistics of target variable (Zhu et al. 2021; Li and Lim 2018; Huber and Stuckenschmidt 2020), Principal Component Analysis (PCA) (Huang et al. 2021), Dynamic Time Warping (DTW) (Zhu et al. 2021), Latent Dirichlet Allocation (Iftikhar and Khan 2020), Dynamic Wavelet Transformation (DWT) (Jaipuria and Mahapatra 2014), Neural network trained to extract features (Cai et al. 2021)
Other demand forecasting	Multiplicative time series representation (Venkatesh et al. 2014), Fractional Factorial Design(Maeng et al. 2020), Autoencoder (Bega et al. 2019)
Other general methods	Time series forests, generic time series extraction approaches

Table 7: Comparison of the applied feature construction and extraction methods by field of application and gaps to general methods applied in other forecasting disciplines.

3.7.2 Feature selection

Section 3.5 observes the variety of model input variables. In addition, the feature construction techniques in Section 3.7.1 make it clear that the number of features can grow significantly. However, when we expose a model to too many input variables, it is likely that it will largely fit the noise of the variables without learning the relevant dependencies in the data (Li et al. 2017). Therefore, studies aim to determine which features best determine the model – assuming that an optimal subset of features exists amongst all features that were defined in the previous process steps. The level of complexity of the search is high and Guyon and Elisseeff and Piramuthu show that an optimal subset of features is not necessarily the combination of the best individual features as the features frequently complement each other and simple filter approaches typically fail to reveal the complementary dependencies of features (Guyon and Elisseeff 2003; Piramuthu 2004).

In this context, the literature uses the terms feature selection (FS) or *feature evaluation* and we review three principal categories of FS algorithms:

- Filter: These approaches determine whether a feature is a model input candidate based on the statistical characteristics of the feature itself or statistical characteristics of the feature and the target variables—independently of the forecasting model. For example, variable ranking through correlation (Hall 1999) or mutual information criteria (Battiti 1994; Koller and Sahami 1996); stepwise multiple regression (Piramuthu 2004); (partial) auto-correlation function, stepwise regression, and spectral analysis for an autoregressive time lag selection (Crone and Kourentzes 2010).
- Wrapper: A search approach is built around a forecasting model that systematically explores the forecasting performance of different feature subsets. The forecasting model functions as a black- or gray box. This includes, among others, general search space optimization algorithms (e.g., branch and bound search (Narendra and Fukunaga 1977), genetic algorithms (Yang and Honavar 1998) or particle swarm optimization (Sheikhan and Mohammadi 2013)); FOCUS, ID3, MIN-FEATURES (Almuallim and Dietterich 1991); sequential forward selection (SFS), or sequential backward elimination (SBE) (Devijver and Kittler 1982; Koller and Sahami 1996).
- Embedding: Some forecasting models already incorporate feature selection approaches as part of the forecasting algorithm, including regularization (e.g., LASSO (Tibshirani 1996), RIDGE (Marquardt and Snee 1975)), deep feature selection, and decision tree feature importance (Breiman 2001).

When applying feature selection techniques, one should be aware that different approaches are used depending on the type of input and target variable (categorical, ordinal, numerical, …). Studies in demand forecasting typically deal with time series and regression tasks and, therefore, this section does not include approaches such as the ANOVA and χ^2-tests which are categorical tests.

Several well-written reviews of general ML FS approaches provide deeper insights into the topic of feature selection (Li et al. 2017; Piramuthu 2004; Guyon and Elisseeff 2003; Blum and Langley 1997; Chandrashekar and Sahin 2014). These studies discuss the advantages and disadvantages of the FS algorithms, for example, Saeys et al. filter all approaches that are least computationally expensive. However, they often make strong assumptions about the data or the forecasting algorithm and are less general. Using wrapper approaches, one can apply a variety of general search algorithms. However, as the search space can be large, the computational times are typically very high. While embedded approaches can be a good balance as they fit well with the forecasting algorithm, not all forecasting algorithms come with embedded FS approaches (Saeys et al. 2007). In their FS review, Guyon and Elisseeff provide a very comprehensive guideline to select the appropriate features in a step-by-step fashion. This first involves using domain knowledge to determine influences that are known to be relevant, then feature construction and feature extraction techniques (e.g., to make feature interactions available to the model), assessment of the variable characteristics in a filter approach (e.g., *variable ranking*), using a simple wrapper or embedding method (computationally not expensive), using a more expensive wrapper or embedding method, and, finally, applying a subsampling (e.g., *bootstrapping*) approach (Guyon and Elisseeff 2003).

A detailed look at the demand forecasting applications of FS shows that most studies apply a mixture of domain knowledge feature selection and implicitly used filter approaches. The work of Hyndman and Fan and Adamowski and Karapatakin provide good examples of how domain knowledge and basic statistical tests can lead to a reasonable feature set (Hyndman and Fan 2010; Adamowski and Karapataki 2010). Huang et al., Liu et al., and Law et al. recommend using *correlation* and *mutual information*-based filter approaches in demand forecasting studies (Huang et al. 2021; Liu et al. 2020b; Law et al. 2019).

The utilization of optimization approaches such as FS wrappers is also common and many authors apply general optimization techniques that will also be reviewed in Section 3.8.2 (Hyper-parameter optimization). For instance, Sheikhan and Mohammadi apply *particle swarm optimization* in electricity demand forecasting (Sheikhan and Mohammadi 2013). We also observe some examples of *genetic algorithms* for feature selection in demand forecasting, for example in the work of Jiang et al. and Bouktif et al. (Bouktif et al. 2018; Jiang et al. 2017).

Ke et al. apply a random forest approach for feature selection even though they use a neural network approach as the actual forecasting model. Using the structure of fitted forests of decision trees is a popular approach, known as *feature importance* (Ke et al. 2017). Further examples are derived from the work of Bouktif et al., Zheng et al., and Rezaali et al. (Zheng et al. 2017; Bouktif et al. 2018; Rezaali et al. 2021).

In general applications, studies apply neural networks for feature selection. Li et al. propose a *deep feature selection network* that applies a regularization term to the connections between the input layer and the first hidden layer (i.e., an embedded approach) (Li et al. 2015). Kabir et al. use a neural network as a learning algorithm in a wrapper approach (Kabir et al. 2010) whereas Cheng et al. apply a random forest in a wrapper approach. However, to our knowledge, no such approaches were proposed in demand forecasting.

Table 8 summarizes the methods applied in the reviewed studies and compares them to the general FS approaches described above.

Feature selection methods in the demand forecasting literature	
Electricity	Wrapper approach combined with main or separate forecasting algorithm (Ahmad and Chen 2018; van der Meer et al. 2018; Chen et al. 2017; Cheng et al. 2017), (partial) autocorrelation function (PACF) as filter for time lagged target variable (Williams and Short 2020; AL-Musaylh et al. 2018a; AL-Musaylh et al. 2018b; Qiu et al. 2017; Ren et al. 2016)
Travel & transportation	Wrapper approach combined with main or separate forecasting algorithm (Xie et al. 2021; Zhang et al. 2020; Law et al. 2019), embedded approach as part of main forecasting algorithm (Zhang et al. 2021), Pearson correlation filter approach (Liu et al. 2020b), random forest feature importance filter approach (Ke et al. 2017)
Water & energy (excl. electricity)	Random forest feature importance filter approach (Rezaali et al. 2021), wrapper approach combined with main or separate forecasting algorithm (Guo et al. 2018), (partial) autocorrelation function (PACF) as filter for time lagged target variable (Guo et al. 2018)
Inventory & logistics	Pearson correlation filter approach (Huang et al. 2021), wrapper approach combined with separate forecasting algorithm (Feizabadi 2022)
Other demand forecasting	Random forest feature importance filter approach (Tsao et al. 2022), wrapper approach combined with main or separate forecasting algorithm (Jiang et al. 2017; Kim et al. 2015), autoencoder approach (functions as feature extraction and selection method) (Bega et al. 2019)
Other general methods	Filter: stepwise multiple regression (Piramuthu 2004); spectral analysis for autoregressive time lag selection (Crone and Kourentzes 2010) Wrapper: general search space optimization algorithms (branch and bound search (Narendra and Fukunaga 1977)); FOCUS, ID3, MIN-FEATURES (Almuallim and Dietterich 1991); sequential forward selection (SFS) or sequential backward elimination (SBE) (Devijver and Kittler 1982; Koller and Sahami 1996) Embedding: Regularization (e.g. LASSO (Tibshirani 1996), RIDGE (Marquardt and Snee 1975)), deep feature selection)

Table 8: Summary of the applied feature selection methods by field of application and gaps to general methods applied in other forecasting disciplines.

3.7.3 Feature scaling

Forecasting algorithms require inputs that are in a certain domain. For example, Shanker et al. showed that neural networks converge faster when input scaling is applied (Shanker et al. 1996). We find multiple examples in demand forecasting literature where feature scaling is applied, for example in Li et al. or Salinas et al. (Salinas et al. 2019; Li et al. 2020b). Although various studies sometimes also refer to feature scaling as *standardization* and *normalization*, these terms have different meanings in statistics.

In this review, we do not address feature scaling in detail, as we do not find particularities in demand forecasting literature feature scaling. However, it should be noted that this approach exists and is also required in demand forecasting (Han et al. 2012).

All fields of applications apply scaling methods in some form and most frequently observe *min-max standardization* for nominal and *one-hot encoding* for categorical variables in the studies while mean normalization is less common and *log normalization* is especially prevalent in travel & transportation studies. The type of scaling is mainly determined by the type of data and the mathematical requirements of the forecasting algorithm. However, we note that almost 80% of the reviewed studies do not state whether scaling was applied, which makes it difficult for other researchers to reproduce the authors' approaches.

Table 9 provides an overview of the methods applied in the reviewed studies.

Feature scaling methods in the demand forecasting literature	
Electricity	Min-max standardization (Li et al. 2020b; AL-Musaylh et al. 2018b; AL-Musaylh et al. 2018a; Williams and Short 2020; Wang et al. 2020; Tan et al. 2020; Pérez-Chacón et al. 2020; Bendaoud et al. 2021; Son and Kim 2020; Qiu et al. 2014; Ryu et al. 2017), one-hot encoding (Tan et al. 2020)
Travel & transportation	Min-max standardization (Yao et al. 2018a; Ke et al. 2017), one-hot encoding (Liu et al. 2020b), log-normalization (Wen et al. 2021a; Assaf et al. 2019; Volchek et al. 2019)
Water & energy (excl. electricity)	Not specified in detail in the studies
Inventory & logistics	Min-max standardization (Cai et al. 2021; Huang et al. 2021; Abbasimehr et al. 2020; Huber and Stuckenschmidt 2020; Güven and Şimşir 2020), log-normalization (Huber and Stuckenschmidt 2020), mean standardization (Bandara et al. 2019; Huber and Stuckenschmidt 2020)
Other demand forecasting	Min-max standardization (Murray 2020), one-hot encoding (Shakarami et al. 2021)

Table 9: Comparison of the applied feature scaling methods by field of application and gaps to general methods applied in other forecasting disciplines.

3.7.4 Summarizing the application of feature engineering

We largely draw the same conclusions concerning the use of feature construction, extraction, selection, and scaling. Overall, studies in demand forecasting apply many state-of-the-art methods to optimize the input for their forecasting models while the fields of applications favor some methods that especially suit their applications (e.g., transformations for electricity forecasting). However, we observe that most studies do not systematically apply feature engineering methods and many studies rather base their model input on expert skill or trial-and-error. While this is certainly a viable way to produce performant forecasting models, it makes it challenging to determine the impact on the forecasting accuracy of the features and methods that are applied.

3.8 The Core of Demand Forecasting: Model Design

Most studies consider model design as the major issue in demand forecasting. This section reviews the key topics around the forecasting model, including the selection of appropriate candidate models (Section 3.8.1), the optimization of the models (Section 3.8.2), and the evaluation of the forecasting results (Section 3.8.3). We assume a basic understanding of the following forecasting models and refer to the original articles and more technical reviews for detailed studies. We conclude Section 3.8 with guidelines on how to select an appropriate model from the multitude of available models (Section 3.8.4).

3.8.1 Forecasting models

This section contains a brief introduction to the historical development of forecasting algorithms whereby it becomes evident that paradigms have shifted from *a-priori* models that make strong assumptions about the data generating processes towards *data-driven* approaches. The latter deduces the data generating process from the data itself. We then provide an overview of the particularities of the industries involved and subsequently present the latest state-of-the-art forecasting models.

Historically, the first applications of the least squares method for linear regression date back centuries as these were already described and applied by Legendre and Gauss in the early 19th century (Legendre 1805; Gauss 1823). In the 1950s, in the face of increasing globalization, the complexity of markets and products rose as goods were increasingly exported globally and the variety of goods dramatically increased. In addition, supplier-driven markets turned into customer-driven markets. All of these aspects resulted in the requirement for companies to predict demands in more detail. *Smoothing models* were already applied in 1957 when Brown predicted the demand for gasoline for the automotive market of the USA (Brown 1957). More complex models that incorporate seasonality, trends, or auto regression followed in subsequent decades including, among others, *ARMA* and *ARCH* and their derivatives (Box and Jenkins 1979; Engle 1982; Bollerslev 1986). *Exponential smoothing* was also improved and adjusted to the needs of demand forecasting, as shown in *Holt-Winters* and *Croston's method* (Holt 2004; Croston 1972; Lapide 2009). *ML* methods have also emerged during the last third of the 20th century and although the preliminary foundations for ANN were already laid by Turing in 1948 (Turing 1948) their practical applications were only available in the late 1960s (Ivakhnenko and Lapa 1966). Due to technical and methodical imperfections, the approaches remained relatively unnoticed until approaches to mitigate, among others, the vanishing and exploding gradient problems were found by the end of the 2000s (Schmidhuber 2015). Until that time, other ML technologies received more attention, such as *support vector machines* (SVM) (Cortes and Vapnik 1995), and *decision trees* and their improved variants such as *random forests* (Ho 1995) and *gradient boosted trees* (Friedman 2001). Since the 2010s, ANN use has drastically increased (Torres et al. 2021; Hyndman 2020).

	Electricity		**Travel & transport**		**Water & energy**		**Logistics & inventory**		**Other**		**Total**	
	a) Main prop. alg.	b) All algorithms	a) Main prop. alg.	b) All algorithms	a) Main prop. alg.	b) All algorithms	a) Main prop. alg.	b) All algorithms	a) Main prop. alg.	b) All algorithms	a) Main prop. alg.	b) All algorithms
Naive	0%	14%	0%	24%	0%	0%	0%	18%	0%	0%	0%	13%
Regression	3%	34%	0%	35%	15%	38%	0%	18%	0%	71%	3%	35%
Smoothing	0%	9%	0%	6%	0%	0%	6%	29%	0%	14%	1%	11%
Holt-Winter	0%	6%	0%	0%	0%	0%	0%	12%	0%	14%	0%	6%
ETS	0%	0%	0%	12%	0%	0%	0%	24%	0%	0%	0%	7%
Croston	0%	0%	0%	0%	0%	0%	0%	29%	0%	0%	0%	6%
SBA	0%	0%	0%	0%	0%	0%	0%	29%	0%	0%	0%	6%
TSB	0%	0%	0%	0%	0%	0%	6%	24%	0%	0%	1%	4%
AR	0%	0%	0%	12%	0%	8%	0%	0%	0%	14%	0%	4%
ARIMA	9%	43%	12%	82%	0%	23%	0%	35%	0%	71%	6%	48%
ARCH	0%	17%	0%	24%	0%	31%	0%	18%	0%	14%	0%	20%
Other statistical	6%	6%	6%	6%	0%	0%	0%	0%	14%	14%	4%	4%
Grey model	0%	0%	0%	6%	8%	15%	0%	24%	0%	0%	1%	8%
k-nearest neighbor	0%	0%	0%	0%	0%	0%	0%	6%	0%	0%	0%	1%
SVM	23%	43%	12%	53%	15%	23%	0%	18%	0%	0%	13%	34%
Random forest	3%	11%	0%	12%	0%	23%	6%	6%	0%	14%	2%	12%
XGBoost	0%	3%	0%	18%	0%	0%	0%	0%	0%	0%	0%	4%
GBRT	0%	0%	0%	6%	0%	0%	6%	6%	14%	14%	2%	3%
ANN	23%	66%	35%	82%	31%	77%	41%	53%	71%	71%	34%	67%
RNN	0%	3%	0%	12%	0%	0%	12%	18%	0%	0%	2%	7%
LSTM	20%	23%	18%	18%	8%	8%	24%	29%	0%	0%	17%	19%
GRU	0%	3%	0%	0%	8%	8%	0%	0%	0%	0%	1%	2%
CNN	0%	6%	6%	12%	0%	0%	0%	0%	0%	0%	1%	4%
ELM	3%	3%	0%	6%	15%	23%	0%	0%	0%	0%	3%	6%
DBN	6%	9%	0%	6%	0%	0%	0%	0%	0%	0%	2%	4%
GAN	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%
Ensemble	0%	11%	12%	12%	0%	8%	0%	12%	0%	14%	2%	11%

Table 10: Overview of the proposed main algorithm in the reviewed literature. The values denote the percentage of studies from one field of application (column) that apply the algorithm (row). The left cell entry a) shows the value for the main proposed algorithm, while the right cell entry b) indicates when the study applies the algorithm as a benchmark.

Table 10 shows which forecasting algorithms studies recommend as the main approach (left entry) and which algorithms they implement and compare overall (right entry) in the different fields of application. Overall, we observe that in the reviewed literature, the studies predominantly use statistical models as benchmark algorithms whereby the ML algorithms SVM and ANN (especially LSTM) make up most of the proposed algorithms.

We also observe that not all industries have adopted the forecasting models to the same degree and that the use of models is industry specific.

Smoothing models (e.g., Holt-Winter's and Croston's method) are frequently used in logistics & inventory demand forecasting only which is due to the demand characteristics in inventory forecasting as intermittent demands with significant periods of zero demand characterize the demand in inventory forecasting. Croston especially developed his method for intermittent demand.

Instead of predicting values for all time periods (with a high probability for zero and only a few observations greater than zero), Croston's method uses exponential smoothing to forecast inter-demand intervals and a separate smoothing approach for the demand value in time intervals with a demand greater than zero. In (Teunter et al. 2011), Teunter et al. could show that Croston's method is fundamentally biased and proposed an adapted approach (Teunter et al. 2011). Even though the approach is successfully tested in another article by Teunter and Duncan (Teunter and Duncan 2009), other approaches achieve good results in forecasting intermittent demands. For example, in 24 real-world use cases Gutierrez et al. demonstrate that their ANN approach is superior to Croston's method (Gutierrez et al. 2008).

Surprisingly, studies in travel & transportation apply ARIMA and other autoregressive models relatively frequently in addition to regression models with econometric variables (Song and Li 2008; Goh and Law 2011). In a recent review of travel demand forecasting, Song et al. assume that time series methods with exogenous (econometric) inputs are preferred by researchers because of the conclusions that can be drawn from the models, namely how external factors influence demand (Song et al. 2019).

Despite the perception derived from popular science and the media, ML approaches are not new in demand forecasting, and generally, we observe that the overall number of mentions of ANN in demand forecasting has drastically increased since the 2000s. For example, Cubero already describes the application of a feedforward neural network in the context of a water demand forecast in 1991 (Cubero 1991). However, the example also illustrates the progress of ANN. The study concludes that the neural network forecast just reached a par with the compared Box-Jenkins model forecast. Later studies report further progress and improvements. For example, Law demonstrates an ANN that outperforms classical approaches in travel demand forecasting by (Law 2000) using the backpropagation technique for network training (Law 2000). Al-Saba and El-Amin come to the same conclusion concerning electricity demand forecasting in (Al-Saba and El-Amin 1999) (Al-Saba and El-Amin 1999). Recent studies apply more advanced evolutions of ANN (see Table 11 for exemplary applications in demand forecasting):

- <u>Long short-term memory (LSTM)</u> ANNs are a type of recurrent neural networks (RNN). RNN generally apply recurrent connections to maintain a temporal state in the network and hence they are suitable for time series forecasting. Early examples of RNN in the 1980s were Elman networks (Elman 1990) and Hopfield networks (Little 1996; Hopfield 1982). LSTM are specialized RNN, with nodes that can control their internal memory through a system of so-called gates, and were first proposed by Hochreiter and Schmidhuber in 1997 (Hochreiter and Schmidhuber 1997). Gated recurrent unit networks (GRU) operate on a similar principle, albeit with an alternative approach (2014 (Cho et al. 2014)).
- <u>Convolutional layer neural networks (CNN)</u> are neural networks that apply convolution operations to calculate a feature map from the inputs. They are widely used for image processing and for time series forecasting. CNN were first proposed by Lecun and Bengio in 1995 (Lecun and Bengio 1995).
- <u>Wavelet neural networks (WNN)</u> use a wavelet transformation before the time series signal is fed into the neural network which technically could also be described as a type of preprocessing. WNN were first proposed by Zhang and Benveniste in 1992 (Zhang and Benveniste 1992).

- Random vector functional link (RVFL) and extreme learning machines (ELM) are similar approaches that apply single hidden layer ANNs. In contrast to other ANN approaches, these have randomized and fixed weights for the connections of input layers to hidden layers. As an advantage, the authors state that both RVFL and ELM do not require training as the non-fixed weights can be determined analytically instead, which requires less computational effort. There is a controversial discussion amongst researchers on whether ELM are a special form of RVFL or a separate type. While the authors who first described ELM see their approach as unique (Huang 2015), other researchers consider it as a type of RVFL (e.g. (Lipo P. Wang and Chunru R. Wan)). RVFL were first proposed by Schmidt et al. in 1992 (Schmidt et al. 1992) and Pao et al. in 1994 (Pao et al. 1994), and ELM by Huang et al. in 2006 (Huang et al. 2006).
- Deep learning approaches do not constitute a single type of network but subsume ensembles of different networks, typically with a large number of hidden layers. CNN and RNN can also be deep learning networks. The term deep learning was first used by Dechter in 1986 (Dechter 1986).
- Deep belief networks (DBN) are special deep neural networks that can be used for supervised and unsupervised learning. In the reviewed literature, they typically consist of restricted Boltzmann machines (RBM).

However, Darbellay and Slama also provide an apt example that ANN are not always the best solution. In their use case of electricity demand forecasting, they point out that the underlying data is predominantly linear, wherefore neural networks cannot be significantly more accurate than classical methods (Darbellay and Slama 2000). This finding is confirmed in a review of studies by Zhang et al. which summarizes scenarios in which ANN do not exhibit advantages over classical statistics despite their potentials: (1.) linear problems with limited complexity; (2.) limited amount of training data; (3.) comparably high computational effort; and (4.) complex problems that require complex networks which are difficult to tune and prevent overfitting (Zhang et al. 1998).

The large number of studies that apply linear regression, auto regression models, and ANN could lead to the false perception that these models are superior to the other approaches. However, relevant examples also show that other ML approaches such as SVM and random forests can outperform the popular approaches. Herrera et al. demonstrate a use case to predict water demand whereby they compare ANN (plain feed-forward), smoothing, SVM, random forests, and a naïve approach. They conclude that SVM, random forests, and smoothing are equally good approaches for their use case and are significantly better than ANN and the naïve approach. Chen and Wang come to a similar conclusion in their study in (Chen and Wang 2007). In a tourism demand forecasting use case, they compare ARIMA, SVM, and ANN (feed-forward). The authors tune the SVM parameters using a genetic algorithm (GA) and they also conclude that it is not the ANN model but the SVM that exhibits the best results (Chen and Wang 2007). However, we note that the ANN models in the two studies both only possess one layer with less than ten nodes, and more complex ANN models, as applied in more recent studies (e.g., Ke et al. (Ke et al. 2017)) might potentially provide better results.

Linear regression	
	Basic literature: derivatives e.g. Ordinary Least Squares (OLS); Partial Least Squares (PLS); Spectral decomposition; Fuzzy regression (Heshmaty and Kandel 1985) **No. of parameters**: Mostly non-parametric (OLS/PLS)
Auto regression	
	Basic literature: ARMA (Box and Jenkins 1979), ARCH (Engle 1982), derivatives e.g. ARIMA (Box and Jenkins 1979); SARIMA; ARMAX; VARMA; GARCH (Bollerslev 1986) **No. of parameters**: ARIMA: p +q; SARIMA: +season; ARMAX: +exog; ARCH: q; GARCH: p+q
Smoothing	
	Basic literature: examples e.g. Holt Winter's (Holt 2004; Winters 1960); Croston's (Croston 1972); TSB (Teunter et al. 2011) **No. of parameters**: Holt Winter's: 3 ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$) + 1 (cycle length); Croston's: 2 (smoothing parameters of demand size and inter-demand interval/demand probability)
Grey model	
	Basic literature: Deng (Deng 1989) **No. of parameters**: 2 (a,b)
Support vector machines (SVM)	
	Basic literature: Cortes and Vapnik (Cortes and Vapnik 1995) **No. of parameters**: Hyper-parameters: 2 (regularization, choice of kernel function) + x (kernel function parameters)
Decision trees	
	Basic literature: derivatives e.g. Random forest (Ho 1995); Gradient boosted random trees (GBRT) (Friedman 2001) **No. of parameters**: Hyper-parameters: Typically ten or more, depending on implementation
ANN	
	Basic literature: derivatives e.g. Multi-layer perceptron (MLP); Convolutional neural networks (CNN) (Lecun and Bengio 1995); Long-short term memory (LSTM) (Hochreiter and Schmidhuber 1997) and gated recurrent units (GRU) (Cho et al. 2014); Wavelet networks (WNN) (Zhang and Benveniste 1992); Randomized vector functional link (RFVL) (Lipo P. Wang and Chunru R. Wan; Schmidt et al. 1992) and extreme learning machines (ELM) (Huang 2015); Deep Learning (Dechter 1986) **No. of parameters**: Hyper-parameters: High number in topology choice (# layers, # nodes, node type) as well as optimizer-parameters and more

Table 11: Hierarchical overview of the forecasting models with the foundational studies of the technique.

We also highlight the study by Burger et al. that applies the unusual method of symbolic regression in combination with a genetic algorithm (Burger et al. 2001).

Studies show that combinations of models improve forecasting quality over the results of single models 13,215 and the literature sometimes interchangeably refers to such models as *composite, hybrid, ensemble,* or *combined models*. In a tourism demand forecasting use case, Oh and Morzuch use four different approaches and average the forecasts. They find that the averaged result is always better than the least accurate model and sometimes even better than the best (Oh and Morzuch 2005). Tiwari and Adamowski train separate WNN on randomly sampled subsets of the data and use the combined forecasting results of the separate WNN to forecast water demand (Tiwari and Adamowski 2013). Hyndman and Fan combine two models, whereby the first forecasts annual data and the second half-hour data (Hyndman and Fan 2010). We find several further studies that combine models to ensembles using multiple datasets for comparison, for example, Qiu et al. and Qiu et al. (Qiu et al. 2017; Qiu et al. 2014).

We observe that many studies compare their elaborate and complex proposed forecasting algorithm to simple versions of state-of-the-art approaches. As Table 10 shows, most studies compare their proposed algorithms to naïve approaches, simple linear regression, or basic autoregressive or ARIMA approaches. However, few studies compare their results to other recent elaborate approaches (e.g., the latest deep ANN, gradient boosting machine, or LSTM paired with elaborate feature engineering methods). We refer to this point in Section 3.9.1.

3.8.2 Hyper-parameter optimization

In this section, we review optimization algorithms that are either used for forecasting model hyper-parameter optimization (HPO) in particular, optimization of all parameters within the demand forecasting process, or any subset of these parameters.

To use concise terminology, we differentiate between the terms *parameters* and *hyper-parameters*. During model fitting, one adjusts model parameters to fit the model to a certain problem or data set, for example smoothing coefficients in exponential smoothing or seasonality coefficients in ARMA models. In the context of classical statistical models, the literature uses the term *parameter estimation* for the process of fitting the model parameters to the data set, whereas in ML, researchers rather use the terms *model training* or *model fitting*. In the context of ANN, this especially comprises the process of adjusting the weights of neural connections. In contrast, hyper-parameters are external model parameters used for the configuration of the model and not estimated by the data set, i.e., a regularization hyper-parameter.

Table 11 shows the number of parameters and hyper-parameters for each forecasting model. We observe that complex approaches such as ANN exhibit a large number of hyper-parameters, for example, compared to SVM. Furthermore, ANN are especially sensitive to slight changes in their parameterization and they react almost capriciously. Furthermore, while the research yielded systematic analytical approaches to estimate the parameters of classical statistical models such as ARIMA/ARCH, Holt Winter's, and Croston's (plus their derivatives), there is little knowledge about the complex relationships between the forecasting performance and hyper-parameters of ML algorithms – most prominently relating to ANN.

Before taking a closer look at HPO approaches, we highlight the difference between ML HPO and ANN HPO in particular. All ML models exhibit a set of hyper-parameters that we can use to finetune the model and the training process. With ANN, we must also consider the network topology itself as a major hyper-parameter. This includes the entire structure of the neural network including which type of neuron to use (e.g., regular perceptrons, LSTM, GRU, CNN, etc.), which activation function per node, which connections, how many layers with how many neurons, and several other aspects. The literature also calls this approach a *neuronal architecture search* (NAS) while the general search for the best hyper-parameters is called HPO.

As described in Section 3.7, studies apply optimization algorithms to both HPO and feature engineering. For example, Escalante et al. apply *particle swarm optimization* (PSO) to the problems of feature selection, feature extraction, feature scaling, and forecasting model optimization (Escalante et al. 2009). El-Telbany and El-Karmi apply PSO but not on hyper-parameters but to update weights and biases in ANN (El-Telbany and El-Karmi 2008).

In the literature, optimization algorithms exhibit two components, namely the *exploration* of parameter space yet unknown to the algorithm and *exploitation* of what was already found to be a good (yet potentially not optimal) area in the parameter space. Good algorithms find an optimal balance between the two: exploitation to quickly dive into minima in the search space, and exploration to ensure not getting lost in local minima too early.

Most studies apply **explorative** techniques such as *grid search* and *random search*. In grid search, discrete values are defined for each parameter to be tested and during the optimization, each parameter value combination is tested, which defines a grid in the parameter space. Random search is the randomized version of grid search whereby any value from the parameter domain can be chosen and not only the predefined steps of the grid. Bergstra and Bengio point out that random search is generally more efficient than grid search (Bergstra and Bengio 2012) although evidently, neither of these approaches takes the structure of the problem into account which makes them easy to implement even though they have no capability for systematic exploitation. More advanced approaches adopt biology-inspired methods. PSO uses a model of swarm animals (e.g., birds) and their behavior, while *evolutionary algorithms* make use of basic evolutionary concepts such as survival of the fittest, reproduction, and mutation. Many studies apply PSO and evolutionary algorithms and report that they offer a good balance between exploitation and exploration and several studies report the successful implementation of the two approaches.

The comparison of the state-of-the-art in demand forecasting HPO with AutoML shows that demand forecasting studies apply comparably few of the available HPO techniques. Four remarkable streams in the HPO literature are Bayesian optimization, simultaneous optimistic optimization (SOO), reinforcement learning (RL), and gradient descent. The recent development in AutoML shows that model-based optimization approaches (e.g., Bayesian optimization and SOO) can optimize very complex hyper-parameter spaces.

- <u>Bayesian optimization</u> is a popular approach in HPO. It assumes a Gaussian process surrogate model of the relationship between parameter space and model performance whereby each tested and evaluated parameter combination adds a known point to the Gaussian process model. Iteratively, one can then determine parameter combinations that are likely to exhibit good model performance. In this way, Bayesian optimization exploits known parameter combinations very efficiently. A few examples of the application of Bayesian optimization can be found in the demand forecasting literature, e.g. Candelieri et al. (Candelieri et al. 2019).
- <u>Simultaneous optimistic optimization</u> is a tree-based branch-and-bound approach proposed by Munos in (Munos 2011).
- <u>Reinforcement neural networks</u> inductively search the parameter space for optimal parameter combinations. The general concept of a reinforcement neural network is that the network takes situation-based actions and receives rewards or penalties depending on the reaction of the environment. The network thereby approximates the relationship between

the parameters and rewarding function. The same principle can be applied in HPO, where the network learns the relationship between hyper-parameters and forecasting results.

Table 12 structures the proposed and applied HPO approaches in the reviewed demand forecasting studies and provides the theoretical basis for the approaches. The comprehensive overview of the reviewed studies in Table 17: Comprehensive overview of all studies reviewed in this survey. shows the HPO approach applied to each study in the review.

Class	**Approach**	**Theoretical background**	**Reviewed studies**
Explorative	Grid search	Bergstra et al. (Bergstra et al. 2012; Bergstra and Bengio 2012)	(Wang et al. 2020), (Pérez-Chacón et al. 2020), (Qiu et al. 2014), (Zhang et al. 2021), (Zhang et al. 2020), (Liu et al. 2020b), (Xenochristou and Kapelan 2020), (Cheng et al. 2017), (Yue et al. 2010)
	Random search	Bergstra et al. (Bergstra et al. 2012; Bergstra and Bengio 2012)	(Zhu et al. 2021), (Abbasimehr et al. 2020), (Babai et al. 2020), (Huang 2016)
Heuristic search	Evolutionary algorithm	Castillo et al. (Castillo et al. 2000)	(Panapakidis and Dagoumas 2017), (Romano and Kapelan 2014), (Chen and Wang 2007)
	Particle swarm optimization	Escalante et al. (Escalante et al. 2009)	(Ofori-Ntow Jnr et al. 2021), (Guo et al. 2020), (AL-Musaylh et al. 2018b), (Huang 2016), (Jiang et al. 2014), (Jiang et al. 2014),
	Greedy search	Huang et al. (Huang et al. 2018)	(Powers et al. 2005) (feature selection)
Model based optimization	Bayesian optimization and sequential model-based optimization	Snoek et al. (Snoek et al. 2012), Shahriari et al. (Shahriari et al. 2016)	(Candelieri et al. 2019), (Bandara et al. 2019)
	Simultaneous Optimistic optimization	Munos (Munos 2011), Valko et al. (Valko et al. 2013)	No demand forecasting example
Reinforcement learning			No demand forecasting example
Gradient descent			(Jiang et al. 2017), (Kulshrestha et al. 2020), (Bedi and Toshniwal 2019), (Ryu et al. 2017)

Table 12: HPO approaches for ML forecasting models.

Summarizing the literature, we observe that most studies do not explicitly address HPO and some state that the hyper-parameter setup resulted from a trial-and-error or expert knowledge process.

When HPO is described, grid search is the most common approach. Particle swarm optimization constitutes the most common advanced approach that we observe in the reviewed studies. Interestingly, we note that an emphasis on HPO is observed in other disciplines of forecasting and highlight that demand forecasting could benefit from adopting these approaches. Moreover, describing the process that led to a successful model design and hyper-parameter setup is not only relevant for evaluating the results but also for learning how to perform efficient HPO.

3.8.3 Evaluation

When we evaluate and compare forecasting models, we must address two issues. First, we need to define a metric that measures the performance of the forecast. Second, we must determine what data to evaluate and how. We will show that especially the second question is more complex in demand forecasting than one might expect.

The field of demand forecasting uses metrics that are based on the forecasting error $e_t = y_t - \hat{y}_t$. **Absolute metrics**, such as the *mean absolute error* (MAE) measures use the mean operation to calculate an average error across all observations and forecasts. In principle, all error metrics are defined for the arithmetic mean, median, or geometric mean. With the *rooted mean squared error* (RMSE), one can punish extreme error values more than small errors when compared to metrics that apply absolute error measures. Again, most errors can be defined analogously with absolute or squared errors and absolute error metrics such as MAE and RMSE can best be used in a scenario with only one time series and different forecasting models that are compared, whereby the model with the lowest error score indicates the highest performance. These absolute metrics are easy to interpret from an academic as well as from a business standpoint.

Relative metrics relate the error to another value while **percentage errors** relate the error to the observation values. Other than absolute errors, percentage errors are consistent for different data sets, as shown by Chatfield after the first M competition (Chatfield 1988). However, for a long time, absolute measures were preferred by researchers and practitioners, as shown by Armstrong and Collopy (Armstrong and Collopy 1992). For example, the *mean absolute percentage error* (MAPE) is a common percentage error that is straightforward in its interpretation, as also demonstrated by Boylan and Syntetos (Boylan and Syntetos 2006). However, researchers still disagree about the symmetry of MAPE (e.g., Hyndman and Koehler (Hyndman and Koehler 2006)). Furthermore, MAPE is not suitable when the values of the observation are zero (due to undefined division by zero) and different versions of a *symmetric MAPE* (sMAPE) were developed to overcome these issues as shown in (Armstrong 1985; Makridakis 1993; Chen and Yang 2004). As pointed out by Hyndman and Koehler, these metrics also exhibit drawbacks in certain situations, which motivated the authors to propose the use of a *mean absolute scaled error* (MASE) (Hyndman and Koehler 2006) which is suitable for zero value observations (also near zero values), albeit with the disadvantage that the metric is difficult to interpret from a business standpoint. Here, the *weighted absolute percentage error* (WAPE, also called *MAD/mean ratio* in the literature) is recommended as a more straightforward metric to interpret (Kolassa and Schütz 2007).

Other metrics compare the forecasting error to the forecasting error of another forecasting model, i.e. a baseline model. An example is the *mean relative absolute error* (MRAE), which is the MAE of a forecast model result divided by the baseline model MAE. The *percentage better* (PB) metric calculates the number of times that one forecasting model exhibits lower errors than another model. Relative measures help to compare different forecasting models using varied data sets based on a common baseline. The well-known *coefficient of determination* (R^2) follows the same principle by comparing the (squared) forecasting error with the variance of the data set (i.e., taking the mean observation value as a baseline) (Adamowski et al. 2012).

It is evident that no metric is the best metric in every situation and most forecasting competitions, e.g., the M-competitions, prefer straightforward metrics such as MAE, MAPE, or PB (Makridakis et al. 1982; Makridakis et al. 1993; Makridakis et al. 2020; Makridakis and Hibon 2000). Hyndman and Koehler recommend MAE for the interpretability of single time series data sets and alternately propose MAPE for non-intermittent contexts and MASE for intermittent demands (Hyndman and Koehler 2006). In (Armstrong and Collopy 1992), Armstrong and Collopy evaluated RMSE, PB, MAPE, MdAPE, GMRAE, and MdRAE in five categories and also conclude that no metric is superior in all situations. However, they prefer RMSE, MAPE, and GMRAE for model calibration due to their high sensitivity and MdRAE and MdAPE for model selection due to their reliability (Armstrong and Collopy 1992).

In order to correctly apply evaluation metrics to obtain reliable results, studies also consider how to integrate the metric into the whole model design process, whereby the following questions depict the typical conceptual challenges of model evaluation:

- Should in- or out-of-sample data be used for evaluation?
- Can all data points be used for evaluation?
- How can data be split into different sets for testing and how can they be cross-validated?
- What happens when hyperparameter tuning comes into play?

We differentiate between in- and out-of-sample and the data that we use to fit the forecasting model is *in-sample data*, whereas other data that is not used for model fitting is *out-of-sample data*. When using out-of-sample data to evaluate the model performance, one can determine how the model generalizes the correlations learned based on the in-sample data. From the evaluation of the M competitions, Makridakis and Winkler conclude that there is a strong discrepancy between the performance of models for in-sample or out-of-sample data (Makridakis and Winkler 1989). This deliberate split between in- and out-of-sample data is also called the *train- and test-split*.

As previously highlighted, demand forecasting is analogous to time series prediction in most cases. This means that observations are not independent of each other and, on the contrary, many models explicitly assume autoregressive dependencies between observations (e.g., ARMA). The literature proposes three approaches to account for **dependent observations**. Several authors recommend *h-blocked validation* to exclude two symmetric intervals of h observations from the training set around one observation used as a test set. They assume that outside this interval, the dependence is negligible (Hart and Wehrly 1986; C.-K. Chu and Marron 1991; Burman and Nolan 1992; Burman et al. 1994). Analogously, *hv-blocked validation* is defined by two symmetric intervals of h observations excluded from training sets around an interval of v test observations (Racine 2000). McQuarrie and Tsai recommend identifying and *excluding all observations* that statistically depend on test observations from the training set (McQuarrie and Tsai 1998). One drawback of the highlighted approaches is that the time series is interrupted by the omitted observations which can prevent models from being fitted properly.

The approaches above show how to determine dependent observations between the training and test sets while the literature also provides ways to choose **representative test sets** to evaluate models.

- Static hold out: Devroye and Wagner propose to define a portion of the data as a test set throughout the entire process (Devroye and Wagner 1979) although the validation results strongly depend on this data set.
- Time series leave-one-out: One (or a p) observation is selected as a test set and while previous observations are used as training data, observations later in time are not considered at all. Hereby, the approach accounts for the fact that future observations are not available in time series forecasting and each observation is iteratively used once as a test observation (if sufficient training observations are available). This approach is analogous to the *leave-one-out* (LOO) or *leave-p-out* (LPO) approaches in non-time series forecasting (Shao 1993; Lachenbruch and Mickey 1968; Geisser 1975) whereby two determinants for variations exist: (1.) whether to always start training from the first observation (*fixed origin* (Hyndman and Athanasopoulos 2014)) or to use a window of the same length prior to the test observation (*rolling window* (Tashman 2000)) and (2.) whether to fit the model in each iteration (*recalibration policy*) or not (*update policy*) (Tashman 2000). In a comprehensive comparative study, Tashman shows the best results for the rolling window recalibration approach.
- k-fold: Each observation is allocated to one of k groups, either block-wise or randomly, and each group is used as a test set in one iteration, and as training in all remaining iterations. Even though there is a dispute over whether it is viable to validate a time series in a non-consecutive order, several studies show that there is no empirical proof that this approach is biased (Bergmeir and Benítez 2012; Bergmeir et al. 2018).

Cases that involve HPO are more complex and for these, the literature also recommends the use of independent data sets for model fitting (*training set*), HPO validation (*development set*, short *dev set*, or *validation set*), and model selection (*test set*). Here, the state-of-the-art approach is *nested cross-validation* and one can use all of the above-mentioned approaches in nested cross-validation. In an outer loop, model selection is applied with a train-test-split while in the inner loop, a train-dev-split is used. However, it is evident that the computational effort is increased by a magnitude. We find examples of this approach in demand forecasting, e.g. in Adamowski and Karapataki (Adamowski and Karapataki 2010).

Table 13, Table 14, and Table 15 represent how frequently evaluation metrics, cross-validation schemes, and significance tests have been applied in the reviewed studies.

From Table 13, we observe that, overall, MAE, RMSE, and especially MAPE are very common metrics applied in the studies. The numbers include variants of the metrics, e.g., we subsume normalized RMSE under RMSE whereby the R2 metric is employed relatively infrequently. More recent developments such as MASE and RMSSE are also rarely applied, even in studies from the last three years.

The most frequently used evaluation metrics in the reviewed studies
Percentage values indicate how many studies of the field of application used the metric. Several studies use multiple metrics. Not all observed metrics are listed in this table. Refer to Table 17 for a detailed overview of all methods used per study.

	MAE	MAPE	MASE	RMSE	R2
Electricity	39%	83%	5%	63%	15%
Travel & transportation	45%	90%	0%	80%	20%
Water & energy	40%	47%	0%	47%	27%
Logistics & inventory	35%	55%	25%	35%	5%
Other	40%	30%	0%	30%	10%

Table 13: Evaluation metrics applied in the reviewed studies.

With respect to cross-validation (see Table 14), the basic train-test split is dominant in the reviewed studies. In *travel & transportation* demand forecasting, we more frequently observe rolling window time series validation approaches, while this approach is rather uncommon in other fields of application. Overall, considering that proper cross-validation is eminent to ensure representative results and prevent data leakage from training to testing, it is surprising that little emphasis is placed on the description of the validation scheme. Moreover, we observe that especially in studies where hyperparameter-optimization is applied (i.e., especially ML studies), the validation set is often either rather superficially described or not at all.

Application of cross-validation schemes in the reviewed studies
Percentage values indicate how many studies of the field of application used the scheme. Several studies use multiple schemes. Refer to Table 17 for a detailed overview of all methods used per study.

	train-test split	train-test split + validation split	rolling window	k-fold	any kind of cross-validation
Electricity	75%	20%	7%	12%	95%
Travel & transportation	60%	10%	35%	5%	100%
Water & energy	85%	27%	7%	7%	87%
Logistics & inventory	65%	10%	5%	10%	95%
Other	59%	10%	0%	0%	80%

Table 14: Cross-validation schemes applied in the reviewed studies.

In addition to evaluation metrics and cross-validation, significance tests are another element in the repertoire of demand forecasting evaluation. Significance tests can be applied in demand forecasting to underline the fact that the samples that are taken and the number of randomized experiments in empirical studies to compare two or more approaches are sufficiently large to consider them as not being the result of lucky sampling. The comparative study by Koning et al. systematically shows how statistical testing is applied to compare the contestants of the M3 forecasting competition. The authors especially observe that multiple comparisons must be corrected for alpha inflation (Koning et al. 2005).

In 2007 and 2008, the demand forecasting community argued whether such significance tests are suitable for application in empirical studies. Some influential researchers highlighted the difficulties with significance testing in the case studies as the tests are frequently not applied correctly or misinterpreted, thereby providing a false sense of security. This is illustrated in, among others, the studies by Armstrong, Goodwin, or Kostenko and Hyndman (Armstrong 2007; Kostenko and Hyndman 2008; Goodwin 2007).

We observe that only some of the reviewed studies apply significance tests at all (see Table 15), which may be a reaction to the abovementioned disagreements in the field. The Diebold-Mariano test is the only test that is occasionally applied while the Friedman test, t-test, and Kolmogorov–Smirnov tests are rarely found in the reviewed studies. Furthermore, correction for alpha-inflation in multiple hypothesis testing is seen even more rarely.

Application of statistical result significance tests in the reviewed studies
Percentage values indicate how many studies of the field of application used the test. Some studies use multiple tests. Refer to Table 17 for a detailed overview of all methods used per study.

	Diebold-Mariano test	Friedman test	t-test	Kolmogorov–Smirnov test
All studies	20%	7%	2%	2%

Table 15: Significance tests applied in the reviewed studies.

3.8.4 How to choose an appropriate model

Throughout this section, it becomes evident that various models can be applied and that the literature cannot provide a recipe for the best procedure. However, we find general principles that help researchers to identify appropriate approaches.

Section 3.8.1 shows the vast number of models that studies apply in demand forecasting. This broad choice may seem overwhelming when taking the first steps in this area of research. As a general guideline, studies name Occam’s razor as a parsimony principle to follow. In brief, when we have the choice between two models that yield the same results, it is reasonable to choose the less complex model and hence it is advisable to first apply models of low complexity. The number of hyper-parameters is an indicator of the model complexity although it is not the only one. This idea is also reflected by the evaluation metrics, for example, AIC punishes models by the number of parameters (Blumer et al. 1987).

Researchers should also consider previous studies in their area of research. As shown in Section 3.8.1, some forecasting models are especially well-suited for specific problem classes, for example, Croston’s method for intermittent demands. In their study, Petropoulos et al. demonstrate how seven characteristics of a time series affect forecasting accuracy in demand forecasting and which models are suitable depending on the characteristics (Petropoulos et al. 2014).

Furthermore, it is a common standard in demand forecasting to develop, implement, and then thoroughly compare different models. In this step, it is important to use a bias-free validation method to compare the forecasting results as shown in Section 3.8.3.

3.9 Summary of the Findings and Open Issues

This section provides an overview of the recent open issues in the demand forecasting literature and we thus first summarize the findings and open issues identified in the previous sections before comparing the findings to those highlighted in the reviewed studies and especially in existing surveys on demand forecasting (Section 3.3.2 and Table 4).

3.9.1 Summary of open issues from the previous sections

We summarized and compared the target variables and additional inputs used in the studies in Section 3.5. We observe that the fields of applications usually apply comparable inputs to reach one forecasting goal and propose that studies should aim to collect and assess all additional inputs available and describe why some have been sorted out in the process (see next paragraph). As pointed out in Section 3.5, studies should also describe their data set's properties in detail to allow other researchers to compare their results.

Section 3.7.4 concludes the findings from feature engineering and scaling in which we highlighted that although we can find most state-of-the-art methods in demand forecasting studies, most studies do not systematically apply the available feature engineering methods and describe why they rule out some of them and use others.

In Section 3.8.1, we summarized the forecasting algorithms used in the fields of application in demand forecasting. We can conclude that the studies generally use a broad range of algorithms and follow the principles of scientific work by comparing their proposed approaches to state-of-the-art methods. However, as highlighted in this section, most studies rather compare their proposed method to basic variants of the state-of-the-art approaches than to other elaborate methods described in recent studies on similar topics.

In addition, we observe that studies mainly review related work in their field of application instead of reviewing cross-application.

In Section 3.8.2, we observe that most studies do not describe the application of systematic methods for HPO or apply trial-and-error to optimize their hyper parameters. As indicated in this section, it would be helpful for the research community to know what steps led to a successful HPO. In addition, we recommend the use of HPO methods, for example, the PSO or other related methods described in this Section 3.8.2 to improve the approaches' results.

Section 3.8.3 shows that the reviewed studies mostly apply an appropriate set of basic evaluation metrics (MAE, MAPE, RMSE) while overall, scaled metrics are less common. We observed that the studies typically use variations of the basic evaluation metrics. In this context, we highlight two issues that can arise. Firstly, studies need to be clear on how they apply cross-validation, especially when using algorithms that require HPO, and secondly, studies should report their evaluation results using different metrics, so that the results are easily comparable, and the reader can observe influences and biases of specific metrics.

Taken together, all the points that are addressed make it challenging to systematically compare studies with their data sets, approaches, and results on a quantitative basis, and perhaps for this reason, we do not find meta-analysis comparing studies on a large scale – except for the forecasting competitions introduced in Section 3.3.2. These competitions ensure the same parameters for all contestants: identical data set, availability of additional inputs, and results based on the same evaluation metrics and cross-validation standards.

The existing surveys that we outline in the next section confirm several of the findings of this section.

3.9.2 Open issues identified in the demand forecasting literature

This section summarizes the open issues from existing surveys in demand forecasting from Table 4. We compare the findings between the fields of applications in Table 16.

The major open issues identified by the existing surveys in the context of **data availability** not only involve the availability of data as such but also the incorporation and the necessary methods to deal with big data. Song et al. recommend exploiting consumer behavior data on a micro-level to increase the forecasting quality in travel & transportation studies. Other than the aggregate econometric models used, micro-level models can derive more detailed consumer behavior insights (Song et al. 2019). Aslam et al. and Meade and Islam highlight the availability of quality big data sources for research as an open issue in demand forecasting (Meade and Islam 2015; Aslam et al. 2021).

All surveys suggest that the researchers should incorporate more detailed **input data** in their approaches whereby the data specified depends on the application, for example, Ahmad et al. and Raza and Khosravi recommend advanced meteorological and climatical features for electricity demand forecasting (Raza and Khosravi 2015; Ahmad et al. 2020) and Ghalehkhondabi et al. for energy demand forecasting (Ghalehkhondabi et al. 2017a); Ghalehkhondabi et al. and Song and Li propose multi-level seasonality (Song and Li 2008; Ghalehkhondabi et al. 2019); and Pinçe et al. prefer anomalies detection and model incorporation in supply chain forecasting (Pinçe et al. 2021).

Pinçe et al. also indicate that demand forecasting research can benefit from a better description of the input data used (Pinçe et al. 2021) while Mediavilla et al. recommend the systematic use of feature engineering techniques (Mediavilla et al. 2022). Our findings from Section 3.9.1 support both of their arguments.

With regard to the **improvement** of the main **forecasting algorithms**, many surveys advocate the hybridization of existing approaches as a promising direction for future research (Raza and Khosravi 2015; Sison et al. 2021; Ghalehkhondabi et al. 2019; Song et al. 2019; Ghalehkhondabi et al. 2017a; Ghalehkhondabi et al. 2017b). The surveys also suggest increasing research in ANN (including all variants) and to master the complexity of ANN models (Ghalehkhondabi et al. 2017a; Ghalehkhondabi et al. 2017b; Mediavilla et al. 2022) and ML approaches in general (Masdari and Khoshnevis 2020). Consistent with our findings from Section 3.9.1, Zhu et al.

observe that an increased standardization of algorithms and their description in studies can improve the generation of insights by promoting the comparability of studies (Zhu et al. 2019).

In the area of **evaluation** of results, we observe two major findings in the surveys. Sison et al., Ghalehkhondabi et al., Pinçe et al., and Seyedan and Mafakheri identify the need to introduce evaluation metrics that place a greater emphasis on the practical impact of forecasting errors, e.g., cost impacts (Seyedan and Mafakheri 2020; Ghalehkhondabi et al. 2017b; Sison et al. 2021; Pinçe et al. 2021). The surveys also emphasize that standardized error measures can help to compare results between studies (Pinçe et al. 2021; Masdari and Khoshnevis 2020), which aligns with our findings in Section 3.9.1.

Finally, the surveys agree on identifying the restrictions imposed by data privacy amongst the **trends** for future research in demand forecasting (Sison et al. 2021; Zhu et al. 2019) and in federated learning (Masdari and Khoshnevis 2020; Sison et al. 2021).

Electricity	**Travel & transportation**	**Water & energy**	**Logistics & inventory**	**Other**
Data availability				
General availability of big data for research (Aslam et al. 2021)	Development and application of data quality assurance methods (Ghalehkhondabi et al. 2019), use and availability of big data on micro-level (Song et al. 2019)		Incorporation of big data information (installed base, expert knowledge) (Pinçe et al. 2021) and agriculture and transportation (Aamer et al. 2020)	General data availability for research (Meade and Islam 2015)
Additional inputs				
Short term: regional influences; mid-/long term climate change(Ahmad et al. 2020) and meteorological factors (Raza and Khosravi 2015), smart grid and smart building inputs (Raza and Khosravi 2015)	Incorporation of qualitative inputs (Ghalehkhondabi et al. 2019); multi-level seasonality (Ghalehkhondabi et al. 2019; Song and Li 2008); extreme events (Ghalehkhondabi et al. 2019; Song and Li 2008), consumer behavior on micro-level (Song et al. 2019)	Saturation factors (Ghalehkhondabi et al. 2017a), advanced climate influences (Ghalehkhondabi et al. 2017a)	Additional inputs (Mediavilla et al. 2022), systematical feature engineering (Mediavilla et al. 2022), comparable description of input data (Pinçe et al. 2021), identification and processing of anomalies (Pinçe et al. 2021)	Incorporation of big data (Zhu et al. 2019), long horizon seasonal variations (Masdari and Khoshnevis 2020)
Forecasting improvement				
Hybridization of algorithms (Raza and Khosravi 2015)	ML development (Sison et al. 2021); hybridization of algorithms (Sison et al. 2021; Ghalehkhondabi et al. 2019; Song et al.	Hybridization of algorithms (Ghalehkhondabi et al. 2017a; Ghalehkhondabi et al. 2017b), mastery of ANN	Incorporation of whole supply chain into forecasting (Mediavilla et al. 2022; Pinçe et al. 2021; Seyedan and Mafakheri 2020; Aamer et	Standardization and comparability of forecasting models (Zhu et al. 2019), improvement of ML models (Masdari and Khoshnevis 2020),

	2019); adoption of techniques from other forecasting disciplines (Sison et al. 2021), impact of aggregation-disaggregation(Song and Li 2008; Song et al. 2019)	(Ghalehkhondabi et al. 2017a), development of more performant ANN (Ghalehkhondabi et al. 2017b), unsupervised learning (Ghalehkhondabi et al. 2017b), adoption of methods from other disciplines (Ghalehkhondabi et al. 2017b; Suganthi and Samuel 2012)	al. 2020), mastery of complex ANN (Mediavilla et al. 2022)	light weight models (e.g., for IoT) (Masdari and Khoshnevis 2020)
Evaluation				
	Incorporation of cost impact (Sison et al. 2021)	Improved error metrics (Ghalehkhondabi et al. 2017a), systematic measuring of input factor effectiveness (Ghalehkhondabi et al. 2017b)	Unification of evaluation metrics (Pinçe et al. 2021), inventory cost and performance measure (Pinçe et al. 2021; Seyedan and Mafakheri 2020)	Diversification of target variables and metrics (Zhu et al. 2019), improved comparability (Meade and Islam 2015)
Recommendations for actions and guidelines derived from forecast data				
Deduction of active demand management (Raza and Khosravi 2015)		Development of practical policies for planners and politicians (Suganthi and Samuel 2012)		Improvement of resource allocation based on forecast (Masdari and Khoshnevis 2020)
Trends				
New energy sources forecasting(Aslam et al. 2021)	Data privacy; federated learning (Sison et al. 2021)	Peak energy demand (Ghalehkhondabi et al. 2017a)		Data privacy (Zhu et al. 2019)

Table 16: Summary of the open issues from the existing demand forecasting surveys.

3.9.3 Critical discussion

The overview of the literature on demand forecasting exhibits a very **heterogeneous picture** as when comparing industries and applications, some are shown to be deeply evolved, while others are only superficially studied and each industry applies characteristic approaches which are often not observed in other industries. For example, in electricity demand forecasting, ANN are predominantly used (Hernandez et al. 2014) while, in contrast, travel demand forecasting studies mainly apply classical time series and econometric models (Song et al. 2019). However, when comparing studies from the same industry, the forecasting approaches are often very similar and almost uniformly apply similar techniques. This also coincides with the authors of the studies in question, who typically dedicate their research to certain topics where they advance knowledge and research intensively and deeply whereby entirely novel approaches are rare. This provides an impetus for future cross-disciplinary studies that assess the transfer of forecasting approaches from one industry to another. As indicated in the previous sections, the underlying theoretical research from other disciplines such as inter alia time series forecasting and AutoML uses approaches that demand forecasting studies have not yet utilized even though they provide potential solutions to issues in demand forecasting that have proven useful in other general contexts.

From our findings in Section 3.9.1 and other authors' statements summarized in Section 3.9.2, we conclude that research in demand forecasting can benefit from a **systematic framework** that makes it possible to transparently compare future studies in each process step from input data to final results without restricting scientific research. In this context, we especially emphasize the significance of a) a detailed description of the data set, b) a detailed description of all data processing, feature engineering, etc., steps, and c) a detailed evaluation of results with common and application-relevant evaluation measures. Even the deliberate omission of methods in b) can help other researchers to understand why some methods were not relevant when the rationale is adequately explained. Systematically probing the state-of-the-art methods for each step will make it easier for both readers and authors to track which methods were chosen and which have been rejected.

We contribute to this development with the generalized process structure introduced in Section 3.4. At best, the demand forecasting research community should be able to conduct meta-study-like comparisons, as Koning et al. did for the M3 competition (Koning et al. 2005) – even when the studies were not conducted in the same context (e.g., a competition).

3.10 Conclusion and Contribution

In this study, we provide a universal framework for the classification of studies and a comprehensive overview of the state-of-the-art in demand forecasting. To our knowledge, no recent survey provides a comparable overview of demand forecasting. In this study, we applied a systematic literature review approach by using four research questions and a reproducible reviewing process.

In detail, this study's framework structures the literature by (1) the fields of applications in demand forecasting (see RQ1, Section 3.3.2) and (2) the generalized process of demand forecasting (see RQ2, Section 3.4).

Based on this framework, this study systematically reviews 116 studies from five major fields of application throughout the process of demand forecasting. Hereby, this study describes the state-of-the-art methods in detail, compares the fields of application, and demonstrates exemplary alternative approaches from other forecasting disciplines (see RQ3, Sections 3.5 to 3.8).

Moreover, this survey systematically identifies the open issues through the individual process steps and application comparison, compares the open issues raised by other existing surveys from the applications of demand forecasting, and finally provides an outline of open issues and future developments in demand forecasting (see RQ4, Section 3.9).

Thereby, this survey contributes to the research in demand forecasting. Firstly, it provides a comprehensive and intuitive overview of the current state for researchers who are new to this discipline. Secondly, this paper introduces a structure to compare and classify studies. Thirdly, the survey presents and compares the recent state-of-the-art in demand forecasting and it highlights and summarizes open issues while providing an outlook for future research directions.

3.11 Appendix

#	Reference	Target variable and additional inputs	Preprocessing	Forecasting algorithm	Evaluation
Electricity					
1	(Velasquez et al. 2022)	Target variable: Energy demand (120-month ahead) per region Additional input: Region, date & time	Preprocessing: (Partial) autocorrelation analysis (P)ACF for seasonality detection	Algorithms: Regression, Exponential Smoothing, ARIMA* HPO: not required	Metric: ME, MAE, MSE, MPE, MAPE Cross validation: train-test split (compared vs. baseline forecast of official authority instead actual values)
2	(Alasali et al. 2021)	Target variable: Electricity demand (24-half hours ahead, 7-days ahead, 12-months ahead) Additional input: Temperature, average last seasons target variable, last time period value, holiday and weekend	Preprocessing: Partial Autocorrelation Analysis (PACF) for time lagged variable determination	Algorithms: Naive seasonal, ARIMAX, ANN, stochasti ARIMAX* HPO: Not specified	Metric: MAPE Cross validation: train-test split
3	(Ofori-Ntow Jnr et al. 2021)	Target variable: Electricity demand (hourly)	Feature extraction: Discrete Wavelet Transformation (DWT)	Algorithms: Extreme Learning Machine, Radial Basis Function and Backpropagation ANN* HPO: Particle Swarm Optimization (PSO), Self-adapting Differential Evolution	Metric: MAPE, MAE, RMSE Cross validation: train-test split
4	(Mansoor et al. 2021)	Target variable: Short term electricity demand (24-hours ahead) Additional input: Date & time (incl. holidays, weekends), weather (temperature, humidity, irradiance)	Preprocessing: k-means clustering by hourly load profile patterns	Algorithms: ANN, Echo State Network (ESN)* HPO: trial-and-error	Metric: MAPE, weighted MAE, envelope-weighted MAE, RMSE, normalized RMSE Cross validation: train-test split
5	(Liu and Lin 2021)	Target variable: Electricity demand () Additional input: COVID19 (no. of tests, lockdown events), weather (temperature, wind speed), renewable energy production	No particular preprocessing specified	Algorithms: bidirectional LSTM* HPO: Trial-and-error, grid search	Metric: MSE, Mean Squared Logarithmic Error (MSLE), RMSE Cross validation: train-test split

6	(Kim and Kim 2021)	Target variable. Electric vehicle electricity demand (hourly, daily) Additional input: Date & time (incl. weekend & holiday), weather (temperature), region and charging station	Preprocessing: Box-Cox Transformation	Algorithms: Trigonometric Exponential Smoothing (TBATS), ARIMA, ANN, LSTM*	Metric: MAPE, MSE Cross validation: train-test split
7	(Dittmer et al. 2021)	Target variable: Short term electricity demand (48-hours ahead) Additional input: None specified	Preprocessing: Augmented Dickey-Fuller (ADF) for stationarity test	Algorithms: Seasonal naive, ARIMA, Trigonometric seasonality Box-Cox transformation ARMA errors Trend and Seasonal (TBATS)	Metric: MAPE, RMSE Cross validation: Not specified in detail
8	(Bendaoud et al. 2021)	Target variable: Short term electricity demand (24-hours ahead) Additional input: Weather (temperature), date & time (month, day of week)	Scaling: Min-max standardization	Algorithms: ANN, Generative Adversarial Networks (GAN)* - Conditional, Deep Convolutional, Wasserstein	Metric: MAPE, RMSE Cross validation: train-test split + validation split
9	(Tan et al. 2020)	Target variable: Ultra-short sliding block (15min) electricity demand (1-5-minutes ahead and 1-24-hours ahead) Additional input: Date & time (hour, month, day of week)	Preprocessing: Autocorrelation function (ACF) and Related Factor Analysis Scaling: Min-max standardization (numerical values), one-hot encoding (categorical values)	Algorithm: SVR, Restricted Boltzmann Machine (RBM), LSTM, Random Forest, XGBoost, LSTM with data bootstrapping, random subset sampling and boosting* HPO: Not specified	Metric: MAPE, MAE normalizedRMSE, Peak Absolute Percentage Error (PAPE) Cross validation: train-test split
10	(Wang et al. 2020)	Target variable: Short term household electricity demand response to Load Aggregator incentive programs Additional input: Weather (high-low temperature), date & time (season, day of week, weekend, time), monetary reward, baseline energy demand	Missing data treatment: Deletion of data point Feature extraction: Principal Component Analysis (PCA) Scaling: Min-max standardization	Algorithms: ANN, CNN, SVM* HPO: Grid search	Metric: MAE, APE, MAPE, RMSE Cross validation: train-test split
11	(Li et al. 2020b)	Target variable: Electricity demand (forecast horizon unclear) Additional input: not specified	Feature extraction: Adaptive Fourier Decomposition (AFD), Fast Fourier Transformation	Algorithms: Naive, ARMA, SVM+SCA*, ANN, ELM HPO: Sine Cosine	Metric: Standard Deviation Error (SDE), sMAPE, MASE

			(FFT) Scaling: Min-max standardization	Optimization Algorithm (SCA)	Cross validation: train-test split
12	(Williams and Short 2020)	Target variable: Short term (half-hourly) electricity demand (four-hours ahead)	Feature selection: Partial Autocorrelation Function (PACF) for identification of periodicity Preprocessing: Removal of trend through ordinary least square regression, smoothing of time series with simple averaging Scaling: Min-max standardization Feature extraction: Piecewise Aggregation Approximation (PAA) using mean aggregation and Symbolic Aggregate Approximation (SAX) based on cubic interpolation	Algorithms: Seasonal Naive, Holt-Winter Exponential Smoothing (ES)	Metric: MAE, MAPE, RMSE, R2 Cross validation: train-test split
13	(Pérez-Chacón et al. 2020)	Target variable: Short term (hourly) electricity demand (one-hour ahead) Additional input: Date & time (year, month, day of month&week, time, holiday)	Preprocessing: outlier removal Scaling: Min-max standardization Clustering of similar patterns with k-means++ algorithm for labeling (no aggregation)	Algorithms: ARIMA, ANN, CART decision tree, Gradient Boosted Machine, Pattern Sequence based Forecasting (PSF - sampling from similar historic sequences), big data bigPSF* HPO: Grid search	Metric: MAE, MAPE, RMSE Cross validation: train-test split + validation split
14	(Son and Kim 2020)	Target variable: Monthly electricity demand (24-months ahead) Additional input: Social (electricity price, consumer price index), weather (wind speed, temperatures, solar radiation, daylight time, cooling degrees, vapor pressure)	Scaling: Min-max standardization	Algorithms: SVR, ANN, ARIMA, linear regression, LSTM*	Metric: MAE, RMSE, MAPE, Mean Bias Error MBE, Unpaired Peak Accuracy UPA Cross validation: train-test split + validation split
15	(Chapagain et al. 2020)	Target variable: Short term half-hourly electricity demand (half-hour ahead) Additional input: Weather	Not specified in detail	Algorithms: Ordinary and General Least Square (OLS, GLS), ANN, deepANN HPO: trial-and-error	Metric: MAE, MAPE Cross validation: train-test set

		(temperature), date&time (work day vs weekend)			
16	(Choi et al. 2020)	Target variable: Electricity demand (3-days ahead)	Feature extraction: Exploitation of Mixed-Data Sampling (MIDAS) weight assignment to data points	Algorithms: MIDAS, LSTM, MIDAS-LSTM* HPO: trial-and-error	Metric: MAPE, RMSE, R2 Cross validation: train-test split Significance test: Friedman test
17	(Del Real et al. 2020)	Target variable: Electricity demand Additional input: Weather (forecasted temperature), date&time (week of year, hour, day, holiday)	Preprocessing: Eliminate outliers and null values Feature extraction: CNN for feature extraction as part of forecasting algorithm	Algorithms: ARIMA, ANN, CNN+ANN* HPO: trial-and-error	Cross validation: train-test split + validation split
18	(Yukseltan et al. 2020)	Target variable: Hourly electricity demand (one-day, one-week, one-year ahead)	Not specified in detail	Algorithms: Fourier Series Expansion*	Metric: MAPE, RMSPE
19	(Khan and Jayaweera 2020)	Target variable: Half-hourly electricity demand (one-week ahead) Additional input: Only time lagged target variable	Feature extraction: Clustering of load profiles using euclidean distance and k-means clustering for separate algorithm training per cluster on cluster averages linearized using Taylor series representation	Algorithms: Multiple linear regression, ANN HPO: not specified	Metric: MAPE Cross validation: train-test split
20	(Taieb et al. 2021)	Target variable: Half-hourly electricity demand (one-day ahead) Additional input: Geographical hierarchy	Not specified in detail	Algorithms: Kernel Density Estimation (KDE), MinT hierarchical forecasting	Metric: Continuous Ranked Probability Score (CRPS), RMSE Cross validation: train-test split + validation split Significance test: two-sample Kolmogorov–Smirnov (KS) test
21	(Bedi and Toshniwal 2019)	Target variable: Short term verage and peak load demand Additional inputs: Date & time	Data cleaning and transformation (not specified in detail) and framework specific clustering	Algorithms: SVM, ANN, RNN, LSTM* HPO: Stochastic Gradient Descend	Metric: R2, RMSE, MAPE Cross validation: time series validation
22	(Jahangir et al. 2019)	Target variable: Electricity demand for electric vehicles	Not specified in detail	Algorithms: Recurrent ANN, rough ANN* HPO: trial-and-error	Metric: MAE, MAPE, RMSE, R2 Cross validation: train-test split + validation split

23	(Johannesen et al. 2019)	Target variable: Short term (half-hour) electricity demand (30-min to 24-hours ahead) Additional input: Date&time (hour, day of week, month, season), weather (humidity, wet/dry-bulb temperature, dew-point), time lagged target variable	Not specified in detail	Algorithms: Random Forest, k-Nearest Neighbor, linear regression HPO: None, i.e., default package values	Metric: MAPE Cross validation: k-fold cross validation
24	(Hwangbo et al. 2019)	Target variable: Hourly electricity demand and supply (6-hours ahead)	Feature extraction: Empirical Mode Decomposition (EMD)	Algorithms: seasonalARIMA, LSTM, GRU LSTM, deep ANN* HPO: trial-and-error	Metric: RMSE, MASE Cross validation: Rolling window time series validation and train-test split
25	(Salinas et al. 2019)	Target variable: Five different data sets from electricity demand, automotive parts demands, car traffic Additional input: Depending on data set, not specified in detail	Feature scaling: Log normalization	Algorithms: Croston, ETS, Snyder, ISSM, RNN, autoregressive RNN DeepAR* HPO: Grid search	Metric: RMSE, Normalized Difference (ND) Cross validation: Rolling window time series validation
26	(Ahmad and Chen 2018)	Target variable: Load demand of heat pumps 7 day ahead Additional input: Temperatures, wind speed & direction, date & time	Correlation analysis for input factor selection, no further data preprocessing specified	Algorithms: Linear regression, random trees (boosted, bagged)*, ANN HPO: not specified in detail	Metric: RMSE, MSE, MAE, R2, CV, MAPE Cross validation: train-test split
27	(AL-Musaylh et al. 2018a)	Target variable: Short term electricity demand (0.5-24-hours ahead) Additional input: Time lagged target variable	Feature extraction: Partial Autocorrelation Function (PACF) for time lagged target variable Scaling: Min-max standardization	Algorithms: Multivariate Regression Spline (MARS)*, SVM*, ARIMA	Metric: RMSE, R, MAE, normalizedRMSE, normalizedMAE, Willmotts Index (WI), Nash–Sutcliffe coefficients Cross validation: train-test split Significance test: t-test
28	(Bouktif et al. 2018)	Target variable: Electricity demand (two-weeks to four-months ahead) Additional input: Weather (temperature, humidity, wind speed)	Preprocessing: Cleansing and structure change, not further specified Feature scaling: Min-max standardization	Algorithms: Linear regression, ridge regression, k-nearest neighbor, Random Forest, Gradient Boosted Trees, MLP	Metric: MAE, RMSE, Coefficient of Variation (CV) Cross validation: Rolling window time series validation

			Feature selection: Genetic Algorithm (GA) to determine optimal time lag of variables, recursive feature elimination, extra tree regression	ANN, LSTM ANN* HPO: Grid search	
29	(Xie et al. 2018)	Target variable: Electricity demand (24h-hours, one-week, one-month, one-year) Additional input: Weather (relative humidity)	Not specified in detail	Algorithms: Linear regression model with different levels of relative humidity variables as additional input, SVR, ANN	Metric: MAPE Cross validation: k-fold, rolling window time series validation
30	(AL-Musaylh et al. 2018b)	Target variable: Electricity demand (one-day, one-week, one-month ahead) Additional input: Date & time	Feature extraction: Empirical Mode Decomposition (EMD), Partial Autocorrelation Function (PACF) Scaling: Min-max standardization	Algorithms: SVR, Multivariate Adaptive Regression Spline (MARS),-M5 model tree HPO: Particle Swarm Optimization (PSO)	Metric: RMSE, MAE, relative RMSE, MAPE, Willmott's Index (WI), Legates-McCabe Index (ELM), Nash–Sutcliffe coefficients Cross validation: train-test split + validation split
31	(van der Meer et al. 2018)	Target variable: Electricity demand and generation	Feature selection: Wrapper approach within the forecasting approach, including cross validation	Algorithms: ARIMA, Gaussian Process (GP)*	Metric: MAE, MAPE, RMSE, normalized RMSE, Prediction Interval Coverage Probability (PICP), Prediction Interval Normalized Average Width (PINAW), normalized Continous Ranked Probability Score (CRPS) Cross validation: train-test split + validation split in combination with k-fold cross validation for time series
32	(Chen et al. 2017)	Target variable: Short term (2 hours ahead) electricity demand in office buildings Additional inputs: Weather (dry bulb temperature), date & time, office working hours	Missing data interpolation Feature selection by SVR-wrapper	Algorithms: Variants of averaging, regression, SVR* HPO: not specified in detail	Metric: AE, ME, MAE Cross validation: train-test split

33	(Cheng et al. 2017)	Target variable: Short term (30 min) household electricity demand Additional inputs: Date & time, weather (temperature, wind speed and bearing, dew point, humidity, precipitation, air pressure)	Feature selection with wrapper (Random Forest Recursive Filter Elimination, including assessment of weights to the features	Algorithms: SVR, Gradient Boosting Trees, LSTM* HPO: Grid search	Metric: MSE*, MAPE Cross validation: train-test split
34	(Qiu et al. 2017)	Target variable: Short term electricity demand (half-hour and one-day ahead) Additional input: Date & time	Pre-analysis: Autocorrelation (ACF) to identify lag variables Feature extraction: Empirical Mode Decomposition (EMD)	Algorithms: SVR, ANN, DBN (Deep Belief Network), Ensemble DNB, EMD-SVR, EMD-ANN, EMD Random Forest, EMD-DBN*	Metric: RMSE, MAPE Cross validation: train-test split as well as k-fold
35	(Ryu et al. 2017)	Target variable: Short term electricity load profile (24 hours ahead) Additional input: Date & time (season, day of week), weather (temperature, humidity, solar radiation, cloud, wind speed), customer information (province, industry category)	Scaling: Min-max standardization Missing data intraploation (average)	Algorithms: shallow ANN, Holt-Winters, Deep Belief Network (DBN) with Restricted Boltzmann Machines (RBM) or ReLU perceptrons* HPO: Stochastic gradient descend, not specified in detail	Metric: MAPE; RRMSE (relative RMSE, ie. percentage error) Cross validation: train-test split
36	(Zheng et al. 2017)	Target variable: Electricity demand (one-day ahead) Additional input: Weather (temperature, humidity, precipitation, wind speed), date & time (day of week, special day indicator)	Feature extraction: Empirical mode decomposition (EMD), similar day selection (SD) based on gradient boosted k-means algorithm	Algorithms: ARIMA, Back Propagation ANN, SVR, LSTM ANN* HPO: not specified	Metric: MAPE Cross validation: train-test split + validation split
37	(Amini et al. 2016)	Target variable: Electricity demand for electric cars (one-day ahead) Additional input: Vehicle battery capacity, charging speed, market share, range (each per vehicle class), expected departure time	Not specified	Algorithms: ARIMA*	Metric: MAE, MAPE, MSE Cross validation: not specified
38	(Hassan et al. 2016)	Target variable: Electricity demand (24-hour ahead)	Feature selection: Autoregression	Algorithms: ANN, Adaptive Neuro Furzzy Inference	Metric: MAPE, RMSE

		Additional input: Time lagged target variable	analysis for time lagged target variable	System (ANFIS), Hoybrid of Fuzzy Logic Inference System + Extreme Learning Machine*	Cross validation: train-test split
39	(Huang 2016)	Target variable: Electricity demand (annual) Additional input: not specified	No further preprocessing specified	Algorithms: Naive, ARIMA, Hybrid of SVR different HPO* HPO: Particle Swarm Optimization (PSO)*, Genetic Algorithm	Metric: MAPE Cross validation: rolling time series validation + separate test set
40	(Keitsch and Bruckner 2016)	Target variable: Short term electricity demand (one-hour and one-day ahead) Additional input: Weather (temperature, radiation, precipitation, clouds, wind speed), economic production index, date & time	No further preprocessing specified	Algorithms: Ensemble of base forecast algorithms (linear regression, TypeDay, ANN, SVR*) and fuzzy logic composition to consolidate model results for different seasons HPO: not specified in detail	Metric: MAPE, normalized RMSE Cross validation: train-test split
41	(Ren et al. 2016)	Target variable: Short term electricity demand (one-day ahead) Additional input: Time lagged target variable only	Feature selection: Partial Autocorrelation (PACF) for time lagged variables	Algorithms: Persistence, sARIMA, ANN, Random Forest, Single Hidden Layer Random Weights ANN*	Metric: normalized RMSE Cross validation: rolling window time series validation
42	(Felice et al. 2015)	Target variable: Electricity demand (one-/two-months ahead) Additional input: Climate forecast (temperature), region	Preprocessing: Trend removal by second-order regression model	Algorithms: linear regression, SVR HPO: not specified	Metric: R2, MAPE, MSE variants (Brier Skill Score, Reliability, Resolution) Cross validation: leave-one-out (i.e. k-fold)
43	(Qiu et al. 2014)	Target variable: Electricity demand (one-hour ahead)	Scaling: Min-max standardization	Algorithms: Ensemble of Deep Belief Network (DBN) + SVR* HPO: Grid search	Metric: RMSE, MAPE Cross validation: train-test split
44	(Shang 2013)	Target variable: Ultra-short time electricity demand (half-hours ahead) Additional input: Deliberately none, as explanatory variables are expected to be not available in ultra-short term forecasting	Preprocessing: Outlier treatment (Functional Principal Component Analysis FPCA) Feature extraction: FPCA	Algorithms: Point forecasts: Random Walk, Mean Predictor, ARIMA, SARIMA, ANN, Penalized Linear Regression (PLS)*; Prediction intervals: Bootstrapping* HPO: trial-and-error, grid search	Metric: MAPE Cross validation: train-test split + validation split
45	(Sheikhan and Mohammadi 2013)	Target variable: Electricity demand (24-hour ahead)	Feature selection: Genetic Algorithm	Algorithms: Mulit-Layer Perceptron ANN (MLP)	Metric: MAPE Cross validation: train-test split

		Additional input: Date & time (season, week, day of week, weekend indicator)	(GA), Ant Colony Optimization (ACO)	HPO: Particle Swarm Optimization (PSO)	
46	(Hyndman and Fan 2010)	Target variable: Daily electricity peak demand (one-day ahead over one year with simulated weather data) Additional input: Date&time (day of week, season, holiday), weather (temperature, wind speed, luminosity)	Not specified in detail	Algorithms: non-linear regression model*	Metric: normalized RMSE, MAPE Cross evaluation: train-test split
Travel & transportation					
47	(Xie et al. 2021)	Target variable: Cruise ship travel demand Additional input: Search query volume (cruise industry related search terms), economic indices (Purchasing Manager Index, Consumer Confidence Index, exchange rates)	Feature selection: Gravitational Search Algorithm (GSA) (wrapper approach of main forecasting algorithm) Lag determination: Pearson correlation	Algorithms: ARIMA, ANN, Radial Basis Function Networks (RBF), SVR* HPO: trial-and-error, grid search	Metric: MSE, MAPE, RMSE, Willmott's Index of Agreement (WIA) Cross validation: k-fold for parameter optimization, train-test split for time series testing
48	(Kulshrestha et al. 2020)	Target variable: Quarter yearly tourist arrivals to Singapore Additional input: Date & time (quarter year), tourist departure country GDP per capita, relative price index	Stationarity test: Augmented Dickey-Fuller test (ADF), Kwiatkowski-Phillips-Schmidt-Shin test (KPSS)	Algorithms: LSTM, SVR, Radial Basis ANN, Autoregressive Distributed Lag Model (ADLM), Bayesian Bidirectional LSTM (BBiLSTM)* HPO: Bayesian Optimization	Metric: RMSE, MAE, MAPE, ratioRMSE Cross validation: train-test split
49	(Li et al. 2020a)	Target variable: Weekly tourism demand (by tourist arrivals) to Mount Siguniang CN (1-12-weeks ahead) Additional input: Keyword volume on search engines (selected keywords), online platform customer review (numerical rating and review volume)	No preprocessing or feature engineering specified in detail	Algorithms: Seasonal Naive, ETS, ARIMA, ARIMAX, SVM, Random Forest HPO: Not specified	Metric: MAE, RMSE, MAPE, DM test Cross validation: rolling window time series validation Significance test: Diebold-Mariano

50	(Zhang et al. 2021)	Target variable: Monthly Hong Kong tourism demand (tourist arrivals) (1-6 months ahead) Additional input: Search engine keyword volume	Feature extraction: STL approach (Seasonal and Trend Decomposition using Loess) Feature selection: Attention layer as part of ANN forecasting algorithm for features and time lagged target variable	Algorithms: SVR, Gradient Boosted Regression Tree (GBRT), Deep ANN* HPO: Grid search	Metric: RMSE, MAE, MAPE Cross validation: Rolling window time series validation Significance test: Diebold-Mariano test
51	(Wen et al. 2021a)	Target variable: Monthly Hong Kong tourism demand (tourist arrivals) (one-month ahead) Additional input: Search engine keyword volume	Preprocessing: Log transformation Feature extraction: Generalized Dynamic Factor Model (GDFM)	Algorithms: Seasonal Naive, ETS, SARIMA, SARIMAX, Mixed Data Sampling (MIDAS), MIDAS-AR, MIDAS-SARIMA*	Metric: MAE, MSE, RMSPE, Theil's U statistic Cross validation: Rolling window time series validation Significance test: Diebold-Mariano test
52	(Zhang et al. 2020)	Target variable: Monthly tourist arrivals (one-month ahead) Additional input: Region, search engine keyword volume (Google Trend)	Feature extraction: Seasonal trend decomposition, Dynamic Time Warping (DTW) in combination with clustering Feature selection: Time lagged target variable selection through attention layer as part of forecasting deep network	Algorithms: Multivariate Exponential Smoothing (MES), ARIMA, SVR, XGBTR), Deep Learning ANN, Group Pooling deep ANN (GP-DLM)* HPO: Grid search	Metric: RMSE, MAPE Cross validation: Rolling window time series validation ("walk-forward validation")
53	(Liu et al. 2020b)	Target variable: Taxi demand Additional input: Waether (air quality, air gas composition, wind speed, temperature, humidity, precipitation)	Missing value treatment: Deletion of data point Feature selection: Filter approach using pariwise Pearson correlation Scaling: one-hot encoding of week number	Algorithms: Random Forest, Ridge Regression, Combination Forecast* HPO: Grid search	Metric: R2, MSE, MAE Cross validation: train-test split
54	(Yao and Cao 2020)	Target variable: Tourist arrivals (one-month ahead) Additional input: Region	Feature extraction: Decomposition of trend, seasonality through Hodrick-Prescott (HP) filter Stationarity test: Augmented Dickey-Fuller (ADF) test	Algorithms: SARIMA, Neural network enhanced hidden Markov Structural Time Series Model (NehM-STSM) - trend modelling with ANN, seasonality modeling with Markov	Metric: MAPE, RMSE Cross validation: Rolling window time series validation and train-test split

				process* HPO: Not specified	
55	(Huang and Hao 2021)	Target variable: Monthly tourist arrival (one-month ahead) Additional input: Search engine keyword volume (Google Trend and Baidu)	Feature selection: trial-and-error	Algorithms: Random Walk, ARIMAX, SVR, ANN, ensemble of bagged Deep Belief Network (DBN) as weak learners and SVR as ensemble learner* HPO: trial-and-error	Metric: normalized RMSE, MAPE Cross validation: train-test split Significance test: Kolmogorov-Smirnov Predictive Accuracy (KSPA)
56	(Li and Law 2020)	Target variable: Monthly zourist arrivals (one-month ahead) Additional input: Search engine keyword volume (Google Trends)	Feature extraction: Ensemble Empirical Mode Decomposition (EEMD)	Algorithm: linear regression, Autoregressive Model with Explanatory variables (ARX)	Metric: MAPE, Improvement Ratio (IR) Cross validation: Rolling window time series validation
57	(Geng et al. 2019)	Target variable: Ride hail demand Additional input: Date & time, geospatial network, points-of-interest	Data cleaning and preprocessing not specified Framework specific graph representation generation	Algorithm: Regression, Autoregressive model, Gradient boosted trees, ANN (Convolutional-Graph Network)* HPO: not described in detail	Metric: RMSE, MAPE Cross validation: train-test split
58	(Ke et al. 2019)	Target variable: Ride hail demand Additional input: Date & time, geospatial network, demand/supply, traffic, weather	Feature extraction of additional inputs, no further data preprocessing specified	Algorithms: Regression, ANN, Gradient boosted trees, hexagonal-CNN* HPO: not specified in detail	Metric: R2, RMSE; MAE, MAPE Cross validation: train-test split
59	(Assaf et al. 2019)	Target variable: Travel demand (tourist arrivals) (1-4 quarter-years ahead) Additional input: Region, relative consumer price index, exchange rate, GDP	Preprocessing: Log-transformation	Algorithms: ARMA, VAR, BVAR, GVAR, Bayesian Global Vector Autoregressive Model (BGVAR)*	Metric: RMSE, MAE, MAPE Cross validation: Not specified Significance test: Diebold-Mariano (DM) test
60	(Law et al. 2019)	Target variable: Monthly, travel demand (tourist arrivals) (one-month ahead) Additional input: Search engine keyword volume (Google Trends, Baidu)	Feature selection: Filter approach using Pearson correlation and Maximal Information Coefficient (MIC), as well as Attention layers as part of the forecasting algorithm	Algorithms: seasonal Naive, ARIMAX, ARIMA, SVR, ANN, deepANN HPO: Not specified	Metric: RMSE, MAE, MAPE Cross validation: Rolling time series cross validation
61	(Xu et al. 2018)	Target variable: Bike sharing demand per area, next time step of	Technical description of raw data processing, no further	Algorithms: Simple, average, ARIMA, XGBoost, SVM,	Metric: MAPE Cross validation: train-test split

		10-30 minutes Additional input: weather (temperature, dew point, wind, clouds, rain), air quality, adjacent area demands, points-of-interest	preprocessing methods specified	ANN, RNN, LSTM-ANN* HPO: Not specified	
62	(Yao et al. 2018a)	Target variable: Taxi demand Additional input: Location	Scaling: Min-max standardization Feature selection: CNN (spatial features), LSTM (temporal features), graph of Dynamic Time Warping (DTW) time series similarity	Algorithms: Historical Average (HA), ARIMA, linear regression, MLP ANN, XGBoost, STResNet, Deep Multi-View Spatial-Temporal Network (DMVST-Net)* HPO: Not specified	Metric: MAPE, RMSE Cross validation: train-test split + validation split
63	(Volchek et al. 2019)	Target variable: Tourist arrivals (1-6-months ahead) Additional inputs: Search engine keyword volume (Google Trend)	Preprocessig: First order differencing to achieve stationarity, logarithm Stationarity test: Augmented Dickey-Fuller (ADF), Phillips-Perron (PP), Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests, Canova-Hansen (CH) seasonality test	Algorithms: (seasonal) Naive, ARMA, ANN, ARMAX/ARIMAX-MIDAS	Metric: RMSE, MAPE Cross validation: Rolling window time series validation
64	(Ke et al. 2017)	Target variable: On-demand ride hailing demand per area Additional input: Date & time (including peak / non-peak indicator, weekends) travel time speed, weather (temperature, humidity, wind speed, visibility)	Feature selection: Random Forest feature importance (filter approach) Feature scaling: Min-max standardization	Algorithms: simple average, ARIMA, ANN, LSTM, CNN, XGBoost, Convolutional LSTM ANN (FCL-Net)*	Metric: RMSE*, R2, MAE, MAPE Cross validation: train-test split
65	(Claveria et al. 2015)	Target variable: Tourism	Stationary test: Augmented Dickey-Fuller test, Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test	Algorithms: Multi Layer Perceptron (MLP) ANN*, Radial Basis Function (RBF) ANN*, Elman RNN	Metric: MAPE, RMSE Cross validation: train-test split + validation split
66	(Jiang et al. 2014)	Target variable: Train transportation passenger demand (one-day ahead) Additional input:	Preprocessing: Detrended Fluctuation Analysis (DFA) Feature extraction: Empirical Mode Decomposition (EMD)	Algorithms: ARIMA, SVM, Grey SVM*+EMD HPO: Particle Swarm Optimization (PSO)	Metric: MAPE, R2, RMSE Cross validation: train-test split

		Date & time (national holiday)			
67	(Carson et al. 2011)	Target variable: Air travel demand Additional input: Geo-economic factors (unemployment rate, population growth, kerosine and crude oil prices), date & time (month of year)	Preprocessing: Logit transformation and first order differentiation of variables	Algorithms: Linear regression econometric model on different aggregation levels	Metric: MAE, RMSE Cross validation: Rolling window time series validation
68	(Hyndman et al. 2011)	Target variable: Travel demand (tourist arrivals, one to eight-steps ahead) Additional input: Purpose of travel, area, capital city	No preprocessing steps stated	Algorithms: Reconciled top-down / bottom-up ARIMA*	Metric: MAPE Cross validation: Rolling window time series validation Significance tests: Pairwise t-tests
Water & energy					
69	(Raza et al. 2022)	Target variable: Total domestic long term (annual) energy demand Additional input: Population (total & growth rate), GDP (total & growth rate in different top level industries)	Only data gathering, no preprocessing specified	Algorithms: Scenario-based econometric regression	No error measurement applied
70	(Kurek et al. 2021)	Target variable: Heating distribution demand (72-hours ahead) Additional input: Date & time, region, weather (temperature, wind speed & direction, humidity, solar radiation, clouds),	Missing data interpolation: averaging/zero Feature construction: Time lagged rolling window means for weather data variables, squares and roots of variables Rule-set based data validation	Algorithms: Linear regression, ridge regression, Autoregression with exogenous variables (ARX), Fuzzy model, partially linear model, ANN HPO: trial-and-error	Metric: MAPE, R2, MAE Cross validation: train-test split
71	(Rezaali et al. 2021)	Target variable: Short term water demand (1-24-hours ahead) Additional input: Weather (temperature, wind speed, clouds), date & time (holiday indicator)	Time lagged variables: Partial Autocorrelation Function (PACF) Feature selection: Random Forest feature importance measure Feature extraction: Discrete wavelet transformation	Algorithms: ANN, SVM, Regularized Extreme Learning Machine (RELM), Random Forest HPO: trial-and-error	Metric: RMSE for point forecast; Prediction Interval Coverage Probability (PICP), Prediction Interval Average Width (PIAW), Average Coverage Error (ACE) for probability interval forecasting Cross validation: train-test split

72	(Zubaidi et al. 2020)	Target variable: Monthly water demand	Feature extraction: Singular Spectrum Analysis (SSA)	Algorithms: Autoregressive model (AR)* HPO: Grid search	Metric: R, MAE, MSE, Fitness Cross validation: train-test split
73	(Guo et al. 2020)	Target variable: Annual water demand of agriculture, industry and residents (one-year ahead) Additional input:	Not specified in detail	Algorithms: Linear model with linear, exponential and logarithmic coefficients HPO: Whale Optimization Algorithm (WOA), improved WOA with social learning strategy	Metric: Relative Error (RE), Mean RE Cross validation: train-test split
74	(Sun and Zhao 2020)	Target variable: Short term wind power generation (15-60-min ahead) Additional input: not specified in detail	Feature extraction: Variational Mode Decomposition (VMD)	Algorithms: ANN, LSTM, Elman, Convolutional LSTM* HPO: authors' expertise / trial-and-error	Metric: RMSE, MRE, MAE, MSE Cross validation: train-test split + validation split
75	(Xenochristou and Kapelan 2020)	Target variable: Water demand Additional input: Date&time (season, day of week), location (post code), weather (sunshine hours, temperature, humidity, day with no rain)	Preprocessing: Removal of errors, null values, outliers	Algorithms: ANN, Generalized Linear Model, Random Forest, Gradient Boosting Machine, deep ANN HPO: grid search and random search	Metric: R2, MSE, MAPE Cross validation: train-test split + k-fold cross validation for HPO
76	(Candelieri et al. 2019)	Target variable: Hourly water demand Additional inputs: Date & time	Clustering by cosine similarity	Algorithms: SVM HPO: Parallel Global Optimization	Metric: MAPE
77	(Wang et al. 2018)	Target variable: Long term energy demand Additional input: None, only univariate time series	Not specified in detail	Algorithms: Grey GM(1,1), ARIMA, Rolling Metabolic Grey Model (MGM), Adaptive Network-based Fuzzy Inference System (ANFIS)	Metric: MSE, MAPE, Mean Square Percent Error (MSPE) Cross validation: train-test split and rolling window time series validation
78	(Guo et al. 2018)	Target variable: Short term water demand (15-min, 24-hours ahead) Additional input: None, only time lagged univariate target variable	Feature selection: Grid search time lagged target variable in validation data set, (Partial) Autocorrelation Function (P/ACF)	Algorithms: ARIMA, ANN, Gated Recurrent Unit GRU ANN* HPO: trial-and-error, grid search	Metric: MAE, MAPE, RMSE, Nash-Sutcliffe model efficiency (NSE) Cross validation: train-test split + validation split
79	(Panapakidis and Dagoumas 2017)	Target variable: Gas demand (1 day ahead) Additional input: Region, date & time, weather (temperature)	Pre-analysis: Correlation analysis of target variables from different regions as well as autoregression	Algorithms: Hybrid of Adaptive Feuro-Fuzzy Inference System (ANFIS) and ANN* HPO: Genetic	Metric: AE, MAE, MAPE, RMSE, MARNE (Mean Absolute Range Normalized Error)

			Feature extraction: Wavelet transformation (continuous and discrete)	Algorithm (for ANFIS), Levenmerg-Marquardt (ANN)	Cross validation: train-test split
80	(Izadyar et al. 2015)	Target variable: Natural gas heating demand (one-month ahead) Additional input: Date & time (month number) weather (outdoor temperature)	No further preprocessing specified	Algorithms: Extreme Learning Machine (ELM)*, ANN, Genetic Programming HPO: not specified	Metric: RMSE, R2, Person correlation coefficient Cross validation: train-test split
81	(Romano and Kapelan 2014)	Target variable: Water demand (24-hour ahead) Additional input: Date & time (time of day, day of week)	Preprocessing: Missing data interpolation (average)	Algorithms: ANN* HPO: Evolutionary Algorithm	Metric: MAPE, MSE, Nash-Slutcliffe Index Cross validation: train-test split + validation split
82	(Adamowski et al. 2012)	Target variable: Daily water demand (one-day ahead) Additional input: Weather (precipitation, max. temperature,	Feature extraction: Discrete Wavelet Transformation (DWT)	Algorithms: linear regression, non-linear regression, ARIMA, ANN, wavelet ANN HPO: trial-and-error	Metric: R2, Nash-Sutcliffe efficiency coefficient, RMSE, relative RMSE Cross validation: train-test split + validation split
83	(Adamowski and Karapataki 2010)	Target variable: Water demand (one-week ahead) Additional input: Weather (maximum temperature, rainfall)	No preprocessing steps stated	Algorithms: Multiple linear regression, ANN (Levenberg-Marquardt*, resilient back-propagation, Powell-Beale) HPO: trial-and-error	Metric: R2, RMSE, MAE Cross validation: train-test split + validation split
84	(Herrera et al. 2010)	Target variable: Water demand (1-24 hours ahead) Additional input: Weather (temperature, wind speed, precipitation, pressure), time lagged target variable	Not specified in detail	Algorithms: Projection Pursuit Regression (PPR) splines, Multivariate Adaptive Regression Splines (MARS), SVR, Random Forests, ANN HPO: Grid search	Metric: RMSE, MAE, Nash–Sutcliffe efficiency coefficient, Cross validation: train-test split with Monte Carlo simulation
Logistics & inventory					
85	(Cai et al. 2021)	Target variable: Sales volume of skin care goods Additional input: Order data features, sentiment analysis, face value calculation	Time series feature construction: Scaling: Min-max standardization and (numerical data) one-hot encoding (categorical data) Sentiment analysis: Word segmentation and mapping; neural network trained to predict star ratings from word maps	Algorithms: ARIMA, MLP-LSTM, bidirectional LSTM*	Metric: MSE, RMSE, MAE, MAPE Cross validation: train-test split for time series, else k-fold

			Face value calculation: Image recognition using VGG16 Convolutional NN to label skin health type		
86	(Huang et al. 2021)	Target variable: Last mile delivery demand Additional input: Economic indicators (GDP, incomer per capita, logistics employment, expenditures, traffic mileage, e-commerce revenues, etc.)	Scaling: Min-max standardization Feature selection: Correlation of input to target variable (filter approach) Feature extraction: Principal Component Analysis (PCA) Stationarity test of forecasting errors: Augmented Dickey-Fuller (ADF)	Algorithms: Factor Analysis, Grey Model GM(1,1), ANN* HPO: Not specified in detail	Metric: MAE, RMSE, MAPE Cross validation: train
87	(Chandriah and Naraganahalli 2021)	Target variable: Spare parts demand Additional input: Date & time (year, month), car brand, share of cars sold	No preprocessing specified in detail	Algorithms: SES, TSB, SBA, Croston, LSTM with modified ADAM optimizer*	Metric: ME, MSE Cross validation: train-test split
88	(Zhu et al. 2021)	Target variable: Pharmaceuticals demand Additional input: Drug classification, inventory at point of care, supply chain structure	Feature construction: Median demand and demand volatility clustered, time lagged target variable through extensive cross validation Feature extraction: Dynamic time warping (DTW)	Algorithms: Moving Average (MA), linear regression, Exponential Smoothing (ES), Exponential State Space Model (ETS), SVR, Random Forest, ANN, RNN* HPO: Grid search and cross validation	Metric: MAPE, normalized MSE, normalized MAE, Bias Cross validation: Rolling window time series cross validation
89	(Abbasimehr et al. 2020)	Target variable: Furniture demand (one-month ahead)	Preparation: Predictability tests (Augmented Dickey-Fuller, Maximal Lyapunov Exponent, Empirical Mode Decomposition) Scaling: Min-max standardization	Algorithms: ARIMA, ETS, ANN, K-nearest Neighbors (KNN), RNN, SVM, LSTM* HPO: Grid search	Metric: RMSE, sMAPE Cross validation: train-test split Significance test: Friedman test, Hochberg's post hoc test
90	(Huber and Stuckenschmidt 2020)	Target variable: Bakery retail goods demand (one-day ahead) Additional input: Date & time (holiday, days before / after holiday, day of year, month, etc.), location (city, nearby other stores, etc.), store and	Feature construction: statistics of target variable (seasonal median, absolute/relative change) Scaling: Log-normalization, mean-normalization	Algorithms: Seasonal naive, ETS, linear regression, LASSO regression, ANN, LSTM, Gradient Boosted Regression Trees (GBRT)* HPO: trial-and-error with stratified cross validation	Metric: sMAPE, RMSE, relative MAE, MASE Cross validation: train-test split + validation

		product class, time lagged target variable			
91	(Güven and Şimşir 2020)	Target variable: Fashion retail sales () Additional input: Product color*, weather (unclear), customer gender, date & time (special days - unclear), economic factors (un-employment rate, interest, consumer price index), etc.	Scaling: Min-max standardization	Algorithms: ANN*, SVM HPO: trial-and-error	Metric: RMSE Cross validation: Not specified
92	(Iftikhar and Khan 2020)	Target variable: Fashion goods ("shorts") demand (one-day ahead) Additional input: Social media sentiment (from Twitter and Facebook)	Preprocessing: Rule-set based filtering (more than 3 words, user < 2,000 posts, etc.) Pre-analysis:n-gram word frequency analysis Feature extraction: Latent Dirichlet Allocation for topic identification, Naive Bayes and SVM for sentiment analysis	Algorithms: Bass Emotion Model (BEM)*	Metric: Accuracy (unclear) Cross validation: Not specified
93	(Feizabadi 2022)	Target variable: Steel products demand (1-3 months ahead) Additional input: Macroeconomic variables (exports, foreign reserves, imports, industrial production, exchange rate, stock market indices)	Feature selection: Wrapper approach using stepwise improvement of linear regression model	Algorithms: Holt-Winters Exponential Smoothing, ARIMAX, ANN* HPO: trial-and-error	Metric: Accuracy defined as 1-MAPE Cross validation: train-test split + validation split
94	(Hu 2020)	Target variable: Annual industrial raw material (magnesium) demand (one-year ahead) Additional input: None specified	Not specified in detail	Algorithms: ANN, ARIMA, Fuzzy Time Series Analysis (FTS), Grey Model, grey residual modification model ANN GRA-NNGM HPO: trail-and-error	Metric: Average Percentage Error (APE), MAPE Cross validation: train-test split
95	(Khan et al. 2020)	Target variable: Sales Additional input: Date&time (week, year, holidays, sales promotions), product inventory	Not specified in detail	Algorithms: Amazon Deep AR HPO: Not specified	Metric: Accuracy (unclear) Cross validation: train-test split
96	(Babai et al. 2020)	Target variable: Intermittent spare part	Not specified in detail	Algorithms: Single Exponential	Metric: scale-free ME and MSE,

		demand (one-month ahead		Smoothing (SES), Croston, SBA, Wilemain's bootstrapping, Zhou and Viswanathan’s bootstrapping, SBA inspired ANN with Bayesian regularization* HPO: Grid search	MASE Cross validation: train-test split + validation split
97	(Bandara et al. 2019)	Additional input: Date & time (incl. holiday, weekend, season, day of week), product information (type, category)	Missing value treatment: Heuristic identification of false zero demands and imputation by last observation value Scaling: Mean standardization for numerical values, one-hot encoding for categorical values Clustering of products by product demand characteristics and k-means clustering algorithm	Algorithms: Naive Seasonal, Exponential Weighted Moving Average (WEMA), Exponential Smoothing (ETS), ARIMA, Facebook Prophet, LSTM* HPO: Bayesian Global Optimization	Metric: modified-MAPE Cross validation: train-test split
98	(Fu et al. 2018)	Target variable: 1 week ahead semi-conductor demand Additional input: Date & time, product demand properties (recency, frequency, quantity, coefficient of variation)	Preprocessing steps not specified	Algorithms: Moving average, Croston, TSB, SBA, RNN, hybrid of SBA+RNN* HPO: not specified in detail	Metric: RMSE, MAE, MASE Cross validation: train-validate-test split
99	(Li and Lim 2018)	Target variable: Fashion product demand 1 day ahead Additional input: Product-store combination, date, public holidays	Feature extraction: Statistics over time series (length, non-zero demand share, average demand interval) No further preprocessing specified	Algorithms: Croston, TSB, SBA, SES, MAPA; ADIDA, moving average, Hybrid of Holt-Winters + simple exponential smoothing + greedy aggregation heuristic* HPO: not required	Metric: MASE, RMASE, MAPE, MdAPE, MAE Cross validation: train-test split
100	(Kim et al. 2017)	Target variable: Spare part demand (after end of production) Additional input: average product lifetime, sales of product in initial and mature phases, warranty period, spare part cost	No additional preprocessing specified	Algorithms: Exponential Weighed Moving Average, 4 variants of Installed Base (log linear autoregressive model)*	Metric: MAPE, RMSE Cross validation: train-test split

101	(Ferreira et al. 2016)	Target variable: Product demand in fashion with no product sales history Additional input: Product information (price, hierarchy of categories, brand, size, color, manufacturer suggested price) sales events, discount, competing products and their price	No further preprocessing specified	Algorithms: linear regression, PCA, Regression Trees (bagging)* HPO: not specified	Metric: MAE, MdAE, MAPE, R2 Cross validation: train-test split and k-fold (k=5)
102	(Jaipuria and Mahapatra 2014)	Target variable: Product demand in different supply chain case studies (one-month ahead) Additional input: Date & time	Feature extraction: Dynamic Wavelet Transformation (DWT) Feature construction: (Partial) autocorrelation (PACF) for time lagged target variable	Algorithms: ARIMA, ANN+DWT* HPO: not specified	Metric: MSE Cross validation: train-test split
103	(Zied Babai et al. 2014)	Target variable: Spare part demand (one-period ahead) Additional input: None	Not specified	Algorithms: Croston, Syntetos-Boylan (SBA), Teunter-Syntetos-Boylan (TSB)*, Single Exponential Smooting (SES), Naive, Zero Forecast	Metric: ME, MSE, MASE Cross validation: Not specified
104	(Rostami-Tabar et al. 2013)	Target variable: Intermittent spare part demand (one-step ahead)	Feature extraction: Temporal aggregation	Algorithms: AR, MA, ARMA, Single Exponential Smoothing (SES)*	Metric: MSE Cross validation: train-test split
105	(Nikolopoulos et al. 2011)	Target variable: Intermittent spare part demand (month-, quarter- and one-lead time ahead) Additional input: Lead time	Feature extraction: Temporal aggregation, aggregate-disaggregate intermittent demand approach (ADIDA)	Alorithms: Naive, SBA (with and without ADIDA*)	Metric: Bias, MASE, MSE, relativeMSE Cross validation: train-test split + validation split
106	(Gutierrez et al. 2008)	Target variable: Electronics demand Additional input: Not specified in detail	No preprocessing steps stated	Algorithms: Croston, SBA, SES, ANN* HPO: no optimization / trial-and-error	Metric: MAPE, relative geometric RMSE, Percentage Best (PB) Cross validation: train-test split
Other					
107	(Shakarami et al. 2021)	Target variable: Mobile edge computing capacity (delay, energy consumption, uplink)	Preprocessing: Noise injection, missing value interpolation, outlier detection Scaling: Encoding and scaling, not specified in detail	Algorithms: multiple linear regression, Hidden Markov Model, deepANN HPO: Not specified in detail (trial-and-error)	Metric: MSE Cross validation: train-test split

108	(Etemadi et al. 2020)	Target variable: Cloud computing workload	Not specified in detail	Algorithms: linear regression, Autoregressive (AR), Moving Average (MA), ARMA, ARIMA, ANN	Metric: RMSE, MAE, MAPE, R2 Cross validation: train-test split + validation split
109	(Tsao et al. 2022)	Target variable: Sales volume of servers (computing) Additional input: Google Trend search results, date & time	Preprocessing: Autocorrelation analysis (ACF) Feature selection: Gini impurity index (aka feature importance, filter approach)	Algorithms: Holt-Winters Exponential Smoothing, ARIMA as baseline models, "Intelligent Forecasting Approach" as hybrid of Random Forest, Gradient Boosted Regression Trees (GBRT) to classify time series based on external features into three characterizing classes (peak, valley, normal) and regression to determine actual forecast values per class* HPO: trial-and-error and grid search	Metric: MSE, MAE, RMSPE Cross validation: train-test split
110	(Dreger and Wolters 2014)	Target variable: Money demand Additional inputs: Inflation rate, nominal interest rate, GDP (nominal, real), nominal house price indicator	Stationarity test: Unit root test	Algorithms: VAR, econometric model	Metric: MAE, RMSE Cross validation:
111	(Venkatesh et al. 2014)	Target variable: ATM cash demand (one-week ahead) Additional input: Date & time (day of week, seasonality)	Feature extraction: Representation time series by multiplicative time series representation Clustering: Sequence Alignment Method (SAM) and Taylor-Butina clustering	Algorithms: ARIMA, ANN, Radial Basis Function (RBF) or Generalized Regression ANN*, Wavelet ANN, Group Method of Data Handling (GMDH) HPO: not specified	Metric: sMAPE Cross validation: train-test set
112	(Murray 2020)	Target variable: Hospital bed, Intensive Care Unit (ICU) and ventilator day demand (120-days ahead) Additional input: Region, COVID measure events (school closure, business closure, stay-at-home restrictions, travel	Scaling: Min-max standardization	Algorithms: Statistical Gaussian Error Function	No out-of-sample error was calculated

		restrictions), patient age group			
113	(Jiang et al. 2017)	Target variable: number of patients, number orders per 3 different hospital resources (28 days ahead) Additional input: Date & time, weather (temperature, dew point, humidity, air pressure, wind speed and bearing, precipitation, cloud, snow)	Stationarity test (augmented Dickey-Fuller, partial autocorrelation ACF/PACF) Feature construction: mean, max, min and change rate of features Feature selection: Genetic Algorithm wrapper approach	Algorithms: GA, PCA, linear regression, ARIMAX, MLP ANN* HPO: Stochastic Gradient Descend	Metric: RMSE Cross validation: train-test split
114	(Kim et al. 2015)	Target variable: Drug demand (one-month ahead) Additional input: Social media (), wireless sensor data (fine dust concentration)	Feature extraction: topic modeling and trend analysis from text documents using Latent Dirichlet Allocation (LDA) Feature selection: Multicollinearity analysis (filter approach)	Algorithms: VARX*	Metric: not clearly specified, presumably MAPE Cross validation: train-test split
115	(Maeng et al. 2020)	Target variable: 5G mobile data broadband demand Additional input: Customer information (gender, age, monthly household income), service level (data transmission rate, data offer, Internet of Things offer, additional charge)	Feature extraction: Fractional Factorial Design	Algorithm: Discrete choice model multinomial logit function	No ex-post validation
116	(Bega et al. 2019)	Target variable: Telecommunication demand (5G data volume) (5-min ahead) Additional input: Location (spatial network)	Feature selection: CNN-encoder +ANN-decoder as part of the main forecasting algorithm	Algorithms: Specialized mobile telecommunication algorithms (Infocom17, MobiHoC18), DeepCog* HPO: Not specified (trial-and-error)	Metric: MAE, MSE, actual costs incurred by over-&under-provisioning (also employed as ANN loss-function) Cross validation: train-test split

Table 17: Comprehensive overview of all studies reviewed in this survey.

3.12 References

Aamer, Ammar; Eka Yani, Luh Putu; Alan Priyatna, I. Made (2020): Data Analytics in the Supply Chain Management: Review of Machine Learning Applications in Demand Forecasting. In *OSCM: An Int. Journal*, pp. 1–13. DOI: 10.31387/oscm0440281.

Abbasimehr, Hossein; Shabani, Mostafa; Yousefi, Mohsen (2020): An optimized model using LSTM network for demand forecasting. In *Computers & Industrial Engineering* 143, p. 106435. DOI: 10.1016/j.cie.2020.106435.

Adamowski, Jan; Fung Chan, Hiu; Prasher, Shiv O.; Ozga-Zielinski, Bogdan; Sliusarieva, Anna (2012): Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. In *Water Resour. Res.* 48 (1), p. 247. DOI: 10.1029/2010WR009945.

Adamowski, Jan; Karapataki, Christina (2010): Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. In *J. Hydrol. Eng.* 15 (10), pp. 729–743. DOI: 10.1061/(ASCE)HE.1943-5584.0000245.

Ahmad, Tanveer; Chen, Huanxin (2018): Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches. In *Energy and Buildings* 166, pp. 460–476. DOI: 10.1016/j.enbuild.2018.01.066.

Ahmad, Tanveer; Zhang, Hongcai; Yan, Biao (2020): A review on renewable energy and electricity requirement forecasting models for smart grid and buildings. In *Sustainable Cities and Society* 55, p. 102052. DOI: 10.1016/j.scs.2020.102052.

Alasali, Feras; Nusair, Khaled; Alhmoud, Lina; Zarour, Eyad (2021): Impact of the COVID-19 Pandemic on Electricity Demand and Load Forecasting. In *Sustainability* 13 (3), p. 1435. DOI: 10.3390/su13031435.

Almuallim, Hussein; Dietterich, Thomas G. (1991): Learning with Many Irrelevant Features. In : Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2: AAAI Press (AAAI'91), pp. 547–552.

AL-Musaylh, Mohanad S.; Deo, Ravinesh C.; Adamowski, Jan F.; Li, Yan (2018a): Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. In *Advanced Engineering Informatics* 35, pp. 1–16. DOI: 10.1016/j.aei.2017.11.002.

AL-Musaylh, Mohanad S.; Deo, Ravinesh C.; Li, Yan; Adamowski, Jan F. (2018b): Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. In *Applied Energy* 217, pp. 422–439. DOI: 10.1016/j.apenergy.2018.02.140.

Al-Saba, Tawfiq; El-Amin, Ibrahim (1999): Artificial neural networks as applied to long-term demand forecasting. In *Artificial Intelligence in Engineering* 13 (2), pp. 189–197. DOI: 10.1016/S0954-1810(98)00018-1.

Amini, M. Hadi; Kargarian, Amin; Karabasoglu, Orkun (2016): ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation. In *Electric Power Systems Research* 140, pp. 378–390. DOI: 10.1016/j.epsr.2016.06.003.

Armstrong, J. Scott (1985): Long-range forecasting. From crystal ball to computer. 2. ed. New York: Wiley.

Armstrong, J. Scott (2007): Significance tests harm progress in forecasting. In *International Journal of Forecasting* 23 (2), pp. 321–327. DOI: 10.1016/j.ijforecast.2007.03.004.

Armstrong, J.Scott; Collopy, Fred (1992): Error measures for generalizing about forecasting methods: Empirical comparisons. In *International Journal of Forecasting* 8 (1), pp. 69–80. DOI: 10.1016/0169-2070(92)90008-W.

Aslam, Sheraz; Herodotou, Herodotos; Mohsin, Syed Muhammad; Javaid, Nadeem; Ashraf, Nouman; Aslam, Shahzad (2021): A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. In *Renewable and Sustainable Energy Reviews* 144, p. 110992. DOI: 10.1016/j.rser.2021.110992.

Assaf, A. George; Li, Gang; Song, Haiyan; Tsionas, Mike G. (2019): Modeling and Forecasting Regional Tourism Demand Using the Bayesian Global Vector Autoregressive (BGVAR) Model. In *Journal of Travel Research* 58 (3), pp. 383–397. DOI: 10.1177/0047287518759226.

Athanasopoulos, George; Hyndman, Rob J.; Song, Haiyan; Wu, Doris C. (2011): The tourism forecasting competition. In *International Journal of Forecasting* 27 (3), pp. 822–844. DOI: 10.1016/j.ijforecast.2010.04.009.

Au, Siu-Tong; Duan, Rong; Hesar, Siamak G.; Jiang, Wei (2010): A framework of irregularity enlightenment for data pre-processing in data mining. In *Ann Oper Res* 174 (1), pp. 47–66. DOI: 10.1007/s10479-008-0494-z.

Babai, M. Z.; Tsadiras, A.; Papadopoulos, C. (2020): On the empirical performance of some new neural network methods for forecasting intermittent demand. In *IMA Journal of Management Mathematics* 31 (3), pp. 281–305. DOI: 10.1093/imaman/dpaa003.

Bandara, Kasun; Shi, Peibei; Bergmeir, Christoph; Hewamalage, Hansika; Tran, Quoc; Seaman, Brian (2019): Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology. In Tom Gedeon, Kok Wai Wong, Minho Lee (Eds.): Neural Information Processing, vol. 11955. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 462–474.

Barandas, Marília; Folgado, Duarte; Fernandes, Letícia; Santos, Sara; Abreu, Mariana; Bota, Patrícia et al. (2020): TSFEL: Time Series Feature Extraction Library. In *SoftwareX* 11, p. 100456. DOI: 10.1016/j.softx.2020.100456.

Battiti, R. (1994): Using mutual information for selecting features in supervised neural net learning. In *IEEE transactions on neural networks* 5 (4), pp. 537–550. DOI: 10.1109/72.298224.

Bedi, Jatin; Toshniwal, Durga (2019): Deep learning framework to forecast electricity demand. In *Applied Energy* 238, pp. 1312–1326. DOI: 10.1016/j.apenergy.2019.01.113.

Bega, Dario; Gramaglia, Marco; Fiore, Marco; Banchs, Albert; Costa-Perez, Xavier (2019): DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning. In : IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. Paris, France, 29.04.2019 - 02.05.2019: IEEE, pp. 280–288.

Bendaoud, Nadjib Mohamed Mehdi; Farah, Nadir; Ben Ahmed, Samir (2021): Comparing Generative Adversarial Networks architectures for electricity demand forecasting. In *Energy and Buildings* 247, p. 111152. DOI: 10.1016/j.enbuild.2021.111152.

Bergmeir, Christoph; Benítez, José M. (2012): On the use of cross-validation for time series predictor evaluation. In *Information Sciences* 191, pp. 192–213. DOI: 10.1016/j.ins.2011.12.028.

Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo (2018): A note on the validity of cross-validation for evaluating autoregressive time series prediction. In *Computational Statistics & Data Analysis* 120, pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

Bergstra, James; Bengio, Yoshua (2012): Random search for hyper-parameter optimization. In *Journal of Machine Learning Research* 13 (Feb), pp. 281–305.

Bergstra, James S.; Bardenet, Rémi; Bengio, Yoshua; Kégl, Balázs (2012): Algorithms for Hyper-Parameter Optimization. In John Shawe-Taylor (Ed.): Advances in neural information processing systems 24. 25th Annual Conference on Neural Information Processing Systems 2011 ; December 12 - 15, 2011, Granada, Spain. Neural Information Processing Systems Foundation; Annual Conference on Neural Information Processing Systems; NIPS. Red Hook, NY: Curran, pp. 2546–2554. Available online at https://academic.microsoft.com/paper/2106411961.

Billings, R. Bruce; Jones, Clive J. (2008): Forecasting Urban Water Demand (2nd Edition): American Water Works Association.

Blum, Avrim L.; Langley, Pat (1997): Selection of relevant features and examples in machine learning. In *Artificial Intelligence* 97 (1-2), pp. 245–271. DOI: 10.1016/S0004-3702(97)00063-5.

Blumer, Anselm; Ehrenfeucht, Andrzej; Haussler, David; Warmuth, Manfred K. (1987): Occam's Razor. In *Information Processing Letters* 24 (6), pp. 377–380. DOI: 10.1016/0020-0190(87)90114-1.

Bollerslev, Tim (1986): Generalized autoregressive conditional heteroskedasticity. In *Journal of Econometrics* 31 (3), pp. 307–327. DOI: 10.1016/0304-4076(86)90063-1.

Bouktif, Salah; Fiaz, Ali; Ouni, Ali; Serhani, Mohamed (2018): Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches †. In *Energies* 11 (7), p. 1636. DOI: 10.3390/en11071636.

Box, George E. P.; Jenkins, Gwilym M. (1979): Time series analysis. Forecasting and control. Rev. ed., [Nachdr.]. Oakland: Holden-Day (Holden-Day series in time series analysis and digital processing).

Boylan, John; Syntetos, Aris (2006): Accuracy and Accuracy Implication Metrics for Intermittent Demand. In *Foresight: The International Journal of Applied Forecasting* 4, pp. 39–42.

Breiman, Leo (2001): Random Forests. In *Machine Learning* 45 (1), pp. 5–32. DOI: 10.1023/A:1010933404324.

Breunig, Markus M.; Kriegel, Hans-Peter; Ng, Raymond T.; Sander, Jörg (2000): LOF: identifying density-based local outliers. In Maggie Dunham (Ed.): Proceedings of the 2000 ACM SIGMOD international conference on Management of data. the 2000 ACM SIGMOD international conference. Dallas, Texas, United States, 5/15/2000 - 5/18/2000. ACM Special Interest Group on Management of Data. New York, NY: ACM, pp. 93–104.

Brown, Robert G. (1957): Exponential smoothing for predicting demand. In : Operations Research, vol. 5. INST OPERATIONS RESEARCH MANAGEMENT SCIENCES 901 ELKRIDGE LANDING RD, STE …, p. 145.

Burger, C.J.S.C; Dohnal, M.; Kathrada, M.; Law, R. (2001): A practitioners guide to time-series methods for tourism demand forecasting — a case study of Durban, South Africa. In *Tourism Management* 22 (4), pp. 403–409. DOI: 10.1016/S0261-5177(00)00068-6.

Burman, P.; Nolan, D. (1992): DATA-DEPENDENT ESTIMATION OF PREDICTION FUNCTIONS. In *J Time Series Analysis* 13 (3), pp. 189–207. DOI: 10.1111/j.1467-9892.1992.tb00102.x.

Burman, Prabir; Chow, Edmond; Nolan, Deborah (1994): A cross-validatory method for dependent data. In *Biometrika* 81 (2), pp. 351–358. DOI: 10.1093/biomet/81.2.351.

C.-K. Chu; Marron, J. S. (1991): Comparison of Two Bandwidth Selectors with Dependent Errors. In *The Annals of Statistics* 19 (4), pp. 1906–1918. Available online at www.jstor.org/stable/2241910.

Cai, Weiwei; Song, Yaping; Wei, Zhanguo (2021): Multimodal Data Guided Spatial Feature Fusion and Grouping Strategy for E-Commerce Commodity Demand Forecasting. In *Mobile Information Systems* 2021, pp. 1–14. DOI: 10.1155/2021/5568208.

Candelieri, Antonio; Giordani, Ilaria; Archetti, Francesco; Barkalov, Konstantin; Meyerov, Iosif; Polovinkin, Alexey et al. (2019): Tuning hyperparameters of a SVM-based water demand forecasting system through parallel global optimization. In *Computers & Operations Research* 106, pp. 202–209. DOI: 10.1016/j.cor.2018.01.013.

Carbonneau, Real; Laframboise, Kevin; Vahidov, Rustam (2008): Application of machine learning techniques for supply chain demand forecasting. In *European Journal of Operational Research* 184 (3), pp. 1140–1154. DOI: 10.1016/j.ejor.2006.12.004.

Carson, Richard T.; Cenesizoglu, Tolga; Parker, Roger (2011): Forecasting (aggregate) demand for US commercial air travel. In *International Journal of Forecasting* 27 (3), pp. 923–941. DOI: 10.1016/j.ijforecast.2010.02.010.

Castillo, P. A.; Carpio, J.; Merelo, J. J.; Prieto, A.; Rivas, V.; Romero, G. (2000): Evolving Multilayer Perceptrons. In *Neural Processing Letters* 12 (2), pp. 115–128. DOI: 10.1023/A:1009684907680.

Chan, S. C.; Tsui, K. M.; Wu, H. C.; Hou, Yunhe; Wu, Yik-Chung; Wu, Felix (2012): Load/Price Forecasting and Managing Demand Response for Smart Grids: Methodologies and Challenges. In *IEEE Signal Process. Mag.* 29 (5), pp. 68–85. DOI: 10.1109/MSP.2012.2186531.

Chandrashekar, Girish; Sahin, Ferat (2014): A survey on feature selection methods. In *Computers & Electrical Engineering* 40 (1), pp. 16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

Chandriah, Kiran Kumar; Naraganahalli, Raghavendra V. (2021): RNN / LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting. In *Multimed Tools Appl* 80 (17), pp. 26145–26159. DOI: 10.1007/s11042-021-10913-0.

Chang, Pei-Chann; Fan, Chin-Yuan; Lin, Jyun-Jie (2011): Monthly electricity demand forecasting based on a weighted evolving fuzzy neural network approach. In *International Journal of Electrical Power & Energy Systems* 33 (1), pp. 17–27. DOI: 10.1016/j.ijepes.2010.08.008.

Chapagain, Kamal; Kittipiyakul, Somsak; Kulthanavit, Pisut (2020): Short-Term Electricity Demand Forecasting: Impact Analysis of Temperature for Thailand. In *Energies* 13 (10), p. 2498. DOI: 10.3390/en13102498.

Chatfield, Chris (1988): Apples, oranges and mean square error. In *International Journal of Forecasting* 4 (4), pp. 515–518. DOI: 10.1016/0169-2070(88)90127-6.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. (2002): SMOTE: Synthetic Minority Over-sampling Technique. In *jair* 16, pp. 321–357. DOI: 10.1613/jair.953.

Chen, Kuan-Yu; Wang, Cheng-Hua (2007): Support vector regression with genetic algorithms in forecasting tourism demand. In *Tourism Management* 28 (1), pp. 215–226. DOI: 10.1016/j.tourman.2005.12.018.

Chen, Yongbao; Xu, Peng; Chu, Yiyi; Li, Weilin; Wu, Yuntao; Ni, Lizhou et al. (2017): Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings. In *Applied Energy* 195, pp. 659–670. DOI: 10.1016/j.apenergy.2017.03.034.

Chen, Zhuo; Yang, Yuhong (2004): Assessing forecast accuracy measures. In *Preprint Series* 2010, pp. 2004–2010.

Cheng, Yao; Xu, Chang; Mashima, Daisuke; Thing, Vrizlynn L. L.; Wu, Yongdong (2017): PowerLSTM: Power Demand Forecasting Using Long Short-Term Memory Neural Network. In Gao Cong, Wen-Chih Peng, Wei Emma Zhang, Chengliang Li, Aixin Sun (Eds.): Advanced Data Mining and Applications, vol. 10604. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 727–740.

Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014): Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Available online at http://arxiv.org/pdf/1406.1078v3.

Choi; Cho; Kim (2020): Power Demand Forecasting using Long Short-Term Memory (LSTM) Deep-Learning Model for Monitoring Energy Sustainability. In *Sustainability* 12 (3), p. 1109. DOI: 10.3390/su12031109.

Christ, Maximilian; Braun, Nils; Neuffer, Julius; Kempa-Liehr, Andreas W. (2018): Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). In *Neurocomputing* 307, pp. 72–77. DOI: 10.1016/j.neucom.2018.03.067.

Christ, Maximilian; Kempa-Liehr, Andreas W.; Feindt, Michael (2016): Distributed and parallel time series feature extraction for industrial big data applications. Available online at http://arxiv.org/pdf/1610.07717v3.

Claveria, Oscar; Monte, Enric; Torra, Salvador (2015): Tourism Demand Forecasting with Neural Network Models: Different Ways of Treating Information. In *Int. J. Tourism Res.* 17 (5), pp. 492–500. DOI: 10.1002/jtr.2016.

Cohen (2013): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences: Routledge.

Cortes, Corinna; Vapnik, Vladimir (1995): Support-vector networks. In *Mach Learn* 20 (3), pp. 273–297. DOI: 10.1007/BF00994018.

Crone, Sven F.; Hibon, Michèle; Nikolopoulos, Konstantinos (2011): Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. In *International Journal of Forecasting* 27 (3), pp. 635–660. DOI: 10.1016/j.ijforecast.2011.04.001.

Crone, Sven F.; Kourentzes, Nikolaos (2010): Feature selection for time series prediction – A combined filter and wrapper approach for neural networks. In *Neurocomputing* 73 (10-12), pp. 1923–1936. DOI: 10.1016/j.neucom.2010.01.017.

Croston, J. D. (1972): Forecasting and Stock Control for Intermittent Demands. In *Journal of the Operational Research Society* 23 (3), pp. 289–303. DOI: 10.1057/jors.1972.50.

Cubero, Robert Griñó (1991): Neural networks for water demand time series forecasting. In Alberto Prieto (Ed.): Artificial Neural Networks, vol. 540. Berlin/Heidelberg: Springer-Verlag (Lecture Notes in Computer Science), pp. 453–460.

Cybenko, G. (1989): Approximation by superpositions of a sigmoidal function. In *Math. Control Signal Systems* 2 (4), pp. 303–314. DOI: 10.1007/BF02551274.

Darbellay, Georges A.; Slama, Marek (2000): Forecasting the short-term demand for electricity. In *International Journal of Forecasting* 16 (1), pp. 71–83. DOI: 10.1016/S0169-2070(99)00045-X.

de Gooijer, Jan G.; Hyndman, Rob J. (2006): 25 years of time series forecasting. In *International Journal of Forecasting* 22 (3), pp. 443–473. DOI: 10.1016/j.ijforecast.2006.01.001.

Dechter, Rina (1986): Learning While Searching in Constraint-Satisfaction-Problems. In, pp. 178–185.

Del Real, Alejandro J.; Dorado, Fernando; Durán, Jaime (2020): Energy Demand Forecasting Using Deep Learning: Applications for the French Grid. In *Energies* 13 (9), p. 2242. DOI: 10.3390/en13092242.

Deng, Houtao; Runger, George; Tuv, Eugene; Vladimir, Martyanov (2013): A time series forest for classification and feature extraction. In *Information Sciences* 239, pp. 142–153. DOI: 10.1016/j.ins.2013.02.030.

Deng, Julong (1989): Introduction to grey system theory. In *The Journal of grey system* 1 (1), pp. 1–24.

Devijver, Pierre A.; Kittler, Josef (1982): Pattern recognition : a statistical approach.

Devroye, L.; Wagner, T. (1979): Distribution-free performance bounds for potential function rules. In *IEEE Trans. Inform. Theory* 25 (5), pp. 601–604. DOI: 10.1109/TIT.1979.1056087.

Dittmer, Celina; Krümpel, Johannes; Lemmer, Andreas (2021): Power demand forecasting for demand-driven energy production with biogas plants. In *Renewable Energy* 163, pp. 1871–1877. DOI: 10.1016/j.renene.2020.10.099.

Donkor, Emmanuel A.; Mazzuchi, Thomas A.; Soyer, Refik; Alan Roberson, J. (2014): Urban Water Demand Forecasting: Review of Methods and Models. In *J. Water Resour. Plann. Manage.* 140 (2), pp. 146–159. DOI: 10.1061/(ASCE)WR.1943-5452.0000314.

Drackley, Adam; Newbold, K. Bruce; Paez, Antonio; Heddle, Nancy (2012): Forecasting Ontario's blood supply and demand. In *Transfusion* 52 (2), pp. 366–374. DOI: 10.1111/j.1537-2995.2011.03280.x.

Dreger, Christian; Wolters, Jürgen (2014): Money demand and the role of monetary indicators in forecasting euro area inflation. In *International Journal of Forecasting* 30 (2), pp. 303–312.

Du Preez, Johann; Witt, Stephen F. (2003): Univariate versus multivariate time series forecasting: an application to international tourism demand. In *International Journal of Forecasting* 19 (3), pp. 435–451. DOI: 10.1016/S0169-2070(02)00057-2.

Elman, Jeffrey L. (1990): Finding Structure in Time. In *Cognitive Science* 14 (2), pp. 179–211. DOI: 10.1207/s15516709cog1402_1.

El-Telbany, Mohammed; El-Karmi, Fawwaz (2008): Short-term forecasting of Jordanian electricity demand using particle swarm optimization. In *Electric Power Systems Research* 78 (3), pp. 425–433. DOI: 10.1016/j.epsr.2007.03.011.

Engle, Robert F. (1982): Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. In *Econometrica* 50 (4), p. 987. DOI: 10.2307/1912773.

Escalante, Hugo Jair; Montes, Manuel; Sucar, Luis Enrique (2009): Particle swarm model selection. In *Journal of Machine Learning Research* 10 (Feb), pp. 405–440.

Etemadi, Masoumeh; Ghobaei-Arani, Mostafa; Shahidinejad, Ali (2020): Resource provisioning for IoT services in the fog computing environment: An autonomic approach. In *Computer Communications* 161, pp. 109–131. DOI: 10.1016/j.comcom.2020.07.028.

Feizabadi, Javad (2022): Machine learning demand forecasting and supply chain performance. In *International Journal of Logistics Research and Applications* 25 (2), pp. 119–142. DOI: 10.1080/13675567.2020.1803246.

Felice, Matteo de; Alessandri, Andrea; Catalano, Franco (2015): Seasonal climate forecasts for medium-term electricity demand forecasting. In *Applied Energy* 137, pp. 435–444. DOI: 10.1016/j.apenergy.2014.10.030.

Ferreira, Kris Johnson; Lee, Bin Hong Alex; Simchi-Levi, David (2016): Analytics for an Online Retailer: Demand Forecasting and Price Optimization. In *M&SOM* 18 (1), pp. 69–88. DOI: 10.1287/msom.2015.0561.

Feurer, Matthias; Klein, Aaron; Eggensperger, Katharina; Springenberg, Jost; Blum, Manuel; Hutter, Frank (2015): Efficient and Robust Automated Machine Learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.): Advances in Neural Information Processing Systems 28: Curran Associates, Inc, pp. 2962–2970. Available online at http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.

Fildes, Robert; Ord, Keith (2007): Forecasting Competitions: Their Role in Improving Forecasting Practice and Research. In Michael P. Clements, David F. Hendry (Eds.): A companion to economic forecasting. [Malden, Mass.]: Blackwell Pub, pp. 322–353.

Forestier, Germain; Petitjean, Francois; Dau, Hoang Anh; Webb, Geoffrey I.; Keogh, Eamonn (2017): Generating Synthetic Time Series to Augment Sparse Datasets. In Vijay Raghavan (Ed.): 17th IEEE International Conference on Data Mining. 18-21 November 2017, New Orleans, Louisiana : proceedings. 2017 IEEE International Conference on Data Mining (ICDM). New Orleans, LA, 11/18/2017 - 11/21/2017. IEEE International Conference on Data Mining; Institute of Electrical and Electronics Engineers; IEEE Computer Society; ICDM. Piscataway, NJ: IEEE, pp. 865–870.

Friedman, Jerome H. (2001): Greedy Function Approximation: A Gradient Boosting Machine. In *The Annals of Statistics* 29 (5), pp. 1189–1232. DOI: 10.2307/2699986.

Fu, Wenhan; Chien, Chen-Fu; Lin, Zih-Hao (2018): A Hybrid Forecasting Framework with Neural Network and Time-Series Method for Intermittent Demand in Semiconductor Supply Chain. In Ilkyeong Moon, Gyu M. editor Lee, Jinwoo Park, Dimitris Kiritsis, Gregor von Cieminski (Eds.): Advances in production management systems. Part II. Smart manufacturing for Industry 4.0 : IFIP WG 5.7 International Conference, APMS 2018, Seoul, Korea, August 26-30, 2018, Proceedings / edited by Ilkyeong Moon, Gyu M. Lee, Jinwoo Park, Dimitris Kiritsis, Gregor von Cieminski, vol. 536. Cham, Switzerland: Springer (IFIP advances in information and communication technology, 1868-4238, 536), pp. 65–72.

Fulcher, Ben D.; Little, Max A.; Jones, Nick S. (2013): Highly comparative time-series analysis: the empirical structure of time series and their methods. In *Journal of the Royal Society, Interface* 10 (83), p. 20130048. DOI: 10.1098/rsif.2013.0048.

García Valverde, Diego; Quevedo Casín, Joseba Jokin; Puig Cayuela, Vicenç; Saludes Closa, Jordi: Water demand estimation and outlier detection from smart meter data using classification and Big Data methods. In : 2nd New Developments in IT & Water Conference, 8-10 February 2015, Rotterdam (Holland), pp. 1–8.

Gauss, Carl-Friedrich (1823): Theoria combinationis observationum erroribus minimis obnoxiae: Henricus Dieterich (1).

Geisser, Seymour (1975): The Predictive Sample Reuse Method with Applications. In *Journal of the American Statistical Association* 70 (350), p. 320. DOI: 10.2307/2285815.

Geng, Xu; Li, Yaguang; Wang, Leye; Zhang, Lingyu; Yang, Qiang; Ye, Jieping; Liu, Yan (2019): Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting. In *AAAI* 33, pp. 3656–3663. DOI: 10.1609/aaai.v33i01.33013656.

Ghalehkhondabi, Iman; Ardjmand, Ehsan; Weckman, Gary R.; Young, William A. (2017a): An overview of energy demand forecasting methods published in 2005–2015. In *Energy Syst* 8 (2), pp. 411–447. DOI: 10.1007/s12667-016-0203-y.

Ghalehkhondabi, Iman; Ardjmand, Ehsan; Young, William A.; Weckman, Gary R. (2017b): Water demand forecasting: review of soft computing methods. In *Environmental monitoring and assessment* 189 (7), p. 313. DOI: 10.1007/s10661-017-6030-3.

Ghalehkhondabi, Iman; Ardjmand, Ehsan; Young, William A.; Weckman, Gary R. (2019): A review of demand forecasting models and methodological developments within tourism and passenger transportation industry. In *JTF* 5 (1), pp. 75–93. DOI: 10.1108/JTF-10-2018-0061.

Gilliland, Mike; others (2010): Defining" Demand" for Demand Forecasting. In *Foresight: The International Journal of Applied Forecasting* (18), pp. 4–8.

Goh, Carey; Law, Rob (2011): The Methodological Progress of Tourism Demand Forecasting: A Review of Related Literature. In *Journal of Travel & Tourism Marketing* 28 (3), pp. 296–317. DOI: 10.1080/10548408.2011.562856.

Goodwin, Paul (2007): Should we be using significance tests in forecasting research? In *International Journal of Forecasting* 23 (2), pp. 333–334. DOI: 10.1016/j.ijforecast.2007.01.008.

Guo, Guancheng; Liu, Shuming; Wu, Yipeng; Li, Junyu; Zhou, Ren; Zhu, Xiaoyun (2018): Short-Term Water Demand Forecast Based on Deep Learning Method. In *J. Water Resour. Plann. Manage.* 144 (12), Article 04018076. DOI: 10.1061/(ASCE)WR.1943-5452.0000992.

Guo, Wenyan; Liu, Ting; Dai, Fang; Xu, Peng (2020): An improved whale optimization algorithm for forecasting water resources demand. In *Applied Soft Computing* 86, p. 105925. DOI: 10.1016/j.asoc.2019.105925.

Gutierrez, Rafael S.; Solis, Adriano O.; Mukhopadhyay, Somnath (2008): Lumpy demand forecasting using neural networks. In *International Journal of Production Economics* 111 (2), pp. 409–420. DOI: 10.1016/j.ijpe.2007.01.007.

Güven, İlker; Şimşir, Fuat (2020): Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. In *Computers & Industrial Engineering* 147, p. 106678. DOI: 10.1016/j.cie.2020.106678.

Guyon, Isabelle; Elisseeff, André (2003): An introduction to variable and feature selection. In *J. Mach. Learn. Res.* 3 (null), pp. 1157–1182.

Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2009): The WEKA data mining software. In *SIGKDD Explor. Newsl.* 11 (1), p. 10. DOI: 10.1145/1656274.1656278.

Hall, Mark Andrew (1999): Correlation-based feature selection for machine learning.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data mining. Concepts and techniques. 3rd ed. Amsterdam, Boston: Elsevier/Morgan Kaufmann (Morgan Kaufmann series in data management systems).

Hart, Jeffrey D.; Wehrly, Thomas E. (1986): Kernel Regression Estimation Using Repeated Measurements Data. In *Journal of the American Statistical Association* 81 (396), pp. 1080–1088. DOI: 10.1080/01621459.1986.10478377.

Harzing, Anne-Wil (2007): Publish or Perish. Available online at https://harzing.com/resources/publish-or-perish.

Hassan, Saima; Khosravi, Abbas; Jaafar, Jafreezal; Khanesar, Mojtaba Ahmadieh (2016): A systematic design of interval type-2 fuzzy logic system using extreme learning machine for electricity load demand forecasting. In *International Journal of Electrical Power & Energy Systems* 82, pp. 1–10. DOI: 10.1016/j.ijepes.2016.03.001.

He, Haibo; Bai, Yang; Garcia, Edwardo A.; Li, Shutao (2008): ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In : IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008 (IEEE World Congress on Computational Intelligence) : 1-8 June, 2008, [Hong Kong, China. 2008 IEEE International Joint Conference on Neural Networks (IJCNN 2008 - Hong Kong). Hong Kong, China, 6/1/2008 - 6/8/2008. IEEE World Congress on Computational Intelligence. Piscataway, N.J.: IEEE, pp. 1322–1328.

He, Xin; Zhao, Kaiyong; Chu, Xiaowen (2019): AutoML: A Survey of the State-of-the-Art. Available online at http://arxiv.org/pdf/1908.00709v4.

Hernandez, Luis; Baladron, Carlos; Aguiar, Javier M.; Carro, Belen; Sanchez-Esguevillas, Antonio J.; Lloret, Jaime; Massana, Joaquim (2014): A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings. In *IEEE Commun. Surv. Tutorials* 16 (3), pp. 1460–1495. DOI: 10.1109/SURV.2014.032014.00094.

Herrera, Manuel; Torgo, Luís; Izquierdo, Joaquín; Pérez-García, Rafael (2010): Predictive models for forecasting hourly urban water demand. In *Journal of Hydrology* 387 (1-2), pp. 141–150. DOI: 10.1016/j.jhydrol.2010.04.005.

Heshmaty, Behrooz; Kandel, Abraham (1985): Fuzzy linear regression and its applications to forecasting in uncertain environment. In *Fuzzy Sets and Systems* 15 (2), pp. 159–191. DOI: 10.1016/0165-0114(85)90044-2.

Ho, Tin Kam (1995): Random decision forests. In : Proceedings of the third International Conference on Document Analysis and Recognition. August 14-16, 1995, Montréal, Canada. 3rd International Conference on Document Analysis and Recognition. Montreal, Que., Canada, 14-16 Aug. 1995. International Conference on Document Analysis and Recognition; International Association for Pattern Recognition. Los Alamitos, Calif: IEEE Computer Society Press, pp. 278–282.

Hochreiter, S.; Schmidhuber, J. (1997): Long short-term memory. In *Neural computation* 9 (8), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Holt, Charles C. (2004): Forecasting seasonals and trends by exponentially weighted moving averages. In *International Journal of Forecasting* 20 (1), pp. 5–10. DOI: 10.1016/j.ijforecast.2003.09.015.

Hong, Tao; Pinson, Pierre; Fan, Shu (2014): Global Energy Forecasting Competition 2012. In *International Journal of Forecasting* 30 (2), pp. 357–363. DOI: 10.1016/j.ijforecast.2013.07.001.

Hong, Tao; Pinson, Pierre; Fan, Shu; Zareipour, Hamidreza; Troccoli, Alberto; Hyndman, Rob J. (2016): Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. In *International Journal of Forecasting* 32 (3), pp. 896–913. DOI: 10.1016/j.ijforecast.2016.02.001.

Hong, Tao; Xie, Jingrui; Black, Jonathan (2019): Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. In *International Journal of Forecasting* 35 (4), pp. 1389–1399. DOI: 10.1016/j.ijforecast.2019.02.006.

Hopfield, J. J. (1982): Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the United States of America* 79 (8), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554.

Hou, Rui; Zhang, Bi-xi (2005): A Method for Forecasting Regional Logistics Demand Based on MLP Neural Network and Its Application [J]. In *Systems Engineering-theory & Practice* 12.

Hu, Yi-Chung (2020): Constructing grey prediction models using grey relational analysis and neural networks for magnesium material demand forecasting. In *Applied Soft Computing* 93, p. 106398. DOI: 10.1016/j.asoc.2020.106398.

Hua, Goh Bee (1996): Residential construction demand forecasting using economic indicators: a comparative study of artificial neural networks and multiple regression. In *Construction Management and Economics* 14 (1), pp. 25–34. DOI: 10.1080/01446199600000004.

Huang, Bai; Hao, Hao (2021): A novel two-step procedure for tourism demand forecasting. In *Current Issues in Tourism* 24 (9), pp. 1199–1210. DOI: 10.1080/13683500.2020.1770705.

Huang, Guang-Bin (2015): What are extreme learning machines? Filling the gap between Frank Rosenblatt’s dream and John von Neumann’s puzzle. In *Cognitive Computation* 7 (3), pp. 263–278.

Huang, Guang-Bin; Zhu, Qin-Yu; Siew, Chee-Kheong (2006): Extreme learning machine: Theory and applications. In *Neurocomputing* 70 (1-3), pp. 489–501. DOI: 10.1016/j.neucom.2005.12.126.

Huang, Lijuan; Xie, Guojie; Zhao, Wende; Gu, Yan; Huang, Yi (2021): Regional logistics demand forecasting: a BP neural network approach. In *Complex Intell. Syst.* DOI: 10.1007/s40747-021-00297-x.

Huang, Min-Liang (2016): Hybridization of Chaotic Quantum Particle Swarm Optimization with SVR in Electric Demand Forecasting. In *Energies* 9 (6), p. 426. DOI: 10.3390/en9060426.

Huang, Norden E.; Shen, Zheng; Long, Steven R.; Wu, Manli C.; Shih, Hsing H.; Zheng, Quanan et al. (1998): The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proc. R. Soc. Lond. A* 454 (1971), pp. 903–995. DOI: 10.1098/rspa.1998.0193.

Huang, Siyu; Li, Xi; Cheng, Zhi-Qi; Zhang, Zhongfei; Hauptmann, Alexander (2018): GNAS. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen et al. (Eds.): MM'18. Proceedings of the 2018 ACM Multimedia Conference : October 22-26, 2018, Seoul, Republic of Korea. 2018 ACM Multimedia Conference. Seoul, Republic of Korea, 10/22/2018 - 10/26/2018. New York, New York: Association for Computing Machinery, pp. 2049–2057.

Huber, Jakob; Stuckenschmidt, Heiner (2020): Daily retail demand forecasting using machine learning with emphasis on calendric special days. In *International Journal of Forecasting* 36 (4), pp. 1420–1438. DOI: 10.1016/j.ijforecast.2020.02.005.

Hwangbo, Soonho; Nam, KiJeon; Heo, SungKu; Yoo, ChangKyoo (2019): Hydrogen-based self-sustaining integrated renewable electricity network (HySIREN) using a supply-demand forecasting model and deep-learning algorithms. In *Energy Conversion and Management* 185, pp. 353–367. DOI: 10.1016/j.enconman.2019.02.017.

Hyndman, Rob J. (2020): A brief history of forecasting competitions. In *International Journal of Forecasting* 36 (1), pp. 7–14. DOI: 10.1016/j.ijforecast.2019.03.015.

Hyndman, Rob J.; Ahmed, Roman A.; Athanasopoulos, George; Shang, Han Lin (2011): Optimal combination forecasts for hierarchical time series. In *Computational Statistics & Data Analysis* 55 (9), pp. 2579–2589. DOI: 10.1016/j.csda.2011.03.006.

Hyndman, Rob J.; Athanasopoulos, G. (2014): Measuring forecast accuracy. In *Business forecasting: Practical problems and solutions*, pp. 177–183.

Hyndman, Rob J.; Fan, Shu (2010): Density Forecasting for Long-Term Peak Electricity Demand. In *IEEE Trans. Power Syst.* 25 (2), pp. 1142–1153. DOI: 10.1109/TPWRS.2009.2036017.

Hyndman, Rob J.; Koehler, Anne B. (2006): Another look at measures of forecast accuracy. In *International Journal of Forecasting* 22 (4), pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Hyndman, Rob J.; Wang, Earo; Laptev, Nikolay Pavlovich (2015): Large-Scale Unusual Time Series Detection. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1616–1619.

Iftikhar, Rehan; Khan, Mohammad Saud (2020): Social Media Big Data Analytics for Demand Forecasting. In *Journal of Global Information Management* 28 (1), pp. 103–120. DOI: 10.4018/JGIM.2020010106.

Ivakhnenko, Alekseĭ Grigorʹevich; Lapa, Valentin Grigorévich (1966): Cybernetic predicting devices. PURDUE UNIV LAFAYETTE IND SCHOOL OF ELECTRICAL ENGINEERING.

Ivanov, Dmitry; Tsipoulanidis, Alexander; Schönberger, Jörn (2021): Demand Forecasting. In Dmitry Ivanov, Alexander Tsipoulanidis, Jörn Schönberger (Eds.): Global Supply Chain and Operations Management. Cham: Springer International Publishing (Springer Texts in Business and Economics), pp. 341–357.

Izadyar, Nima; Ong, Hwai Chyuan; Shamshirband, Shahaboddin; Ghadamian, Hossein; Tong, Chong Wen (2015): Intelligent forecasting of residential heating demand for the District Heating System based on the monthly overall natural gas consumption. In *Energy and Buildings* 104, pp. 208–214. DOI: 10.1016/j.enbuild.2015.07.006.

Jahangir, Hamidreza; Tayarani, Hanif; Ahmadian, Ali; Golkar, Masoud Aliakbar; Miret, Jaume; Tayarani, Mohammad; Gao, H. Oliver (2019): Charging demand of Plug-in Electric Vehicles: Forecasting travel behavior based on a novel Rough Artificial Neural Network approach. In *Journal of Cleaner Production* 229, pp. 1029–1044. DOI: 10.1016/j.jclepro.2019.04.345.

Jaipuria, Sanjita; Mahapatra, S. S. (2014): An improved demand forecasting method to reduce bullwhip effect in supply chains. In *Expert Systems with Applications* 41 (5), pp. 2395–2408. DOI: 10.1016/j.eswa.2013.09.038.

Jiang, Shancheng; Chin, Kwai-Sang; Wang, Long; Qu, Gang; Tsui, Kwok L. (2017): Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. In *Expert Systems with Applications* 82, pp. 216–230. DOI: 10.1016/j.eswa.2017.04.017.

Jiang, Xiushan; Zhang, Lei; Chen, Xiqun (2014): Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China. In *Transportation Research Part C: Emerging Technologies* 44, pp. 110–127. DOI: 10.1016/j.trc.2014.03.016.

Johannesen, Nils Jakob; Kolhe, Mohan; Goodwin, Morten (2019): Relative evaluation of regression tools for urban area electrical energy demand forecasting. In *Journal of Cleaner Production* 218, pp. 555–564. DOI: 10.1016/j.jclepro.2019.01.108.

Jutten, Christian; Herault, Jeanny (1991): Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. In *Signal Processing* 24 (1), pp. 1–10. DOI: 10.1016/0165-1684(91)90079-X.

Kabir, Md Monirul; Islam, Md Monirul; Murase, Kazuyuki (2010): A new wrapper feature selection approach using neural network. In *Neurocomputing* 73 (16-18), pp. 3273–3283.

Ke, Jintao; Yang, Hai; Zheng, Hongyu; Chen, Xiqun; Jia, Yitian; Gong, Pinghua; Ye, Jieping (2019): Hexagon-Based Convolutional Neural Network for Supply-Demand Forecasting of Ride-Sourcing Services. In *IEEE Trans. Intell. Transport. Syst.* 20 (11), pp. 4160–4173. DOI: 10.1109/TITS.2018.2882861.

Ke, Jintao; Zheng, Hongyu; Yang, Hai; Chen, Xiqun (2017): Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. In *Transportation Research Part C: Emerging Technologies* 85, pp. 591–608. DOI: 10.1016/j.trc.2017.10.016.

Keitsch, Krischan A.; Bruckner, Thomas (2016): Modular electrical demand forecasting framework — A novel hybrid model approach. In : 13th International Multi-Conference on Systems, Signals & Devices. March 21-24, 2016 in Leipzig, Germany. 2016 13th International Multi-Conference on Systems, Signals & Devices (SSD). Leipzig, Germany, 3/21/2016 - 3/24/2016. Piscataway, NJ: IEEE, pp. 454–458.

Khan, Muhammad Adnan; Saqib, Shazia; Alyas, Tahir; Ur Rehman, Anees; Saeed, Yousaf; Zeb, Asim et al. (2020): Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. In *IEEE Access* 8, pp. 116013–116023. DOI: 10.1109/ACCESS.2020.3003790.

Khan, Zafar A.; Jayaweera, Dilan (2020): Smart Meter Data Based Load Forecasting and Demand Side Management in Distribution Networks With Embedded PV Systems. In *IEEE Access* 8, pp. 2631–2644. DOI: 10.1109/ACCESS.2019.2962150.

Kim, Thai Young; Dekker, Rommert; Heij, Christiaan (2017): Spare part demand forecasting for consumer goods using installed base information. In *Computers & Industrial Engineering* 103, pp. 201–215. DOI: 10.1016/j.cie.2016.11.014.

Kim, Wooju; Won, Jung Hoon; Park, Sangun; Kang, Juyoung (2015): Demand Forecasting Models for Medicines through Wireless Sensor Networks Data and Topic Trend Analysis. In *International Journal of Distributed Sensor Networks* 11 (9), p. 907169. DOI: 10.1155/2015/907169.

Kim, Yunsun; Kim, Sahm (2021): Forecasting Charging Demand of Electric Vehicles Using Time-Series Models. In *Energies* 14 (5), p. 1487. DOI: 10.3390/en14051487.

Kolassa, Stephan; Schütz, Wolfgang (2007): Advantages of the MAD/Mean Ratio over the MAPE. In *Foresight: The International Journal of Applied Forecasting* (6), pp. 40–43. Available online at https://ideas.repec.org/a/for/ijafaa/y2007i6p40-43.html.

Koller, Daphne; Sahami, Mehran (1996): Toward Optimal Feature Selection. Stanford InfoLab. Available online at http://ilpubs.stanford.edu:8090/208/.

Koning, Alex J.; Franses, Philip Hans; Hibon, Michèle; Stekler, H. O. (2005): The M3 competition: Statistical tests of the results. In *International Journal of Forecasting* 21 (3), pp. 397–409. DOI: 10.1016/j.ijforecast.2004.10.003.

Kostenko, Andrey V.; Hyndman, Rob J. (2008): Forecasting without significance tests? In *manuscript, Monash University, Australia*.

Krawiec, Krzysztof (2002): Genetic Programming-based Construction of Features for Machine Learning and Knowledge Discovery Tasks. In *Genet Program Evolvable Mach* 3 (4), pp. 329–343. DOI: 10.1023/A:1020984725014.

Kulshrestha, Anurag; Krishnaswamy, Venkataraghavan; Sharma, Mayank (2020): Bayesian BILSTM approach for tourism demand forecasting. In *Annals of Tourism Research* 83, p. 102925. DOI: 10.1016/j.annals.2020.102925.

Kurek, Teresa; Bielecki, Artur; Świrski, Konrad; Wojdan, Konrad; Guzek, Michał; Białek, Jakub et al. (2021): Heat demand forecasting algorithm for a Warsaw district heating network. In *Energy* 217, p. 119347. DOI: 10.1016/j.energy.2020.119347.

Lachenbruch, Peter A.; Mickey, M. Ray (1968): Estimation of Error Rates in Discriminant Analysis. In *Technometrics* 10 (1), p. 1. DOI: 10.2307/1266219.

Landry, Michel D.; Hack, Laurita M.; Coulson, Elizabeth; Freburger, Janet; Johnson, Michael P.; Katz, Richard et al. (2016): Workforce Projections 2010-2020: Annual Supply and Demand Forecasting Models for Physical Therapists Across the United States. In *Physical therapy* 96 (1), pp. 71–80. DOI: 10.2522/ptj.20150010.

Lapide, Larry (2009): History to demand-driven forecasting. In *Journal of Business Forecasting Methods and Systems* 28 (2), p. 18.

Law, Rob (2000): Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. In *Tourism Management* 21 (4), pp. 331–340. DOI: 10.1016/S0261-5177(99)00067-9.

Law, Rob; Li, Gang; Fong, Davis Ka Chio; Han, Xin (2019): Tourism demand forecasting: A deep learning approach. In *Annals of Tourism Research* 75, pp. 410–423. DOI: 10.1016/j.annals.2019.01.014.

Lecun, Yann; Bengio, Y. (1995): Convolutional Networks for Images, Speech, and Time-Series. In *The Handbook of Brain Theory and Neural Networks*.

Lee, John A.; Verleysen, Michel (2007): Nonlinear Dimensionality Reduction. New York, NY: Springer Science+Business Media LLC (Information Science and Statistics). Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10214400.

Legendre, Adrien Marie (1805): Nouvelles méthodes pour la détermination des orbites des comètes: F. Didot.

Leshno, Moshe; Lin, Vladimir Ya.; Pinkus, Allan; Schocken, Shimon (1993): Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. In *Neural Networks* 6 (6), pp. 861–867. DOI: 10.1016/S0893-6080(05)80131-5.

Levy, Yair; Ellis, Timothy J. (2006): A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. In *InformingSciJ* 9, pp. 181–212. DOI: 10.28945/479.

Li, Chongshou; Lim, Andrew (2018): A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. In *European Journal of Operational Research* 269 (3), pp. 860–869. DOI: 10.1016/j.ejor.2018.02.029.

Li, Hengyun; Hu, Mingming; Li, Gang (2020a): Forecasting tourism demand with multisource big data. In *Annals of Tourism Research* 83, p. 102912. DOI: 10.1016/j.annals.2020.102912.

Li, Jundong; Cheng, Kewei; Wang, Suhang; Morstatter, Fred; Trevino, Robert P.; Tang, Jiliang; Liu, Huan (2017): Feature Selection. In *ACM Comput. Surv.* 50 (6), pp. 1–45. DOI: 10.1145/3136625.

Li, Ranran; Jiang, Ping; Yang, Hufang; Li, Chen (2020b): A novel hybrid forecasting scheme for electricity demand time series. In *Sustainable Cities and Society* 55, p. 102036. DOI: 10.1016/j.scs.2020.102036.

Li, Xin; Law, Rob (2020): Forecasting Tourism Demand with Decomposed Search Cycles. In *Journal of Travel Research* 59 (1), pp. 52–68. DOI: 10.1177/0047287518824158.

Li, Yifeng; Chen, Chih-Yu; Wasserman, Wyeth W. (2015): Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. In Teresa M. Przytycka (Ed.): Research in Computational Molecular Biology. 19th Annual International Conference, RECOMB 2015, Warsaw, Poland, April 12-15, 2015, Proceedings, vol. 9029. Cham: Springer International Publishing; Imprint; Springer (Lecture Notes in Computer Science, 9029), pp. 205–217.

Lin, J.; Keogh, E.; Fu, A.; van Herle, H. (2005): Approximations to Magic: Finding Unusual Medical Time Series. In Alexey Tsymbal, Pádraig Cunningham (Eds.): Proceedings, 18th IEEE Symposium on Computer-Based Medical Systems. 23-24 June 2005, Dublin, Ireland. 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05). Dublin, Ireland, 23-24 June 2005. IEEE Symposium on Computer-Based Medical Systems; IEEE Computer Society; Science Foundation Ireland; Trinity College (Dublin, Ireland). Los Alamitos, Calif: IEEE Computer Society, pp. 329–334.

Lipo P. Wang; Chunru R. Wan: Comments and Replies Comments on “The Extreme Learning Machine”.

Little, W. A. (1996): The Existence of Persistent States in the Brain. In Blas Cabrera, H. Gutfreund, Vladimir Kresin (Eds.): From High-Temperature Superconductivity to Microminiature Refrigeration. N. Boston: Springer US, pp. 145–164.

Liu, Bo; Zhang, Zhenguo; Cui, Rongyi (2020a): Efficient Time Series Augmentation Methods. In : 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Chengdu, China, 17.10.2020 - 19.10.2020: IEEE, pp. 1004–1009.

Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (2008): Isolation Forest. In Fosca Giannotti (Ed.): Eighth IEEE International Conference on Data Mining, 2008. ICDM '08 ; Pisa, Italy, 15 - 19 Dec. 2008. 2008 Eighth IEEE International Conference on Data Mining (ICDM). Pisa, Italy, 12/15/2008 - 12/19/2008. Institute of Electrical and Electronics Engineers; IEEE International Conference on Data Mining; ICDM. Piscataway, NJ: IEEE, pp. 413–422.

Liu, Xiaolei; Lin, Zi (2021): Impact of Covid-19 pandemic on electricity demand in the UK based on multivariate time series forecasting with Bidirectional Long Short Term Memory. In *Energy* 227, p. 120455. DOI: 10.1016/j.energy.2021.120455.

Liu, Zhizhen; Chen, Hong; Li, Yan; Zhang, Qi (2020b): Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots. In *Journal of Advanced Transportation* 2020, pp. 1–13. DOI: 10.1155/2020/1302586.

Maeng, Kyuho; Kim, Jihwan; Shin, Jungwoo (2020): Demand forecasting for the 5G service market considering consumer preference and purchase delay behavior. In *Telematics and Informatics* 47, p. 101327. DOI: 10.1016/j.tele.2019.101327.

Maier, Tobias; Afentakis, Anja (2013): Forecasting supply and demand in nursing professions: impacts of occupational flexibility and employment structure in Germany. In *Human resources for health* 11, p. 24. DOI: 10.1186/1478-4491-11-24.

Makridakis, S.; Andersen, A.; Carbone, R.; Fildes, R.; Hibon, M.; Lewandowski, R. et al. (1982): The accuracy of extrapolation (time series) methods: Results of a forecasting competition. In *J. Forecast.* 1 (2), pp. 111–153. DOI: 10.1002/for.3980010202.

Makridakis, Spyros (1993): Accuracy measures: theoretical and practical concerns. In *International Journal of Forecasting* 9 (4), pp. 527–529. DOI: 10.1016/0169-2070(93)90079-3.

Makridakis, Spyros; Chatfield, Chris; Hibon, Michèle; Lawrence, Michael; Mills, Terence; Ord, Keith; Simmons, LeRoy F. (1993): The M2-competition: A real-time judgmentally based forecasting study. In *International Journal of Forecasting* 9 (1), pp. 5–22. DOI: 10.1016/0169-2070(93)90044-N.

Makridakis, Spyros; Hibon, Michèle (2000): The M3-Competition: results, conclusions and implications. In *International Journal of Forecasting* 16 (4), pp. 451–476. DOI: 10.1016/S0169-2070(00)00057-1.

Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios (2020): The M4 Competition: 100,000 time series and 61 forecasting methods. In *International Journal of Forecasting* 36 (1), pp. 54–74. DOI: 10.1016/j.ijforecast.2019.04.014.

Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios (2022): M5 accuracy competition: Results, findings, and conclusions. In *International Journal of Forecasting* 38 (4), pp. 1346–1364. DOI: 10.1016/j.ijforecast.2021.11.013.

Makridakis, Spyros; Winkler, Robert L. (1989): Sampling Distributions of Post-Sample Forecasting Errors. In *Applied Statistics* 38 (2), p. 331. DOI: 10.2307/2348063.

Mansoor, Muhammad; Grimaccia, Francesco; Leva, Sonia; Mussetta, Marco (2021): Comparison of echo state network and feed-forward neural networks in electrical load forecasting for demand response programs. In *Mathematics and Computers in Simulation* 184, pp. 282–293. DOI: 10.1016/j.matcom.2020.07.011.

Markovitch, Shaul; Rosenstein, Dan (2002): Feature Generation Using General Constructor Functions. In *Machine Learning* 49 (1), pp. 59–98. DOI: 10.1023/A:1014046307775.

Marquardt, Donald W.; Snee, Ronald D. (1975): Ridge Regression in Practice. In *The American Statistician* 29 (1), pp. 3–20. DOI: 10.1080/00031305.1975.10479105.

Masci, Jonathan; Meier, Ueli; Cireşan, Dan; Schmidhuber, Jürgen (2011): Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In Timo Honkela (Ed.): Artificial neural networks and machine learning-- ICANN 2011. 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, vol. 6791. Berlin, Heidelberg, New York: Springer (LNCS sublibrary. SL 1, Theoretical computer science and general issues, 6792), pp. 52–59.

Masdari, Mohammad; Khoshnevis, Afsane (2020): A survey and classification of the workload forecasting methods in cloud computing. In *Cluster Comput* 23 (4), pp. 2399–2424. DOI: 10.1007/s10586-019-03010-3.

McQuarrie, Allan D. R.; Tsai, Chih-Ling (1998): Regression and Time Series Model Selection: WORLD SCIENTIFIC.

Meade, Nigel; Islam, Towhidul (2015): Forecasting in telecommunications and ICT—A review. In *International Journal of Forecasting* 31 (4), pp. 1105–1126. DOI: 10.1016/j.ijforecast.2014.09.003.

Mediavilla, Mario Angos; Dietrich, Fabian; Palm, Daniel (2022): Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. In *Procedia CIRP* 107, pp. 1126–1131. DOI: 10.1016/j.procir.2022.05.119.

Mitchell, Tom Michael (1997): Machine learning. International ed. New York, NY: McGraw-Hill (McGraw-Hill series in computer science).

Molino, Piero; Dudin, Yaroslav; Miryala, Sai Sumanth (2019): Ludwig: a type-based declarative deep learning toolbox. Available online at http://arxiv.org/pdf/1909.07930v1.

Montgomery, Douglas C.; Jennings, Cheryl L.; Kulahci, Murat (2008): Introduction to time series analysis and forecasting. Hoboken, NJ: Wiley-Interscience (Wiley series in probability and statistics). Available online at http://www.loc.gov/catdir/enhancements/fy0740/2007019891-d.html.

Mostard, Julien; Teunter, Ruud; Koster, René de (2011): Forecasting demand for single-period products: A case study in the apparel industry. In *European Journal of Operational Research* 211 (1), pp. 139–147. DOI: 10.1016/j.ejor.2010.11.001.

Motoda, Hiroshi; Liu, Huan (2002): Feature selection, extraction and construction. In *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol* 5 (67-72), p. 2.

Munos, Rémi (2011): Optimistic optimization of a deterministic function without the knowledge of its smoothness. In : Advances in neural information processing systems, pp. 783–791.

Murray, Christopher J. L. (2020): Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months.

Narendra; Fukunaga (1977): A Branch and Bound Algorithm for Feature Subset Selection. In *IEEE Trans. Comput.* C-26 (9), pp. 917–922. DOI: 10.1109/TC.1977.1674939.

Nasr, George E.; Badr, E. A.; Joun, C. (2002): Cross entropy error function in neural networks: Forecasting gasoline demand. In : FLAIRS conference, pp. 381–384.

Nenni, Maria Elena; Giustiniano, Luca; Pirolo, Luca (2013): Demand Forecasting in the Fashion Industry: A Review. In *International Journal of Engineering Business Management* 5 (1), p. 37. DOI: 10.5772/56840.

Nikolopoulos, K.; Syntetos, A. A.; Boylan, J. E.; Petropoulos, F.; Assimakopoulos, V. (2011): An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. In *Journal of the Operational Research Society* 62 (3), pp. 544–554. DOI: 10.1057/jors.2010.32.

Ofori-Ntow Jnr, Eric; Ziggah, Yao Yevenyo; Relvas, Susana (2021): Hybrid ensemble intelligent model based on wavelet transform, swarm intelligence and artificial neural network for electricity demand forecasting. In *Sustainable Cities and Society* 66, p. 102679. DOI: 10.1016/j.scs.2020.102679.

Oh, Chi-Ok; Morzuch, Bernard J. (2005): Evaluating Time-Series Models to Forecast the Demand for Tourism in Singapore. In *Journal of Travel Research* 43 (4), pp. 404–413. DOI: 10.1177/0047287505274653.

Pagallo, Giulia (1989): Learning DNF by decision trees. In. International Joint Conference on Artificial Intelligence, 8/20/1989, pp. 639–644.

Panapakidis, Ioannis P.; Dagoumas, Athanasios S. (2017): Day-ahead natural gas demand forecasting based on the combination of wavelet transform and ANFIS/genetic algorithm/neural network model. In *Energy* 118, pp. 231–245. DOI: 10.1016/j.energy.2016.12.033.

Pao, Yoh-Han; Park, Gwang-Hoon; Sobajic, Dejan J. (1994): Learning and generalization characteristics of the random vector functional-link net. In *Neurocomputing* 6 (2), pp. 163–180. DOI: 10.1016/0925-2312(94)90053-1.

Pazzani, Michael J. (1998): Constructive Induction of Cartesian Product Attributes. In Huan Liu, Hiroshi Motoda (Eds.): Feature Extraction, Construction and Selection, vol. 10. Boston, MA: Springer US, pp. 341–354.

Pearson, Karl (1901): LIII. On lines and planes of closest fit to systems of points in space. In *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11), pp. 559–572. DOI: 10.1080/14786440109462720.

Pérez-Chacón, R.; Asencio-Cortés, G.; Martínez-Álvarez, F.; Troncoso, A. (2020): Big data time series forecasting based on pattern sequence similarity and its application to the electricity demand. In *Information Sciences* 540, pp. 160–174. DOI: 10.1016/j.ins.2020.06.014.

Petropoulos, Fotios; Makridakis, Spyros; Assimakopoulos, Vassilios; Nikolopoulos, Konstantinos (2014): ‘Horses for Courses’ in demand forecasting. In *European Journal of Operational Research* 237 (1), pp. 152–163. DOI: 10.1016/j.ejor.2014.02.036.

Pinçe, Çerağ; Turrini, Laura; Meissner, Joern (2021): Intermittent demand forecasting for spare parts: A Critical review. In *Omega* 105, p. 102513. DOI: 10.1016/j.omega.2021.102513.

Piramuthu, Selwyn (2004): Evaluating feature selection methods for learning in data mining applications. In *European Journal of Operational Research* 156 (2), pp. 483–494. DOI: 10.1016/S0377-2217(02)00911-6.

Powers, Rob; Goldszmidt, Moises; Cohen, Ira (2005): Short term performance forecasting in enterprise systems. In Robert L. Grossman (Ed.): KDD-2005. Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery and data mining / Edited by: Robert L. Grossman … [et al.]. Proceeding of the eleventh ACM SIGKDD international conference. Chicago, Illinois, USA, 8/21/2005 - 8/24/2005. ACM Special Interest Group on Knowledge Discovery and Data Mining. New York, New York, USA: ACM Press, p. 801.

Prescott, Raymond B. (1922): Law of Growth in Forecasting Demand. In *Journal of the American Statistical Association* 18 (140), pp. 471–479. DOI: 10.1080/01621459.1922.10502490.

Qiu, Xueheng; Ren, Ye; Suganthan, Ponnuthurai Nagaratnam; Amaratunga, Gehan A.J. (2017): Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting. In *Applied Soft Computing* 54, pp. 246–255. DOI: 10.1016/j.asoc.2017.01.015.

Qiu, Xueheng; Zhang, Le; Ren, Ye; Suganthan, P.; Amaratunga, Gehan (2014): Ensemble deep learning for regression and time series forecasting. In : 2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL). 9-12 Dec. 2014, Orlando, Florida, USA ; [part of] IEEE SSCI 2014, 2014 IEEE Symposium Series on Computational Intelligence. 2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL). Orlando, FL, USA, 12/9/2014 - 12/12/2014. IEEE Symposium on Computational Intelligence in Ensemble Learning; CIEL; IEEE Symposium Series on Computational Intelligence; SSCI. Piscataway, NJ: IEEE, pp. 1–6.

Racine, Jeff (2000): Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. In *Journal of Econometrics* 99 (1), pp. 39–61. DOI: 10.1016/S0304-4076(00)00030-0.

Rashid, Khandakar M.; Louis, Joseph (2019): Times-series data augmentation and deep learning for construction equipment activity recognition. In *Advanced Engineering Informatics* 42, p. 100944. DOI: 10.1016/j.aei.2019.100944.

Raza, Muhammad Amir; Khatri, Krishan Lal; Israr, Amber; Ul Haque, Muhammad Ibrar; Ahmed, Manzar; Rafique, Khalid; Saand, Abdul Sattar (2022): Energy demand and production forecasting in Pakistan. In *Energy Strategy Reviews* 39, p. 100788. DOI: 10.1016/j.esr.2021.100788.

Raza, Muhammad Qamar; Khosravi, Abbas (2015): A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. In *Renewable and Sustainable Energy Reviews* 50, pp. 1352–1372. DOI: 10.1016/j.rser.2015.04.065.

Ren, Ye; Suganthan, P. N.; Srikanth, N.; Amaratunga, Gehan (2016): Random vector functional link network for short-term electricity load demand forecasting. In *Information Sciences* 367-368, pp. 1078–1093. DOI: 10.1016/j.ins.2015.11.039.

Rezaali, Mostafa; Quilty, John; Karimi, Abdolreza (2021): Probabilistic urban water demand forecasting using wavelet-based machine learning models. In *Journal of Hydrology* 600, p. 126358. DOI: 10.1016/j.jhydrol.2021.126358.

Rockaway, Thomas D.; Coomes, Paul A.; Rivard, Joshua; Kornstein, Barry (2011): Residential water use trends in North America. In *Journal - American Water Works Association* 103 (2), pp. 76–89. DOI: 10.1002/j.1551-8833.2011.tb11403.x.

Romano, Michele; Kapelan, Zoran (2014): Adaptive water demand forecasting for near real-time management of smart water distribution systems. In *Environmental Modelling & Software* 60, pp. 265–276. DOI: 10.1016/j.envsoft.2014.06.016.

Rostami-Tabar, Bahman; Babai, M. Zied; Syntetos, Aris; Ducq, Yves (2013): Demand forecasting by temporal aggregation. In *Naval Research Logistics* 60 (6), pp. 479–498. DOI: 10.1002/nav.21546.

Roweis, S. T.; Saul, L. K. (2000): Nonlinear dimensionality reduction by locally linear embedding. In *Science (New York, N.Y.)* 290 (5500), pp. 2323–2326. DOI: 10.1126/science.290.5500.2323.

Ryu, Seunghyoung; Noh, Jaekoo; Kim, Hongseok (2017): Deep Neural Network Based Demand Side Short Term Load Forecasting. In *Energies* 10 (1), p. 3. DOI: 10.3390/en10010003.

Saeys, Yvan; Inza, Iñaki; Larrañaga, Pedro (2007): A review of feature selection techniques in bioinformatics. In *Bioinformatics (Oxford, England)* 23 (19), pp. 2507–2517. DOI: 10.1093/bioinformatics/btm344.

Salinas, David; Flunkert, Valentin; Gasthaus, Jan; Januschowski, Tim (2019): DeepAR: Probabilistic forecasting with autoregressive recurrent networks. In *International Journal of Forecasting*. DOI: 10.1016/j.ijforecast.2019.07.001.

Schmidhuber, Jürgen (2015): Deep learning in neural networks: an overview. In *Neural networks : the official journal of the International Neural Network Society* 61, pp. 85–117. DOI: 10.1016/j.neunet.2014.09.003.

Schmidt, Wouter F.; Kraaijveld, Martin A.; Duin, Robert P. W.; others (1992): Feed forward neural networks with random weights. In : International Conference on Pattern Recognition. IEEE COMPUTER SOCIETY PRESS, p. 1.

Setiono, Rudy; Liu, Huan (1998): Feature extraction via Neural networks. In Huan Liu, Hiroshi Motoda (Eds.): Feature Extraction, Construction and Selection. Boston, MA: Springer US, pp. 191–204.

Seyedan, Mahya; Mafakheri, Fereshteh (2020): Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. In *J Big Data* 7 (1). DOI: 10.1186/s40537-020-00329-2.

Shahriari, Bobak; Swersky, Kevin; Wang, Ziyu; Adams, Ryan P.; Freitas, Nando de (2016): Taking the Human Out of the Loop: A Review of Bayesian Optimization. In *Proc. IEEE* 104 (1), pp. 148–175. DOI: 10.1109/JPROC.2015.2494218.

Shakarami, Ali; Shahidinejad, Ali; Ghobaei-Arani, Mostafa (2021): An autonomous computation offloading strategy in Mobile Edge Computing: A deep learning-based hybrid approach. In *Journal of Network and Computer Applications* 178, p. 102974. DOI: 10.1016/j.jnca.2021.102974.

Shang, Han Lin (2013): Functional time series approach for forecasting very short-term electricity demand. In *Journal of Applied Statistics* 40 (1), pp. 152–168. DOI: 10.1080/02664763.2012.740619.

Shanker, M.; Hu, M. Y.; Hung, M. S. (1996): Effect of data standardization on neural network training. In *Omega* 24 (4), pp. 385–397. DOI: 10.1016/0305-0483(96)00010-2.

Shao, Jun (1993): Linear Model Selection by Cross-validation. In *Journal of the American Statistical Association* 88 (422), pp. 486–494. DOI: 10.1080/01621459.1993.10476299.

Sheikhan, Mansour; Mohammadi, Najmeh (2013): Time series prediction using PSO-optimized neural network and hybrid feature selection algorithm for IEEE load data. In *Neural Comput & Applic* 23 (3-4), pp. 1185–1194. DOI: 10.1007/s00521-012-0980-8.

Sison, Nicolai; Li, Lin; Han, Meng (2021): Survey of Machine Learning and Deep Learning Techniques for Travel Demand Forecasting. In : 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI). 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI). Atlanta, GA, USA, 18.10.2021 - 21.10.2021: IEEE, pp. 606–613.

Snoek, Jasper; Larochelle, Hugo; Adams, Ryan P. (2012): Practical bayesian optimization of machine learning algorithms. In : Advances in neural information processing systems, pp. 2951–2959.

Son, Hyojoo; Kim, Changwan (2020): A Deep Learning Approach to Forecasting Monthly Demand for Residential–Sector Electricity. In *Sustainability* 12 (8), p. 3103. DOI: 10.3390/su12083103.

Sondhi, Parikshit (2009): Feature construction methods: a survey. In *sifaka. cs. uiuc. edu* 69, pp. 70–71.

Song, Haiyan; Li, Gang (2008): Tourism demand modelling and forecasting—A review of recent research. In *Tourism Management* 29 (2), pp. 203–220. DOI: 10.1016/j.tourman.2007.07.016.

Song, Haiyan; Li, Gang; Witt, Stephen F.; Fei, Baogang (2010): Tourism Demand Modelling and Forecasting: How Should Demand Be Measured? In *Tourism Economics* 16 (1), pp. 63–81. DOI: 10.5367/000000010790872213.

Song, Haiyan; Qiu, Richard T.R.; Park, Jinah (2019): A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. In *Annals of Tourism Research* 75, pp. 338–362. DOI: 10.1016/j.annals.2018.12.001.

Song, Haiyan; Wong, Kevin K.F.; Chon, Kaye K.S. (2003): Modelling and forecasting the demand for Hong Kong tourism. In *International Journal of Hospitality Management* 22 (4), pp. 435–451. DOI: 10.1016/S0278-4319(03)00047-1.

Suganthi, L.; Samuel, Anand A. (2012): Energy models for demand forecasting—A review. In *Renewable and Sustainable Energy Reviews* 16 (2), pp. 1223–1240. DOI: 10.1016/j.rser.2011.08.014.

Sun, Zexian; Zhao, Mingyu (2020): Short-Term Wind Power Forecasting Based on VMD Decomposition, ConvLSTM Networks and Error Analysis. In *IEEE Access* 8, pp. 134422–134434. DOI: 10.1109/ACCESS.2020.3011060.

Sutton, Richard S.; Matheus, Christopher J. (1991): Learning Polynomial Functions by Feature Construction. In : Machine Learning Proceedings 1991: Elsevier, pp. 208–212.

Taieb, Souhaib Ben; Taylor, James W.; Hyndman, Rob J. (2021): Hierarchical Probabilistic Forecasting of Electricity Demand With Smart Meter Data. In *Journal of the American Statistical Association* 116 (533), pp. 27–43. DOI: 10.1080/01621459.2020.1736081.

Tan, Mao; Yuan, Siping; Li, Shuaihu; Su, Yongxin; Li, Hui; He, Feng He (2020): Ultra-Short-Term Industrial Power Demand Forecasting Using LSTM Based Hybrid Ensemble Learning. In *IEEE Trans. Power Syst.* 35 (4), pp. 2937–2948. DOI: 10.1109/TPWRS.2019.2963109.

Tashman, Leonard J. (2000): Out-of-sample tests of forecasting accuracy: an analysis and review. In *International Journal of Forecasting* 16 (4), pp. 437–450. DOI: 10.1016/S0169-2070(00)00065-0.

Tavenard, Romain; Faouzi, Johann; Vandewiele, Gilles; Divo, Felix; Androz, Guillaume; Holtz, Chester et al. (2020): Tslearn, a machine learning toolkit for time series data. In *J. Mach. Learn. Res.* 21 (118), pp. 1–6.

Taylor, James W. (2010): Triple seasonal methods for short-term electricity demand forecasting. In *European Journal of Operational Research* 204 (1), pp. 139–152. DOI: 10.1016/j.ejor.2009.10.003.

Teunter, R. H.; Duncan, L. (2009): Forecasting intermittent demand: a comparative study. In *Journal of the Operational Research Society* 60 (3), pp. 321–329. DOI: 10.1057/palgrave.jors.2602569.

Teunter, Ruud H.; Syntetos, Aris A.; Zied Babai, M. (2011): Intermittent demand: Linking forecasting to inventory obsolescence. In *European Journal of Operational Research* 214 (3), pp. 606–615. DOI: 10.1016/j.ejor.2011.05.018.

Tibshirani, Robert (1996): Regression Shrinkage and Selection Via the Lasso. In *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Tiwari, Mukesh K.; Adamowski, Jan (2013): Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. In *Water Resour. Res.* 49 (10), pp. 6486–6507. DOI: 10.1002/wrcr.20517.

Toktay, L. Beril; Wein, Lawrence M. (2001): Analysis of a Forecasting-Production-Inventory System with Stationary Demand. In *Management Science* 47 (9), pp. 1268–1281. DOI: 10.1287/mnsc.47.9.1268.9787.

Torres, José F.; Hadjout, Dalil; Sebaa, Abderrazak; Martínez-Álvarez, Francisco; Troncoso, Alicia (2021): Deep Learning for Time Series Forecasting: A Survey. In *Big data* 9 (1), pp. 3–21. DOI: 10.1089/big.2020.0159.

Truong, Anh; Walters, Austin; Goodsitt, Jeremy; Hines, Keegan; Bruss, C. Bayan; Farivar, Reza (2019): Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. Available online at http://arxiv.org/pdf/1908.05557v2.

Tsao, Yu-Chung; Chen, Yu-Kai; Chiu, Shih-Hao; Lu, Jye-Chyi; Vu, Thuy-Linh (2022): An innovative demand forecasting approach for the server industry. In *Technovation* 110, p. 102371. DOI: 10.1016/j.technovation.2021.102371.

Tsay, R. S. (2000): Outliers in multivariate time series. In *Biometrika* 87 (4), pp. 789–804. DOI: 10.1093/biomet/87.4.789.

Turing, Alan (1948): Intelligent machinery (1948). In *The Essential Turing*, p. 395.

Valko, Michal; Carpentier, Alexandra; Munos, Rémi (2013): Stochastic simultaneous optimistic optimization. In : International Conference on Machine Learning, pp. 19–27.

van der Maaten, Laurens; Hinton, Geoffrey (2008): Visualizing data using t-SNE. In *Journal of Machine Learning Research* 9 (Nov), pp. 2579–2605.

van der Meer, D. W.; Shepero, M.; Svensson, A.; Widén, J.; Munkhammar, J. (2018): Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian Processes. In *Applied Energy* 213, pp. 195–207. DOI: 10.1016/j.apenergy.2017.12.104.

Velasquez, Carlos E.; Zocatelli, Matheus; Estanislau, Fidellis B.G.L.; Castro, Victor F. (2022): Analysis of time series models for Brazilian electricity demand forecasting. In *Energy* 247, p. 123483. DOI: 10.1016/j.energy.2022.123483.

Venkatesh, Kamini; Ravi, Vadlamani; Prinzie, Anita; van den Poel, Dirk (2014): Cash demand forecasting in ATMs by clustering and neural networks. In *European Journal of Operational Research* 232 (2), pp. 383–392. DOI: 10.1016/j.ejor.2013.07.027.

Vilar, Juan M.; Raña, Paula; Aneiros, Germán (2016): Using robust FPCA to identify outliers in functional time series, with applications to the electricity market. In *Sort-statistics and Operations Research Transactions* 1 (2), pp. 321–348.

Volchek, Katerina; Liu, Anyu; Song, Haiyan; Buhalis, Dimitrios (2019): Forecasting tourist arrivals at attractions: Search engine empowered methodologies. In *Tourism Economics* 25 (3), pp. 425–447. DOI: 10.1177/1354816618811558.

Wang, Fei; Xiang, Biao; Li, Kangping; Ge, Xinxin; Lu, Hai; Lai, Jingang; Dehghanian, Payman (2020): Smart Households' Aggregated Capacity Forecasting for Load Aggregators Under Incentive-Based Demand Response Programs. In *IEEE Trans. on Ind. Applicat.* 56 (2), pp. 1086–1097. DOI: 10.1109/TIA.2020.2966426.

Wang, Qiang; Li, Shuyu; Li, Rongrong (2018): Forecasting energy demand in China and India: Using single-linear, hybrid-linear, and non-linear time series forecast techniques. In *Energy* 161, pp. 821–831. DOI: 10.1016/j.energy.2018.07.168.

Webster, Jane; Watson, Richard T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In *MIS quarterly* 26 (2), pp. xiii–xxiii. Available online at http://www.jstor.org/stable/4132319.

Wen, Long; Liu, Chang; Song, Haiyan; Liu, Han (2021a): Forecasting Tourism Demand with an Improved Mixed Data Sampling Model. In *Journal of Travel Research* 60 (2), pp. 336–353. DOI: 10.1177/0047287520906220.

Wen, Qingsong; Sun, Liang; Yang, Fan; Song, Xiaomin; Gao, Jingkun; Wang, Xue; Xu, Huan (2021b): Time Series Data Augmentation for Deep Learning: A Survey, pp. 4653–4660. DOI: 10.24963/ijcai.2021/631.

Willemain, Thomas R.; Smart, Charles N.; Schwarz, Henry F. (2004): A new approach to forecasting intermittent demand for service parts inventories. In *International Journal of Forecasting* 20 (3), pp. 375–387. DOI: 10.1016/S0169-2070(03)00013-X.

Williams, Brent D.; Waller, Matthew A. (2010): CREATING ORDER FORECASTS: POINT-OF-SALE OR ORDER HISTORY? In *Journal of Business Logistics* 31 (2), pp. 231–251. DOI: 10.1002/j.2158-1592.2010.tb00150.x.

Williams, Sean; Short, Michael (2020): Electricity demand forecasting for decentralised energy management. In *Energy and Built Environment* 1 (2), pp. 178–186. DOI: 10.1016/j.enbenv.2020.01.001.

Winters, Peter R. (1960): Forecasting Sales by Exponentially Weighted Moving Averages. In *Management Science* 6 (3), pp. 324–342. DOI: 10.1287/mnsc.6.3.324.

Wu, Yi-Leh; Agrawal, Divyakant; El Abbadi, Amr (2000): A comparison of DFT and DWT based similarity search in time-series databases. In Arvin Agah (Ed.): Proceedings of the ninth international conference on Information and knowledge management. the ninth international conference. McLean, Virginia, United States, 11/6/2000 - 11/11/2000. New York, NY: ACM, pp. 488–495.

Xenochristou, Maria; Kapelan, Zoran (2020): An ensemble stacked model with bias correction for improved water demand forecasting. In *Urban Water Journal* 17 (3), pp. 212–223. DOI: 10.1080/1573062X.2020.1758164.

Xie, Gang; Qian, Yatong; Wang, Shouyang (2021): Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. In *Tourism Management* 82, p. 104208. DOI: 10.1016/j.tourman.2020.104208.

Xie, Jingrui; Chen, Ying; Hong, Tao; Laing, Thomas D. (2018): Relative Humidity for Load Forecasting Models. In *IEEE Trans. Smart Grid* 9 (1), pp. 191–198. DOI: 10.1109/TSG.2016.2547964.

Xu, Chengcheng; Ji, Junyi; Liu, Pan (2018): The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. In *Transportation Research Part C: Emerging Technologies* 95, pp. 47–60. DOI: 10.1016/j.trc.2018.07.013.

Yan, Weizhong (2012): Toward automatic time-series forecasting using neural networks. In *IEEE transactions on neural networks and learning systems* 23 (7), pp. 1028–1039. DOI: 10.1109/TNNLS.2012.2198074.

Yang, Jihoon; Honavar, Vasant (1998): Feature Subset Selection Using a Genetic Algorithm. In Huan Liu, Hiroshi Motoda (Eds.): Feature Extraction, Construction and Selection. Boston, MA: Springer US, pp. 117–136.

Yao, Huaxiu; Wu, Fei; Ke, Jintao; Tang, Xianfeng; Jia, Yitian; Lu, Siyu et al. (2018a): Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction. In *AAAI* 32 (1). DOI: 10.1609/aaai.v32i1.11836.

Yao, Quanming; Wang, Mengshuo; Chen, Yuqiang; Dai, Wenyuan; Li, Yu-Feng; Tu, Wei-Wei et al. (2018b): Taking Human out of Learning Applications: A Survey on Automated Machine Learning. Available online at http://arxiv.org/pdf/1810.13306v4.

Yao, Yuan; Cao, Yi (2020): A Neural network enhanced hidden Markov model for tourism demand forecasting. In *Applied Soft Computing* 94, p. 106465. DOI: 10.1016/j.asoc.2020.106465.

Yue, Liu; Zhenjiang, Liao; Yafeng, Yin; Zaixia, Teng; Junjun, Gao; Bofeng, Zhang (2010): Selective and Heterogeneous SVM Ensemble for Demand Forecasting. In : Proceedings the 10th IEEE International Conference on Computer and Information Technology. CIT 2010 : the 7th IEEE International Conference on Embedded Software and Systems (ICESS-2010) : the 10th IEEE International Conference on Scalable Computing and Communications (ScalCom-2010) : 29 June-1 July 2010, Bradford, West Yorkshire, UK. 2010 IEEE 10th International Conference on Computer and Information Technology (CIT). Bradford, United Kingdom, 6/29/2010 - 7/1/2010. IEEE International Conference on Embedded Software and Systems; IEEE International Conference on Scalable Computing and Communications. Los Alamitos, Calif.: IEEE Computer Society, pp. 1519–1524.

Yukseltan, Ergun; Yucekaya, Ahmet; Bilge, Ayse Humeyra (2020): Hourly electricity demand forecasting using Fourier analysis with feedback. In *Energy Strategy Reviews* 31, p. 100524. DOI: 10.1016/j.esr.2020.100524.

Zhang, Guoqiang; Eddy Patuwo, B.; Y. Hu, Michael (1998): Forecasting with artificial neural networks. In *International Journal of Forecasting* 14 (1), pp. 35–62. DOI: 10.1016/S0169-2070(97)00044-7.

Zhang, Q.; Benveniste, A. (1992): Wavelet networks. In *IEEE transactions on neural networks* 3 (6), pp. 889–898. DOI: 10.1109/72.165591.

Zhang, Yishuo; Li, Gang; Muskat, Birgit; Law, Rob (2021): Tourism Demand Forecasting: A Decomposed Deep Learning Approach. In *Journal of Travel Research* 60 (5), pp. 981–997. DOI: 10.1177/0047287520919522.

Zhang, Yishuo; Li, Gang; Muskat, Birgit; Law, Rob; Yang, Yating (2020): Group pooling for deep tourism demand forecasting. In *Annals of Tourism Research* 82, p. 102899. DOI: 10.1016/j.annals.2020.102899.

Zheng, Huiting; Yuan, Jiabin; Chen, Long (2017): Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. In *Energies* 10 (8), p. 1168. DOI: 10.3390/en10081168.

Zhu, Xiaodan; Ninh, Anh; Zhao, Hui; Liu, Zhenming (2021): Demand Forecasting with Supply-Chain Information and Machine Learning: Evidence in the Pharmaceutical Industry. In *Prod Oper Manag* 30 (9), pp. 3231–3252. DOI: 10.1111/poms.13426.

Zhu, Xiaoxin; Zhang, Guanghai; Sun, Baiqing (2019): A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence. In *Nat Hazards* 97 (1), pp. 65–82. DOI: 10.1007/s11069-019-03626-z.

Zied Babai, Mohamed; Syntetos, Aris; Teunter, Ruud (2014): Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence. In *International Journal of Production Economics* 157, pp. 212–219. DOI: 10.1016/j.ijpe.2014.08.019.

Zubaidi, Salah L.; Al-Bugharbee, Hussein; Muhsin, Yousif Raad; Hashim, Khalid; Alkhaddar, Rafid (2020): Forecasting of monthly stochastic signal of urban water demand: Baghdad as a case study. In *IOP Conf. Ser.: Mater. Sci. Eng.* 888 (1), p. 12018. DOI: 10.1088/1757-899X/888/1/012018.

4 Developing an Understanding of External Factors Influencing Demand Forecasting Models using a Case Example

[*This chapter consists of a study published in 2021*: Bauer, Markus; Kiefer, Daniel; Grimm, Florian (2021): Sales Forecasting Under Economic Crisis: A Case Study of the Impact of the COVID19 Crisis to the Predictability of Sales of a Medium-Sized Enterprise. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 163–172. – *further referred to as* Bauer et al. 2021]

Sales Forecasting under Economic Crisis: A Case study of the Impact of the COVID19 Crisis to the predictability of Sales of a Medium-Sized Enterprise

This article describes a case study of sales forecasting with machine learning. Based on the example of a medium-sized enterprise in the field of B2B retailing, the study examines how the effect of global events on the enterprise's sales can be modelled. We conclude that global events show substantial effects on the enterprise's revenues, that economic indicators are suitable to improve sales forecasts and how they should be applied.

4.1 Introduction

The task of forecasting sales is a central element of strategic planning in enterprises. Accurate sales forecasts predict future sales potentials to a high degree of certainty for a future period of weeks and months. These forecasts enable all departments to efficiently allocate resources and capacities to optimally fulfil their customers' demands. Therefore, enterprises put considerable effort into improving their sales forecasting methods as a competitive advantage over other market participants.

In the light of the recent global COVID19 crisis, it becomes evident that rare and exceptional events can have severe impact on enterprises – irrespective of the fact whether they do business internationally or not. Estimations of the WorldBank predict a loss of global Gross Domestic Product (GDP) of 5.2% for the year 2020 and long term effects for all nations (Global Economic Prospects, June 2020 2020).

In this article, we describe the case study of an enterprise and improvements of their sales forecast. While the general target of our research began with the goal to predict sales three month ahead, we focus on the impact of global events on sales forecasting in this study. This article is a descriptive case study on the particular impact of the COVID19 crisis on the enterprise's sales volumes. Furthermore, we introduce a forecasting model for the monthly sales and exploratively show how economic indicators improve the sales forecast and under which circumstances.

We define the following research questions:

RQ1: Was the enterprise's sales affected by the global economic effects of the COVID19 crisis?
RQ2: What is the impact on accuracy of economic business indicators and stock market indices on the enterprise's sales forecast?
RQ3: What constraints or organizational frameworks are necessary to apply better forecasting?

4.2 The State-of-the-Art in Sales Forecasting

In this study, we consider quantitative sales forecasting models that predict future sales.

Literature of sales forecasting focuses on two main aspects of research: The choice of input variables and the choice of forecasting algorithms. Both dimensions are widely studied. As input variables, studies take variables such as weather, promotions, competitors' behavior, product lifecycle, pricing, or macroeconomic data – highly depending on the particular use case and industry. For example, Sagaert et al. employ more than 60,000 macroeconomic indicators in an approach to forecast sales of tires in the EU and the US.

Forecasting algorithms employed range from statistical models like multiple linear regression, ARIMA, GARCH or Bayesian approaches to machine learning approaches like Support Vector Machines (SVM), Decision Trees or Artificial Neural Networks (ANN) inter alia (Liu et al. 2013; Tsoumakas 2019; Pavlyshenko 2019).

In recent time, forecasting competitions have gained increasing popularity in the community of forecasting researchers and practitioners. Based on comparable data sets and evaluation principles, the applied models are easy to compare. Competitions like the Makridakis' M1 to M5 competition and several competitions on Kaggle have therefore boosted the advancement of forecasting models for academic and practical use cases (Hyndman 2020). Analysis of the latest competitions show, that boosted decision tree ensembles provided amongst the best solutions of the competitions. XGBoost (Chen and Guestrin 2016) and LightGBM (Ke et al. 2017) are on the top scores of most competition leaderboards (Bojer and Meldgaard 2020).

Research also addresses sales forecasting under extreme events. One should note that this article does not consider prediction techniques of extreme events (cf. (Makridakis and Bakas 2016)). Instead, extreme events are taken as external shocks. Once they occur, the forecasting models' reaction on these shocks is observed.

In more general terms, we found articles that study resilience of systems under extreme events and ways to forecast the impact of the events on the system. Wang et al. review articles that focus on the forecast of the resilience of power grids to natural desasters. They conclude that the accuracy of such forecasts is low or that the forecasting models are highly instance-specific (Wang et al. 2016). Rajesh applies secondary data of enterprises to predict indicators of resilience of supply chains to extreme events in a case study – using a grey model (Rajesh 2016).

Specifically, in the context of sales forecasting, we find studies that address the impact of global events on sales. Bonham et al. study the impact of the 9/11 event on tourism on Hawaii using vector error correction models (Bonham et al. 2006). They conclude that they observe a positive effect on domestic visitors in the following months after the event and a negative on foreign visitors. Peels et al. propose a system dynamic model to forecast sales of a Dutch chemicals enterprise in response to the financial crisis after 2008 (Peels et al. 2009). Wang et al. apply an adaptive neuro fuzzy inference system to improve sales forecasts in the automotive industry and the impact of the financial crisis (Wang et al. 2011).

Only few studies research how the recent COVID19 pandemic crisis affects sales forecasting. Yang et al. forecast differences in restaurant demand by daily cases of COVID19 infections as result of a panel data model (Yang et al. 2020). Shen et al. show that especially companies in already difficult economic situations suffer from a decrease in sales (Shen et al. 2020).

To our best knowledge, no study so far combines analysis of the financial crisis and COVID19 pandemic crisis in a sales forecasting use case.

Even though various studies and techniques exist to facilitate sales forecasting, the actual degree of application of the techniques is low. Recent studies show that sales forecasting in companies lacks behind the possibilities that are provided by research (Lawrence et al. 2000). Furthermore, literature also shows that companies, especially small- and medium-sized enterprises (SME) are hesitant to implement state-of-the-art machine learning models (Bauer et al. 2020).

4.3 Characterization of the Case study

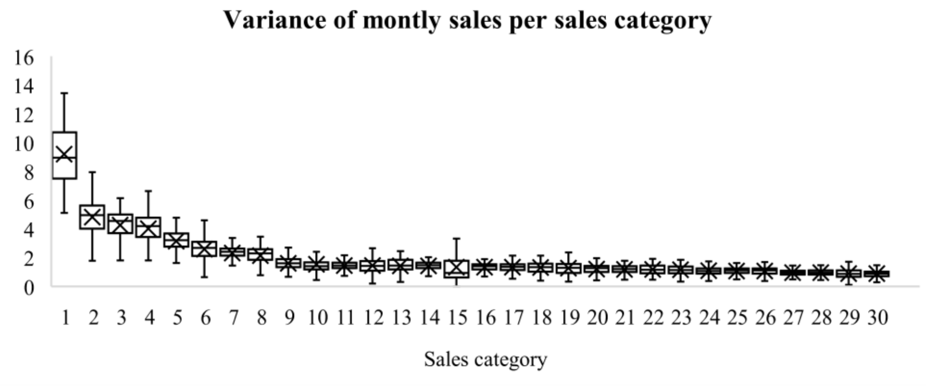


Figure 9: Box-Whisker plot of the monthly sales of the top 30 sales categories (out of 71, by total sales volume), normalized by average monthly sales per sales category. Sales are heterogenous amongst sales categories.

For this article, we study the monthly sales of a German enterprise. The enterprise is a medium-sized dealer of industrial parts in the field of business-to-business (B2B) with annual revenues of approximately 100 million €. The enterprise mainly serves the German market for supplies of industrial C-parts. To comply with confidentiality restrictions of the enterprise, all sales values in the figures below are normalized.

The enterprise divides their customers into 71 sales categories depending on the customer's field of business. We conducted forecasts of monthly sales per sales category for three-months-ahead each month for the enterprise's sales department.

Figure 9 depicts the historic distribution of sales per month and per sales category for the top 30 sales categories (normalized values). We observe a high variance in the monthly sales. Total average normalized sales volume per month and sales category accounts to 1 and the average variance accounts to 2.14.

The enterprise provided detailed sales data per month and sales category for a period from January 2008 to October 2020. Figure 10 depicts the development of annual sales for all sales categories of this period.

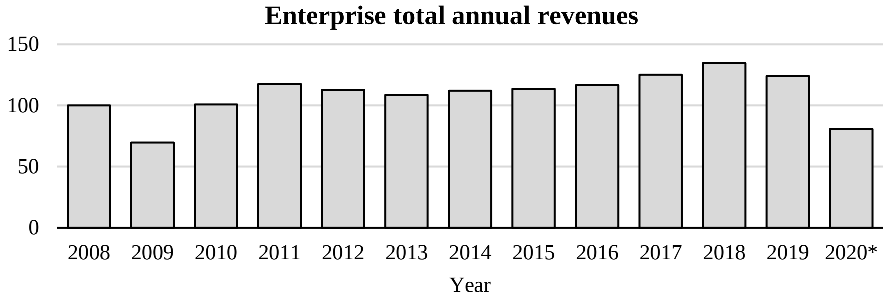


Figure 10: Total annual revenues of the enterprise, normalized to the year 2008 (100 = revenue of the year 2008). * Year 2020 includes months up until October.

4.4 Modelling the Impact of Global Events

To study the research questions as introduced in section 4.1, we apply a model with two indicators for the economic situation: the OECD Business Confidence Indicator (BCI) and stock market values of major market indices (Leading indicators 2017). The indicators exist for different countries.

In the period between 2008 and 2020, two major global events with high impact on economies took place, the global financial crisis in the years 2009 and 2010 (also named subprime mortgage crisis) and the global COVID19 pandemic crisis in the year 2020.

The events took place at different time for different regions; therefore, we consider the indicators BCI and stock market indices for the two major global economies (China and USA) as well as for the target market of the enterprise (Germany). The OECD BCI exists for each of these countries and additionally for the OECD countries in total. To represent the stock markets, we took the greatest stock market index per country: SSE Shanghai Composite for China, Dow Jones Industrial Average for the US and DAX for Germany (see Figure 11 and Figure 12).

The economic indicators clearly indicate the global events named above by extreme decreases in the years 2009, 2010 and 2020.

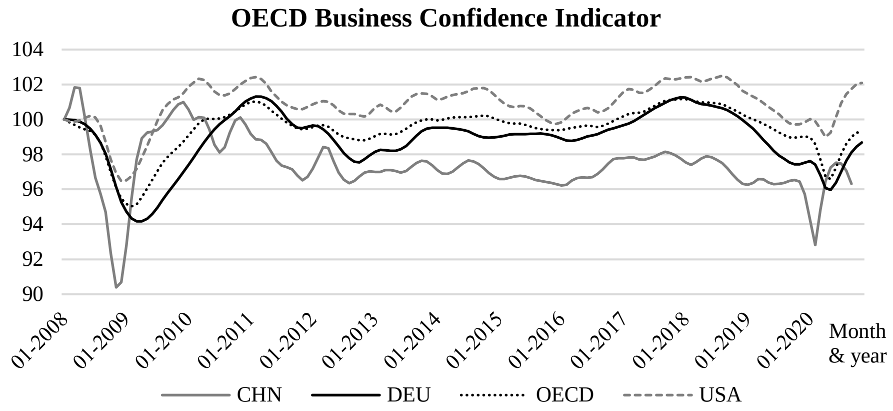


Figure 11: Business Confidence Indicators provided by OECD for PR of China (CHN), Germany (DEU), OECD countries and the United States of America (USA), normalized to January 2008 (100 = index value of January 2008) (Leading indicators 2017)

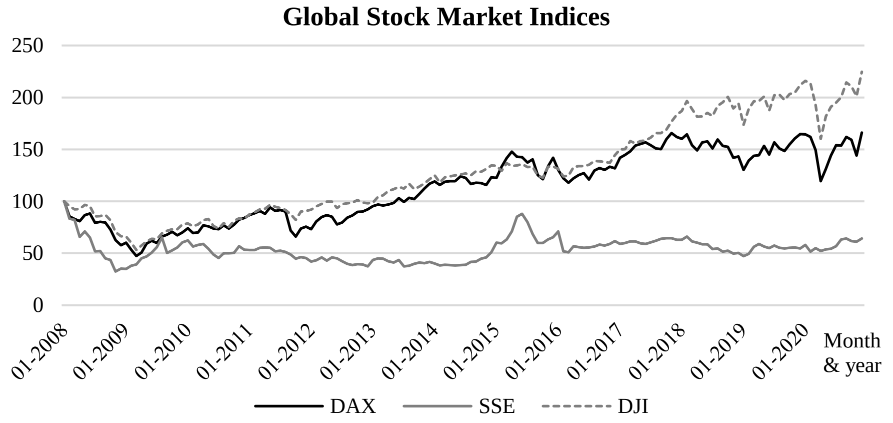


Figure 12: Monthly opening prices of global stock market indices DAX (DAX performance-index ^GDAXI), SSE (SSE Composite Index 000001.SS), DJI (Dow Jones Industrial Average ^DJI), normalized to January 2008 (100 = value of January 2008).

When we compare the enterprise's annual sales (see Figure 10) with the indicators (see Figure 11 and Figure 12), we observe periods of lower sales in the years 2009 and 2010 as well as 2020 which coincide with the global events.

The coefficient of correlation also suggests correlations between the indicators and the sales data – even though the Pearson correlation coefficients are not exceptionally high (see Table 18).

Indicator	OECD Business Confidence Indicator				Stock market indices		
	CHN	DEU	OECD	USA	DAX	SSE	DJI
Coefficient of correlation (Pearson)	-0.28	0.60	0.58	0.38	0.47	0.18	0.42

Table 18: Overview of the correlation coefficients of the economic indicators and the total monthly sales volumes. See Figure 11 and Figure 12 for abbreviations.

The judgement of the sales managers of the enterprise also suggests that the economic recessions during the two global events affected a decrease of the enterprise's sales in the periods.

Conclusion 1: The studied enterprise's sales volumes are affected by global events, namely the financial crisis and the COVID19 pandemic crisis – *see RQ1*.

In the following, we describe the forecasting model used in this case study. The input to the model consists of:

- the monthly revenues per sales category
- the values of the economic indicators (BCI and stock market indices)
- the month and year
- an ID of the sales category.

As forecasting algorithm, we applied a Histogram-based Gradient Boosting Regression Tree (HGBRT) from scikit-learn, which is an implementation of the LightGBM algorithm (Pedregosa et al. 2011; Ke et al. 2017) due to high performance in comparable cases in literature (see section 4.2). The prediction results were cross-validated by kfold cross-validation (Bergmeir et al. 2018). The model's output is a prediction for the sales volume of each sales category for the month "three-month ahead".

To compare the impact of the economic indicator input, we created an experiment design as shown in Table 19. Experiments *Exp01* and *Exp02* do not contain economic indicators. *Exp03* and *Exp04* contain OECD BCI values, *Exp05* and *Exp06* contain stock market indices. *Exp07* and *Exp08* combine all indicators. The experiment design so far contributes to the research questions RQ1 and RQ2.

We also study whether economic indicators themselves enable the model to forecast the impact of global events or if such events must have taken place during the training phase for the algorithm to correctly learn the relationship between sales and economic indicators. Therefore, we train the models in experiments *Exp02*, *Exp04*, *Exp06* and *Exp08* without the period from the years 2008 to 2011 – the period of the global financial crisis. This part of the experiment design contributes to RQ3.

When the models were trained and evaluated, we observed a substantial influence of randomness in the models. The implementation of the HGBRT algorithm takes a random state as input which makes randomness reproducible. To decrease the effect of randomness of the algorithm, each experiment was conducted with the same 1,000 different random states. The model was trained and evaluated with the entire dataset for each random state.

Experiment	Exp01	Exp02	Exp03	Exp04	Exp05	Exp06	Exp07	Exp08
Training period	2008-2019	2011-2019	2008-2019	2011-2019	2008-2019	2011-2019	2008-2019	2011-2019
OECD BCI	No	No	Yes	Yes	No	No	Yes	Yes
Stock market indices	No	No	No	No	Yes	Yes	Yes	Yes
Random states	1,000							

Table 19: The experiment design combines different training periods (from 2008 or from 2011), and different economic indicators (OECD BCI – see Figure 11, stock markets indices – see Figure 12).

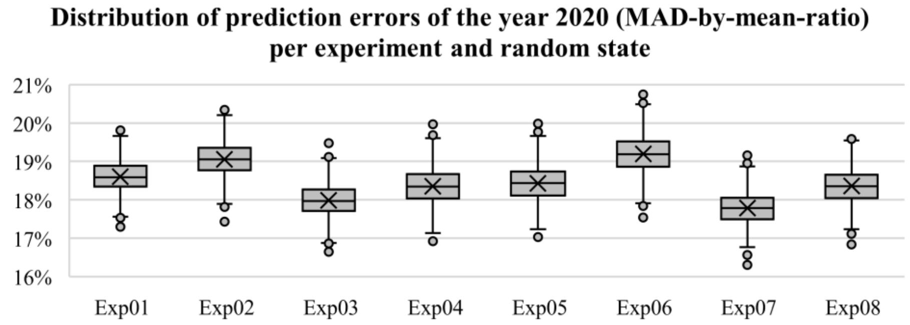
4.5 Discussion of Results

We evaluated all predictions for the year 2020 (each month from January to October) using the MAD-by-mean-ratio as evaluation score, which is the ratio of the mean absolute deviation between prediction and actual target variable value (MAD) divided by the mean value of the target variable (Kolassa and Schütz 2007). For each experiment and random state, we calculated the value of the MAD-by-mean-ratio. The distribution of these values is shown in Figure 13.

Based on the results, we observe that the distributions of the experiment results are relatively homogenous. The results per experiment are approximatively distributed normally around their mean value according to the Kolmogorov–Smirnov test with comparable variance. Therefore, we consider the mean values for comparison in the following.

The mutual paired samples t-test for means as well as an ANOVA test of the experiments suggests at high levels of significance that mean values of the experiments are unlikely to be equal ($\alpha \ll 0.05$). Therefore, we can assume that the mean values of the distributions differ significantly.

First, we compare the evaluation results of the experiments *Exp01*, *Exp03*, *Exp05* and *Exp07*. They vary in the inclusion of the economic indicators but are all trained on the full period from 2008 to 2019 and are therefore suitable to give answers to the research questions *RQ1* and *RQ2*. *Exp03* includes the OECD BCI indicators and exhibits an evaluation score that is an improvement of 3.3% compared to *Exp01*, which includes no economic indicators. *Exp05* includes the stock market indicators and exhibits an improvement of 0.9% compared to *Exp01*. *Exp07*, which includes all indicators, exhibits an improvement of 4.4% over *Exp01*.



Mean value of MAD-by-mean-ratios per experiment

18.61% 19.05% 17.99% 18.36% 18.43% 19.20% 17.79% 18.36%

Figure 13: Overview of the MAD-by-mean-ratio for the sales prediction of the year 2020 (each month from January to October) per experiment based on 1,000 random samples.

From these results, we conclude that the inclusion of the economic indicators helps the model to predict the sales values during the COVID19 crisis compared to the model that does not include any indicators. The contribution of the OECD BCI indicator is higher than of the stock market indices. However, the indicators combined result in even better model predictions than models with only one of the indicators.

Conclusion 2: Economic indicators as additional input for the forecasting model improve the accuracy of the prediction by 3.3% to 4.4% – *see RQ2*.

Second, we compare pairs of models that vary only in the training period, i.e., *Exp01* with *Exp02*, *Exp03* with *Exp04* and so on. We find that the results' evaluation scores of experiments with training periods that include the period of the financial crisis (2008 to 2011) are better than of the experiments that leave this period out. This holds true for each pair of experiments.

The pair of experiments *Exp01* and *Exp02* does not include any economic indicators. The difference in the evaluation scores accounts to 2.3%. This difference cannot be justified by any learning of the model related to the economic indicator. The difference of the evaluation scores of *Exp03* and *Exp04* accounts to 2.0%– therefore we conclude that the effect probably is also not justified by economic indicator related reasons.

However, the pairs *Exp05* / *06* as well as *Exp07* / *08* exhibit higher differences (4.0% and 3.1%). This indicates that the model attributes a higher weight to the input of the economic indicators when a past global event can be linked to an actual reaction of the sales volumes of the enterprise.

Conclusion 3: The results indicate that the models that are trained on past global events in combination with explanatory economic indicators reflect better results in predicting future global events that are also indicated by economic indicators – *see RQ3*.

4.6 Conclusion and Outlook

In this study, we raised three research questions *RQ1* to *RQ3* that were studied based on the case study of one enterprise. We described how the global financial crisis and the COVID19 pandemic crisis – affected the sales volumes of the enterprise of this case study (*RQ1*). Furthermore, in this study we describe that models trained with two specific economic indicators that reflect the general economic consequences of the global events, improve the forecasting accuracy of the enterprise's sales during these events (*RQ2*). Finally, we find that the models trained with past events of the investigated global crisis affecting the enterprise can, in some cases, enable the models to correlate economic indicators and their effects on the enterprise more precisely (*RQ3*).

This article shows a case study where economic indicators are effectively applied to improve sales forecasts of enterprises dependent of global events. While it is reasonable to assume that the effects shown in this study can also be observed in similar cases of sales forecasting, we are aware that this study only exhibits one specific case study and therefore the study cannot yield general statements for every situation. This fact will be motivation for further comparative studies with more enterprises.

Furthermore, this study does not mean that the models are able to forecast future global events before they take place. Instead, this study shows how economic indicators help to predict how the enterprise's sales are affected by global events.

As an outlook, we can expect that global events will continuously occur in the future. Therefore, it is reasonable to evaluate the application of economic indicators in similar enterprises' sales forecasts. The study shows that the indicators enhance forecasting accuracy. However, enterprises will rely on high quality and up to date indicators that are suitable for prediction. Future research should therefore concentrate on fast-reacting and conclusive indicators for global events.

4.7 References

Bauer, Markus; Kiefer, Daniel; Grimm, Florian (2021): Sales Forecasting Under Economic Crisis: A Case Study of the Impact of the COVID19 Crisis to the Predictability of Sales of a Medium-Sized Enterprise. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 163–172.

Bauer, Markus; van Dinther, Clemens; Kiefer, Daniel (2020): Machine Learning in SME: An Empirical Study on Enablers and Success Factors. In : AMCIS 2020 Conference Proceedings.

Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo (2018): A note on the validity of cross-validation for evaluating autoregressive time series prediction. In *Computational Statistics & Data Analysis* 120, pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

Bojer, Casper Solheim; Meldgaard, Jens Peder (2020): Kaggle forecasting competitions: An overlooked learning opportunity. In *International Journal of Forecasting*. DOI: 10.1016/j.ijforecast.2020.07.007.

Bonham, Carl; Edmonds, Christopher; Mak, James (2006): The Impact of 9/11 and Other Terrible Global Events on Tourism in the United States and Hawaii. In *Journal of Travel Research* 45 (1), pp. 99–110. DOI: 10.1177/0047287506288812.

Chen, Tianqi; Guestrin, Carlos (2016): XGBoost. In Balaji Krishnapuram, Mohak Shah, Alex Smola, Charu Aggarwal, Dou Shen, Rajeev Rastogi (Eds.): Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA, 13 08 2016 17 08 2016. New York, NY: ACM, pp. 785–794.

Global Economic Prospects, June 2020 (2020): Washington, DC: World Bank.

Hyndman, Rob J. (2020): A brief history of forecasting competitions. In *International Journal of Forecasting* 36 (1), pp. 7–14. DOI: 10.1016/j.ijforecast.2019.03.015.

Ke, Guolin; Meng, Qi; Finley, Thomas; Wang, Taifeng; Chen, Wei; Ma, Weidong et al. (2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.): Advances in Neural Information Processing Systems, vol. 30: Curran Associates, Inc, pp. 3146–3154. Available online at https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kolassa, Stephan; Schütz, Wolfgang (2007): Advantages of the MAD/Mean Ratio over the MAPE. In *Foresight: The International Journal of Applied Forecasting* (6), pp. 40–43. Available online at https://EconPapers.repec.org/RePEc:for:ijafaa:y:2007:i:6:p:40-43.

Lawrence, Michael; O'Connor, Marcus; Edmundson, Bob (2000): A field study of sales forecasting accuracy and processes. In *European Journal of Operational Research* 122 (1), pp. 151–160. DOI: 10.1016/S0377-2217(99)00085-5.

Leading indicators (2017): OECD.

Liu, Na; Ren, Shuyun; Choi, Tsan-Ming; Hui, Chi-Leung; Ng, Sau-Fun (2013): Sales Forecasting for Fashion Retailing Service Industry: A Review. In *Mathematical Problems in Engineering* 2013 (4), pp. 1–9. DOI: 10.1155/2013/738675.

Makridakis, Spyros; Bakas, Nikolas (2016): Forecasting and uncertainty: A survey. In *RDA* 6 (1), pp. 37–64. DOI: 10.3233/RDA-150114.

Pavlyshenko, Bohdan (2019): Machine-Learning Models for Sales Time Series Forecasting. In *Data* 4 (1), p. 15. DOI: 10.3390/data4010015.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. et al. (2011): Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peels, Robert; Udenio, Maximiliano; Fransoo, Jan C.; Wolfs, Marcel; Hendrikx, Tom; NeoResins, D. S.M. (2009): Responding to the Lehman Wave: Sales forecasting and supply management during the credit crisis. In *Dec* 5 (2697), pp. 1–20.

Rajesh, R. (2016): Forecasting supply chain resilience performance using grey prediction. In *Electronic Commerce Research and Applications* 20 (3), pp. 42–58. DOI: 10.1016/j.elerap.2016.09.006.

Sagaert, Yves R.; Aghezzaf, El-Houssaine; Kourentzes, Nikolaos; Desmet, Bram (2018): Tactical sales forecasting using a very large set of macroeconomic indicators. In *European Journal of Operational Research* 264 (2), pp. 558–569. DOI: 10.1016/j.ejor.2017.06.054.

Shen, Huayu; Fu, Mengyao; Pan, Hongyu; Yu, Zhongfu; Chen, Yongquan (2020): The Impact of the COVID-19 Pandemic on Firm Performance. In *Emerging Markets Finance and Trade* 56 (10), pp. 2213–2230. DOI: 10.1080/1540496X.2020.1785863.

Tsoumakas, Grigorios (2019): A survey of machine learning techniques for food sales prediction. In *Artif Intell Rev* 52 (1), pp. 441–447. DOI: 10.1007/s10462-018-9637-z.

Wang, Fu-Kwun; Chang, Ku-Kuang; Tzeng, Chih-Wei (2011): Using adaptive network-based fuzzy inference system to forecast automobile sales. In *Expert Systems with Applications* 38 (8), pp. 10587–10593. DOI: 10.1016/j.eswa.2011.02.100.

Wang, Yezhou; Chen, Chen; Wang, Jianhui; Baldick, Ross (2016): Research on Resilience of Power Systems Under Natural Disasters—A Review. In *IEEE Trans. Power Syst.* 31 (2), pp. 1604–1613. DOI: 10.1109/TPWRS.2015.2429656.

Yang, Yang; Liu, Hongbo; Chen, Xiang (2020): COVID-19 and restaurant demand: early effects of the pandemic and stay-at-home orders. In *IJCHM* 13 (12), pp. 3809–3834. DOI: 10.1108/IJCHM-06-2020-0504.

5 How Time Series Characteristics Affect the Forecast Quality in State-of-the-Art Algorithms for Intermittent Demands

[*This chapter summarizes three articles published in 2021:*

Kiefer, Daniel; Grimm, Florian; Bauer, Markus; van Dinther, Clemens (2021b): Demand Forecasting Intermittent and Lumpy Time Series: Comparing Statistical, Machine Learning and Deep Learning Methods. In : Hawaii International Conference on System Sciences 2021. Honolulu, HI: University of Hawai'i at Manoa, Hamilton Library, p. 1425. – *further referred to as* Kiefer et al. 2021b

Kiefer, Daniel; Bauer, Markus; Grimm, Florian (2021a): Univariate Time Series Forecasting: Machine Learning Prediction of the Best Suitable Forecast Model Based on Time Series Characteristics. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 152–162. – *further referred to as* Kiefer et al. 2021a

Grimm, Florian; Kiefer, Daniel; Bauer, Markus (2021): Univariate Time Series Forecasting by Investigating Intermittence and Demand Individually. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 143–151. – *further referred to as* Grimm et al. 2021]

The three studies summarized in this chapter primarily address research question RQ III of this dissertation: ”What is the forecasting performance of machine learning methods com-pared to classical approaches for intermittent time series and how can ap-proaches be selected depending on the time series characteristics?”, In particular, the first study (Kiefer et al. 2021b) addresses the comparison of classical and machine learning methods for forecasting intermittent time series. The second study (Kiefer et al. 2021a) focuses on investigating the interrelationships of time series properties and the choice of forecasting method. The third study (Grimm et al. 2021) analyzes the hybridization of Croston's method with machine learning algorithms.

5.1 Comparing the performance of state-of-the-art algorithms on intermittent and non-intermittent demand time series

In Kiefer et al. 2021b, we studied the practical application of demand forecasting algorithms application-oriented demand time series. Our literature research yielded a gap in research to determine what algorithms performs best under which circumstances, especially when the goal is to predict data sets containing intermittent time series. We therefore systematically benchmarked the state-of-the-art algorithms based on a publicly available data set and compared the forecasting performances.

In the study, we raised the following two research questions:

RQ1: Do modern advanced deep learning methods achieve considerably better forecasts than classic, established statistical methods and machine learning methods in forecasting demand for intermittent and lumpy time series?

RQ2: Under which time series characteristics, in particular the degree of intermittent behavior and lumpiness of the time series, do deep learning methods achieve superior results and vice versa?

5.1.1 Related work

The literature research was conducted following the principles of Webster and Watson and Levy and Ellis (Webster and Watson 2002; Levy and Ellis 2006).

The basic foundational research reviewed for this study constitutes the state-of-the-art algorithms applied in this study. First, we review classical approaches for intermittent time series forecasting, like basic ARIMA and Croston's approach (Croston 1972) and its advancements like SBA and TSB (Syntetos and Boylan 2001) as well as general time series forecasting algorithms like Holt-Winter's). Second, we studied general machine learning algorithms (SVR, Random Forest and boosted tree approaches) and deep learning algorithms (multi layer perceptron MLP, LSTM). In addition, we find algorithms further adapted to intermittent demand forecasting, like the approach of Kourentzes or Willemain et al. (Kourentzes 2013; Willemain et al. 1994).

5.1.2 Approach

We propose a systematic empiric approach to compare the state-of-the-art forecasting algorithms on a data set.

As data set, we chose the M5 competition data set. It contains over 100,000 time series with 1,941 time steps each, representing Walmart product sales (Makridakis et al. 2022). The time series were classified as smooth, erratic, intermittent and lumpy, as defined by Williams (Williams 1984). For the experiments, we randomly chose 30 time series from each class, totaling 120 time series per experiment. Only the univariate time series were considered for training and testing of the algorithms, i.e., no additional inputs except for the past values of the target variable were provided.

We applied a rolling window cross validation scheme as proposed by Bergmeir et al., with a forecasting horizon of 28 days ahead. As evaluation metrics, we applied the mean absolute scaled (MASE) metric as proposed by Hyndman and Koehler and used as standard metric in the M5 competition (Hyndman and Koehler 2006; Makridakis et al. 2022). In addition, we applied the stock-keeping-oriented prediction error costs (SPEC) metric, as proposed by Martin et al., which accounts for both the costs of over-stocks as well as under-stocks caused by forecasting errors (Martin et al. 2020).

In the study, we applied three classical statistical algorithms (Croston's method, Holt-Winter triple exponential smoothing, ARIMA with autotuned parameters), three machine learning algorithms (Random Forest, XGBoost, SVR with autotuned parameters) and three deep learning algorithms (MLP, LSTM with two different topologies).

Details concerning the chosen parameters for the setup of the algorithms, evaluation metrics and cross validation are given in the published study (Kiefer et al. 2021b).

5.1.3 Results and conclusions

The results of the experiments are shown in the table below, where Table 20 summarizes the results of all time series and the following tables show the results for smooth (Table 21), intermittent (Table 22), erratic (Table 23) and lumpy (Table 24) time series.

Table 20: Forecasting evaluation results
All time series

	Ø SPEC		**Ø MASE**		**Rank**
Statistic					
Croston	4.75	(0%)	2.15	(0%)	1
Holt-Winter	7.41	(56%)	1.08	(50%)	4
Auto-ARIMA	6.93	(46%)	1.04	(52%)	3
Machine Learning					
Random Forest	8.19	(73%)	1.14	(47%)	5
XGBoost	12.01	(153%)	1.13	(48%)	9
Auto-SVR	9.98	(110%)	1.10	(49%)	7
Deep Learning					
MLP	10.15	(114%)	1.66	(23%)	8
LSTM	5.96	(26%)	0.98	(54%)	2
LSTM-2	8.57	(81%)	1.04	(52%)	6

Throughout all classes, the best ranked algorithm by SPEC metric (i.e., lowest SPEC values) is the Croston algorithm. Second in rank is the LSTM neural network algorithm–except for intermittent time series, where the ARIMA algorithm is in second rank. The Holt-Winter algorithm typically ranks third or fourth, except for lumpy time series.

We observe a different picture when the forecasting accuracy is evaluated under the MASE metric. Here, the LSTM algorithms performs best overall (i.e., lowest MASE values) and especially for smooth, erratic and intermittent time series. Only for lumpy time series, the ARIMA algorithm yields the best rank. Evaluated under the MASE criterion, Croston's algorithm yields the last rank in overall an in all classes. In addition, we observe that only LSTM achieves MASE scores below 1.0, i.e., yields a lower forecasting error than the naïve forecast used by the MASE metric as baseline algorithm. All other algorithms yield scores considerably higher than 1.0.

Table 21: Forecasting evaluation results Smooth time series

	Ø SPEC		Ø MASE		Rank
Statistic					
Croston	2.02	(0%)	1.85	(0%)	1
Holt-Winter	3.45	(71%)	1.06	(43%)	3
Auto-ARIMA	3.76	(86%)	1.03	(44%)	4
Machine Learning					
Random Forest	4.29	(112%)	1.04	(44%)	7
XGBoost	5.25	(160%)	1.16	(37%)	9
Auto-SVR	4.05	(100%)	1.00	(46%)	6
Deep Learning					
MLP	4.31	(113%)	1.54	(17%)	8
LSTM	2.24	(11%)	0.96	(48%)	2
LSTM-2	3.99	(97%)	1.03	(44%)	5

Table 22: Forecasting evaluation results Intermittent time series

	Ø SPEC		Ø MASE		Rank
Statistic					
Croston	10.42	(0%)	1.59	(0%)	1
Holt-Winter	13.45	(29%)	1.16	(27%)	4
Auto-ARIMA	12.32	(18%)	1.12	(29%)	2
Machine Learning					
Random Forest	17.35	(67%)	1.27	(20%)	6
XGBoost	23.21	(123%)	1.20	(24%)	9
Auto-SVR	20.11	(93%)	1.09	(32%)	7
Deep Learning					
MLP	20.97	(101%)	1.49	(6%)	8
LSTM	12.38	(19%)	1.14	(28%)	3
LSTM-2	16.82	(61%)	1.07	(33%)	5

Table 23: Forecasting evaluation results Erratic time series

	Ø SPEC		Ø MASE		Rank
Statistic					
Croston	5.04	(0%)	2.18	(0%)	1
Holt-Winter	9.90	(96%)	1.07	(51%)	5
Auto-ARIMA	9.15	(81%)	1.04	(52%)	4
Machine Learning					
Random Forest	8.65	(72%)	1.15	(47%)	3
XGBoost	15.05	(198%)	1.10	(49%)	9
Auto-SVR	12.35	(145%)	1.06	(52%)	8
Deep Learning					
MLP	11.45	(127%)	1.68	(23%)	7
LSTM	7.29	(45%)	0.97	(56%)	2
LSTM-2	11.12	(121%)	1.04	(52%)	6

Table 24: Forecasting evaluation results Lumpy time series

	Ø SPEC		Ø MASE		Rank
Statistic					
Croston	1.51	(0%)	3.00	(0%)	1
Holt-Winter	2.83	(88%)	1.03	(66%)	6
Auto-ARIMA	2.49	(66%)	0.97	(68%)	5
Machine Learning					
Random Forest	2.46	(64%)	1.11	(63%)	4
XGBoost	4.54	(202%)	1.05	(65%)	9
Auto-SVR	3.40	(126%)	1.27	(58%)	7
Deep Learning					
MLP	3.87	(157%)	1.91	(36%)	8
LSTM	2.02	(34%)	1.00	(67%)	2
LSTM-2	2.35	(56%)	1.02	(66%)	3

In the study, we draw the following conclusions:

Machine and deep learning yield good results, however, not always better than classical algorithms (referring to research question 1.).

Under the SPEC metric, the classical algorithms Croston and ARIMA dominated the other algorithms–also deep learning algorithms–in all classes. Under the MASE metric, the LSTM algorithm scores best overall and in the classes, except for lumpy time series. It is also the only algorithm that performs better than the naïve baseline algorithm (referring to research question 2.).

In addition, we point out the opposing results depending on the metric applied. The best algorithm under SPEC yields the worst rank under MASE. This underlines that the choice of metric must be taken carefully, depending on the interpretation of the metric and the intended use case.

This study contributes to the research by systematically comparing state-of-the-art forecasting algorithms to a publicly available data set and by the close examination of the differences of the forecasting performances for time series with different time series characteristics. We could show under which circumstances classical out-perform machine and deep learning algorithms and which algorithms did not play a distinctive role in the experiments.

5.2 Predicting the performance of algorithms based on time series specific characteristics

The study Kiefer et al. 2021a is based on the insights from Kiefer et al. 2021b. In the previous work, we studied the different forecasting performances of algorithms for the classes of smooth, erratic, intermittent and lumpy time series. In this thereupon based study, we further examine how forecasting algorithms can be optimally chosen based on univariate time series characteristics.

RQ1: Can a machine learning model predict the best suitable algorithm for univariate time series forecasting based on time series characteristics?

RQ2: Could dependencies of demand patterns and model forecast skill be identified?

RQ3: Can such a classification algorithm be used advantageously in the business context?

5.2.1 Related work

As the previous study, this study relates to the basic literature on classical forecasting approaches as ARIMA or Holt-Winters, and machine learning approaches (including deep learning approaches) as SVR or LSTM networks.

This study also relates to the literature studying evaluation metrics, while this study primarily relies on the MAE evaluation metric.

Notable specific literature relevant to this study represent the Croston approach for intermittent time series (Croston 1972), the work on classification of intermittent time series by Syntetos et al. (Syntetos et al. 2005), and the study comparing different intermittent time series forecasting approaches by Kourentzes (Kourentzes 2013). As before mentioned, this study is based on Kiefer et al. 2021b.

5.2.2 Suggested experimental design

The approach is based on the previous study in Kiefer et al.: The time series of a data set is classified into the classes smooth, erratic, intermittent and lumpy. The classification is done based on the metrics average demand interval (ADI) and squared coefficient of variation (CV2) as proposed by Williams and Syntetos et al. (Williams 1984; Syntetos et al. 2005).

In this study, a proprietary data set from a retailer for industrial business-to-business goods is used, comprising 200,000 time series of daily sales with up to 3,960 time steps per time series. For the experiments, 16,035 time series are randomly chosen. Apart from the historic values of the target variable, the ADI and CV2 and the classification, no additional inputs were used to train the forecasting algorithms.

For cross validation, we applied rolling window validation as proposed by Bergmeir et al. (Bergmeir et al. 2018). The mean average error (MAE) was used as evaluation metric for forecasting accuracy.

First, for demand forecasting, we implemented classical algorithms (Holt-Winters, linear regression), machine learning algorithms (XGBoost, random forest, SVR) and deep learning algorithms (LSTM, CNN-LSTM, CONV-LSTM). Details on the topology and parametrization can be found in Kiefer et al. 2021a.

After generating forecasts with the before mentioned algorithms, the results were evaluated with the MAE metric and the best performing (i.e., lowest MAE score) algorithm chosen per time series.

Subsequently in a second step, classifier algorithms were trained to predict the best performing forecasting algorithm. To cross validate the classification prediction, a k-fold scheme with 5 folds was applied.

The following algorithms from the Scikit-learn Python library were benchmarked for classification: BoostClassifier, SupportVectorClassification, RandomForestClassifier, AdaBoostClassifier, GaussianNB, MLPClassifier, KNeighborsClassifier and GradienBoostingClassifier (Pedregosa et al. 2011). We applied the synthetic minority oversampling technique (SMOTE) to deal with the unbalanced distribution of time series in the classes.

5.2.3 Results and conclusion

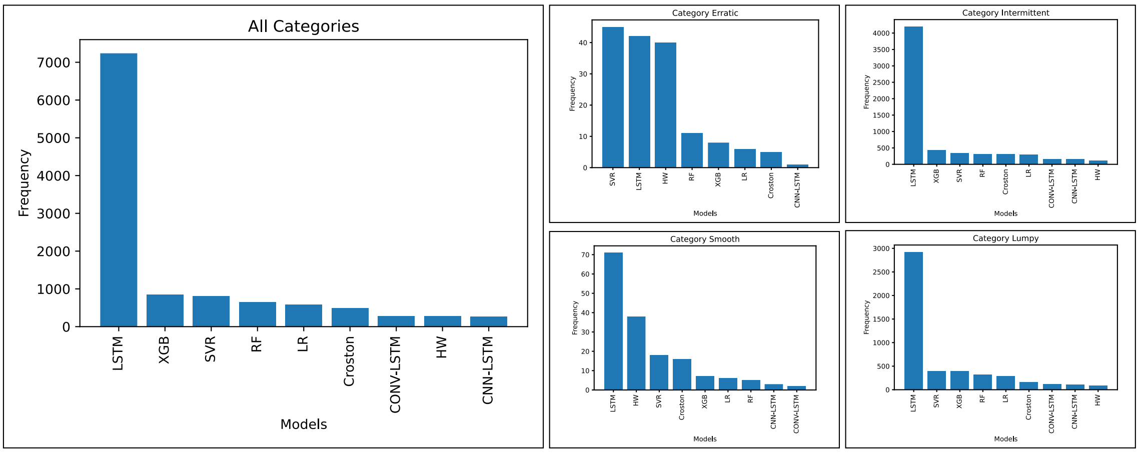


Figure 14: Best MAE score rank of the forecasting algorithms–overall and per class

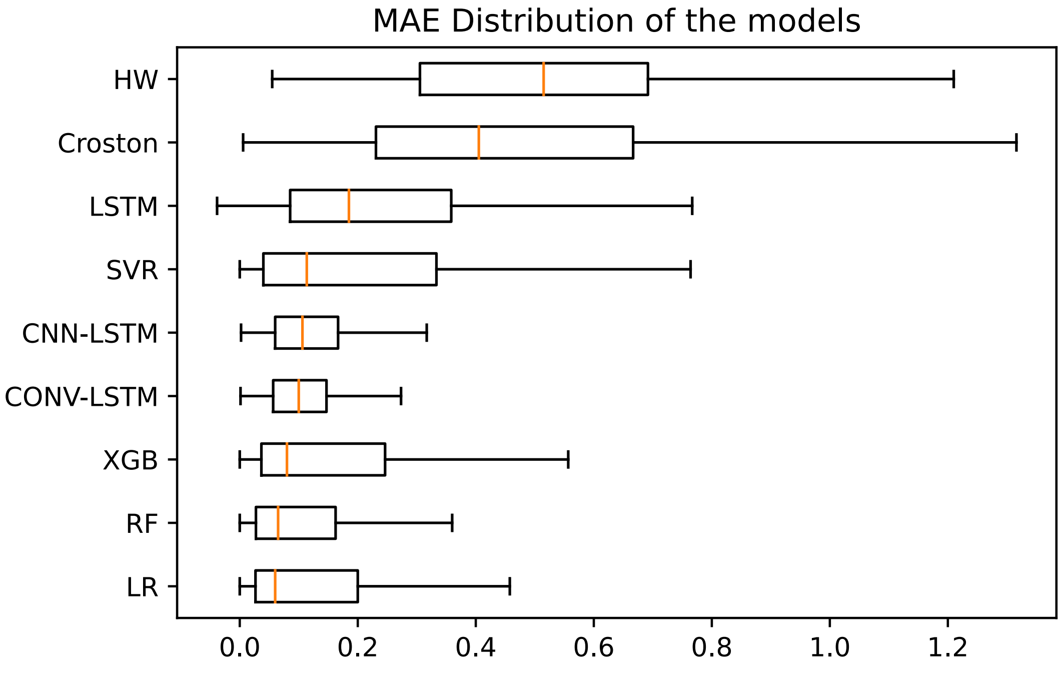


Figure 15: Histogram of the MAE score distribution for all classes

First, we summarize the results of the sales forecasts. Here we first observe the number of times that an algorithm achieves the best (i.e., lowest) MAE score for a time series ("best MAE score rank", see Figure 14). By far best ranking algorithm in this metric is the LSTM algorithm. The XGB and SVR rank second and third, however, the number of time series these algorithms achieve the best MAE score is about one seventh compared to the LSTM algorithm. For erratic time series, SVM, LSTM and Holt-Winters perform comparable forecasting accuracies based on the best MAE score ranking. In addition, we observe the distribution of MAE metric values over all time series (as opposed to the rank of the algorithm, see Figure 15). Here, the Holt-Winters and Croston algorithms perform bad (i.e., high) MAE values overall. CNN-LSTM, CONV-LSTM and the random forest algorithms perform lowest distribution and therefore perform best. Surprisingly, despite performing best in the beforementioned ranking score, the plain LSTM exhibits relatively bad MAE distribution values.

	ROC-AUC	**Accuracy**	**Precision**	**Recall**	**F1**
Mean value over all folds	0.89	0.76	0.76	0.76	0.76

Table 25: Classification accuracy results of the best classifier

Second, we examine the classification accuracy scores of classifier algorithms (Table 25). The evaluation metrics employed are ROC-AUC, accuracy, precision, recall and F1 metric. We concentrated on the ROC-AUC metric primarily. The best classifier algorithm from the tested was the Gradient Boosting Classifier. The mean ROC-AUC value is 0.89–which generally indicates a high classification accuracy.

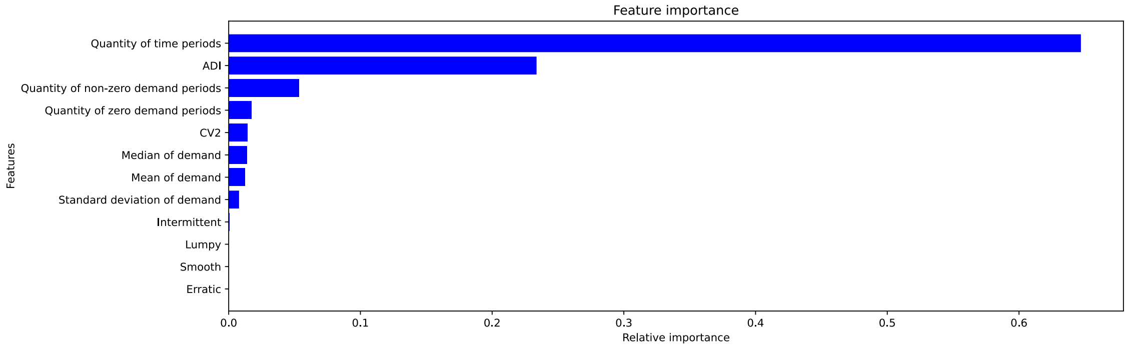


Figure 16: Feature importance scores of the classifier algorithm

To study which of the features provided to the algorithm are most important to predict the best forecasting algorithm, we applied the feature importance score approach (Figure 16). It showed that the quantity of time periods in the time series (i.e., time series length) was the feature with highest importance to the classifier. Second important feature is the ADI. The further relevant features, yet with lower importance scores are the quantity of zero and non-zero demand periods, CV2 and median, mean and standard deviation of the demand size.

We conclude in this study, referring to the initial research questions of the study:

The classifier algorithm predicted the best forecasting algorithm with a high ROC-AUC score of 89%. We therefore conclude for RQ1, that it is possible for a machine learning classifier to predict the best forecasting algorithm with high accuracy based on the time series characteristics of the time series.

For RQ2, we observe that the time series characteristics features identified in the feature importance analysis were used by the classifier algorithm to predict the best forecasting algorithm. We can therefore conclude that time series characteristics are a decisive criterion for which forecasting algorithm yields the best forecast accuracy.

Regarding RQ3, we conclude that the insights from the study can be used to efficiently chose the optimal forecasting algorithm, reducing computational efforts.

The study contributes to the current research by first showing that time series characteristics can be exploited to choose optimal forecasting algorithms. Second, we empirically showed what characteristics can be used, how to exploit them, and third showed what features were important in the given use case.

5.3 Improving the performance of intermittent time series forecasting by combining Croston's approach with machine learning techniques

In Grimm et al. 2021, we studied how Croston's approach can be leveraged by applying modern machine learning algorithms.

In general, Croston's approach is based on the consideration of forecasting the interval between demands and the level of demands separately for intermittent time series. At the time Croston's approach was published, machine learning methods were not yet in use, so Croston used exponential smoothing as a forecasting method for demand intervals and levels. So instead of exponential smoothing using modern machine learning approaches is an approach worth exploring today.

5.3.1 Related work

Croston proposed a two-step approach to forecast intermittent demand time series: First, forecasting of the intervals between non-zero demands and second forecasting the demand size (i.e., quantity of the demand) for the predicted non-zero demand period. In the study, Croston proposed exponential smoothing, for both the demand intervals and demand size (Croston 1972). Syntetos and Boylan, and Teunter et al. revised the approach to fix a systematical bias in Croston's approach and to account for obsolescence of items (Syntetos and Boylan 2001; Teunter et al. 2011).

Kourentzes proposes to improve Croston's approach by applying neural networks for forecasting of the demand intervals and demand size, thereby combining Croston's approach and machine learning. They conclude that their approach cannot achieve higher forecasting accuracy than classical approaches. However, they conclude that based on measuring service levels at increasing stocks, their approach performs better (Kourentzes 2013).

Turkmen et al. further develop Kourentzes' approach. Instead of training two separate models for demand interval and demand size prediction, they propose to use only one neural network which simultaneously predicts both variables and learns from all time series (Turkmen et al. 2019).

5.3.2 Approach

We base our study on the M5 data set, from which we selected randomly 150 time series, equally drawn from the categories food, household and hobby (Makridakis et al. 2022).

To prepare the data, we constructed an inter-demand interval and a non-zero demand size vector for each time series. For training and cross validation, the rolling window approach was chosen.

We use two separate models, of which each is trained on one of the before mentioned vectors and then predicted a value for a one-step ahead forecast. In a first experiment, we used pair of LSTM networks (SplitLSTM). In the second experiment, we used a pair of LGBM models (SplitLGBM).

As baseline algorithms, we used the Croston algorithm, and LSTM and LGMB algorithms which were directly applied to the time series (not the two vectors). The detailed topology and parametrization can be found in Grimm et al. 2021.

For evaluation of the forecasting accuracy, we applied the SPEC (Martin et al. 2020) and RMSSE metrics (Hyndman and Koehler 2006) as well as R2, MSE, MAE, normalized RMSE and MASE.

5.3.3 Results and conclusion

Model	MSE	MAE	R2	MASE	SPEC	nRMSE	RMSSE
CROSTON	7.18	**1.04**	**-0.14**	**0.95**	**16.88**	**1.87**	**0.86**
LSTM	13.41	1.26	-0.39	1.06	31.35	1.98	0.89
LGBM	**5.80**	1.08	-0.49	*1.07*	17.36	2.13	0.97
SplitLSTM	14.31	*1.28*	*-1.03*	1.04	27.73	*2.52*	*1.12*
SplitLGBM	*14.80*	1.24	-0.99	0.99	*39.29*	2.46	*1.12*

Table 26: Comparison of the prediction evaluations by metric and approach. SPEC parameters $\boldsymbol{\alpha_1 = \alpha_2 = 0.5}$. Lower values are better for all metrics.

Table 26 shows the forecasting performances of the experiments. We observe that the Croston algorithm shows best evaluation results for all metrics–except for MSE, where the baseline LGBM algorithm yields a better MSE value than all other algorithms. The SplitLSTM and SplitLGBM approaches do not yield better results but considerably higher error rates.

Model	$\alpha_1 = 0.25, \alpha_2 = 0.75$	$\alpha_1 = 0.5, \alpha_2 = 0.5$	$\alpha_1 = 0.75, \alpha_2 = 0.25$
CROSTON	18.83	**16.88**	14.92
LSTM	*44.09*	31.35	18.62
LGBM	20.06	17.36	**14.67**
SplitLSTM	**17.56**	27.73	37.89
SplitLGBM	22.17	*39.29*	*56.41*

Table 27: Evaluation results for the SPEC metric with different values for the parameters $\boldsymbol{\alpha_1}$ (opportunity costs) and $\boldsymbol{\alpha_2}$ (stock keeping costs).

The results in Table 27 show how different values for the α parameters of the SPEC metric influence the results. While shifting the parameter towards α_1 (note that $\alpha_1 + \alpha_2 \overset{!}{=} 1$), giving opportunity costs for out-of-stock events more importance, the LGBM algorithm improves compared to Croston's algorithm and the split algorithms evaluation results worsen. However, shifting the balance more to α_2 emphasizing the impact of stock keeping costs. In this case, the split algorithms perform better compared to the other parameter settings. SplitLSTM in this case even becomes the algorithm with the lowest (i.e., best SPEC) value.

Further examination led to the understanding that the split algorithms systematically understated demand, while the other algorithms rather overstated demand.

We therefore conclude the following insights in this study: Applying deep learning algorithms to improve Croston's approach does not naturally improve the forecasting results, as we see from the results in Table 26. However, we conclude that it can be beneficial under certain circumstances, as shown in Table 27, where we weighed stock keeping costs higher than out-of-stock opportunity costs. This result is structurally comparable to the findings of Kourentzes; however, they find that their approach shows advantages when out-of-stock costs are emphasized.

5.4 Conclusion of this chapter

The results of the first study (Kiefer et al. 2021b) of this chapter have shown that the tested machine learning algorithms are not necessarily better than the specialized classical algorithms. In fact, the Croston approach showed the best results in most experiments and among the datasets used. Machine Learning approaches do not prove to be superior in the experiments performed. The study contributes to the state of the art by systematically comparing these state of the art algorithms on known and public datasets.

The study also shows that the choice of evaluation metric is critical in determining which method is determined to be the best.

The second study (Kiefer et al. 2021a) showed that a classification algorithm can be used to predict the optimal forecasting method from a selection of methods with high accuracy, based on the time series properties of the forecasted objects. Thus, the study shows that the investigated time series properties are clearly related to the choice of the forecasting method and that this relationship can be quantified.

The third study (Grimm et al. 2021) confirms the results of the previous two in two aspects. First, the hybridized Croston and machine learning method does not outperform the classical method in the experiments performed–just as in the two previous studies other machine learning approaches did not show any systematic advantage. Second, in the third study, as in the other two, it is shown that the choice of the evaluation metric influences which algorithm can be considered the best.

Thus, with respect to the entire dissertation and RQ III. of the dissertation, the conclusion from this chapter is that the applied state of the art machine learning algorithms by no means generally lead to better results than the classical approaches–related to the conducted experiments and underlying data sets.

5.5 References

Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo (2018): A note on the validity of cross-validation for evaluating autoregressive time series prediction. In *Computational Statistics & Data Analysis* 120, pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

Croston, J. D. (1972): Forecasting and Stock Control for Intermittent Demands. In *Journal of the Operational Research Society* 23 (3), pp. 289–303. DOI: 10.1057/jors.1972.50.

Grimm, Florian; Kiefer, Daniel; Bauer, Markus (2021): Univariate Time Series Forecasting by Investigating Intermittence and Demand Individually. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 143–151.

Hyndman, Rob J.; Koehler, Anne B. (2006): Another look at measures of forecast accuracy. In *International Journal of Forecasting* 22 (4), pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Kiefer, Daniel; Bauer, Markus; Grimm, Florian (2021a): Univariate Time Series Forecasting: Machine Learning Prediction of the Best Suitable Forecast Model Based on Time Series Characteristics. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 152–162.

Kiefer, Daniel; Grimm, Florian; Bauer, Markus; van Dinther, Clemens (2021b): Demand Forecasting Intermittent and Lumpy Time Series: Comparing Statistical, Machine Learning and Deep Learning Methods. In : Hawaii International Conference on System Sciences 2021. Honolulu, HI: University of Hawai'i at Manoa, Hamilton Library, p. 1425.

Kourentzes, Nikolaos (2013): Intermittent demand forecasts with neural networks. In *International Journal of Production Economics* 143 (1), pp. 198–206. DOI: 10.1016/j.ijpe.2013.01.009.

Levy, Yair; Ellis, Timothy J. (2006): A systems approach to conduct an effective literature review in support of information systems research. In *Informing Science* 9.

Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios (2022): The M5 competition: Background, organization, and implementation. In *International Journal of Forecasting* 38 (4), pp. 1325–1336. DOI: 10.1016/j.ijforecast.2021.07.007.

Martin, Dominik; Spitzer, Philipp; Kühl, Niklas (2020): A New Metric for Lumpy and Intermittent Demand Forecasts: Stock-keeping-oriented Prediction Error Costs.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. et al. (2011): Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* 12, pp. 2825–2830.

Syntetos, A. A.; Boylan, J. E.; Croston, J. D. (2005): On the categorization of demand patterns. In *Journal of the Operational Research Society* 56 (5), pp. 495–503. DOI: 10.1057/palgrave.jors.2601841.

Syntetos, A.A; Boylan, J.E (2001): On the bias of intermittent demand estimates. In *International Journal of Production Economics* 71 (1-3), pp. 457–466. DOI: 10.1016/S0925-5273(00)00143-2.

Teunter, Ruud H.; Syntetos, Aris A.; Zied Babai, M. (2011): Intermittent demand: Linking forecasting to inventory obsolescence. In *European Journal of Operational Research* 214 (3), pp. 606–615. DOI: 10.1016/j.ejor.2011.05.018.

Turkmen, Ali Caner; Wang, Yuyang; Januschowski, Tim (2019): Intermittent Demand Forecasting with Deep Renewal Processes.

Webster, Jane; Watson, Richard T. (2002): Analyzing the past to prepare for the future: Writing a literature review. In *MIS quarterly*, pp. xiii–xxiii.

Willemain, Thomas R.; Smart, Charles N.; Shockor, Joseph H.; DeSautels, Philip A. (1994): Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. In *International Journal of Forecasting* 10 (4), pp. 529–538. DOI: 10.1016/0169-2070(94)90021-3.

Williams, T. M. (1984): Stock Control with Sporadic and Slow-Moving Demand. In *Journal of the Operational Research Society* 35 (10), pp. 939–948. DOI: 10.1057/jors.1984.185.

6 A New Approach in Hierarchical Demand Forecasting

[*This chapter corresponds to an article submitted to the Journal of Forecasting (JoF). It is currently in second revision status in the acceptance process at JoF:* Bauer, Markus; van Dinther, Clemens; Kiefer, Daniel; Grimm, Florian (2023): Forecasting Intermittent Demand with no given External Hierarchy: An Aggregation-Disaggregation Approach based on Clustering of Time Series Characteristics Similarity. – *further referred to as* Bauer et al. 2023]

Forecasting Intermittent Demand with no given external Hierarchy: An Aggregation-Disaggregation Approach based on Clustering of Time Series Characteristics Similarity

Forecasting intermittent demand time series with periods of zero demands is notoriously difficult. While research shows that forecasting of such time series can be improved by aggregation-disaggregation in cases where a hierarchy between the time series can be exploited (such as an article hierarchy in product demand forecasting), there are practical scenarios where no hierarchy is available. In this empirical study, we introduce an approach to improve intermittent demand forecasting by aggregation-disaggregation of time series based only on time series characteristics similarities: similarity-based time series forecasting (STSF). The results show a significant improvement of forecasting quality in the context of the publicly available M5 and Kaggle store sales forecasting competition data set. Additionally, the time series that benefit from the approach can be accurately predicted by their specific characteristics.

6.1 Introduction and Problem Definition

Time series (TS) are sequences of observations of variables with a temporal discrete order. They are observed in many disciplines of academia (natural sciences, economics, engineering, and many more) as well as in practice (society, business, and more). To all these disciplines, forecasting future, yet unknown values of such variables is a goal of academic research.

In this study, we focus on the case of intermittent time series (iTS). iTS are characterized by considerable sequences of observations of the variable that equal zero. Throughout this article, we employ the example of spare part demand forecasting – relevant for example in inventory planning. Spare parts are often required only occasionally, in periods without any demand. Research has identified forecasting of iTS as difficult because traditional forecasting algorithms apply assumptions that conflict with zero demand periods.

Research has evolved various approaches to forecast intermittent time series (see Section 6.2). Common approaches work by aggregating time series of objects that are known to correlate. For example, in the automobile spare parts industry, car bumper demands for the same type of car but of different colors are likely to correlate. Most known approaches require an external structure or hierarchy of the objects, such as a spare parts family categorization. However, in many cases no hierarchies exist, or they are not available for forecasting.

In this study, we propose a new approach for forecasting of iTS by aggregation-disaggregation with no additional object hierarchy but the information that we can extract from the time series themselves. We call this approach similarity-based time series forecasting (STSF). We use the following research questions to structure the study and the empirical experiment design:

RQ1. Can an aggregation-disaggregation approach based on similar TS characteristics improve forecasting accuracy? If yes, by how much?

RQ2. Can special types of TS be identified that benefit most from the approach? By which characteristics are they best defined and how much is the improvement for these TS?

RQ3. Can these TS be identified in advance and does the approach yield similar accuracy results when only these TS are used?

RQ4. What other parameters of the proposed STSF approach influence the approach and how?

6.2 The State-of-the-Art in Intermittent Demand Forecasting

This section provides a brief overview of the current state-of-the-art of classical approaches for intermittent time series forecasting, aggregation approaches, and introduces the basics of intermittent time series categorization and clustering.

We introduce the following notation. We assume time series as sequences of observations of a variable $x_{t,i}$ with time steps $t \in T = \{0,1,\dots,t_{max}\}$ and objects (e.g., products) $i \in I = \{0,1,\dots\}$. All given observations of this variable $x_{t,i}$ are X. The goal is to find an approximation $\hat{x}_{t+1,i}$ for $x_{t+1,i}$ defined by the unknown function $x_{t+1,i} = f(X,\Phi) + \epsilon_{t+1,i}$. We define Φ as a set of observations of stochastic influences $\epsilon_{t,i}$. The function $f(.)$ may depend on interdependencies of x and Φ. The stochastic random variable $\epsilon_{t+1,i}$ may depend on X and Φ.

As there exist a multitude of approaches to time series and demand forecasting, we will not discuss the general landscape of algorithms in detail. The most relevant approaches shall only be mentioned here, by no means exhaustive, i.e., classical approaches (ARMA, ARCH, Holt-Winter's) and machine learning (support vector machines, decision trees, artificial neural networks). An overview of time series forecasting can be found in Boylan and Syntetos' extensive overview (Boylan and Syntetos 2021).

In the further course of the article, we will use the M5 forecasting competition data set as well as the Kaggle store sales data set for our empirical study. They are recent, publicly available data sets designed to be a challenging data set for forecasting of demand time series that also includes a considerable degree of intermittent time series (see also Section 6.3.1). Studies based on the competing approaches show a superiority of decision tree based algorithms, especially the Light Gradient Boosted Machine (LGBM) as an advancement of the Gradient Boosting Decision Trees (GBDT) (Ke et al. 2017; Makridakis et al. 2022).

6.2.1 Intermittent time series categorization approaches

While the term intermittence is generally applied to time series with considerable sequences of zero values (e.g., demands), there is no generally accepted definition of a metric that quantifies the degree of intermittence or yields a categorization of intermittent and non-intermittent time series. Williams introduces the definition of intermittence by $1/\bar{L}\lambda$, where $\bar{L}$ denotes the mean lead time (time between two demands greater zero) and λ denotes the mean number of demands per time interval, assuming a Poisson arrival process of demands. Large values for $1/\overline{L}\lambda$ are interpreted as high intermittence (also referred to as sporadicity). Furthermore, Williams defines the term "lumpiness" as $C^2/\overline{L}\lambda$, where C^2 is the coefficient of variation of the demand height, demand frequency, and inter-demand intervals. The study provides a categorization of products into four categories (by the values of $1/\overline{L}\lambda$ and $C^2/\overline{L}\lambda$ (Williams 1984).

Based on the study of Williams, Eaves describes an improved approach using the transaction variability, demand size variability and lead time variability instead of transaction variability only (c.f. above) (Eaves 2002). Syntetos et al. further refined this approach and utilize the average inter-demand interval p and the coefficient of variation of demand (c.f. Williams) CV^2 as metrics for iTS categorization (Syntetos et al. 2005).

We introduce the categorization approaches not primarily to classify TS along hard boundaries. Instead, we will adopt the metrics by Syntetos et al. as a form of clustering representation. We introduce TS clustering and representations in Section 6.2.4.

6.2.2 An overview of classical models for intermittent demand forecasting

Croston constitutes a major stream of research in iTS forecasting, especially in the context of spare parts. The first approach is based on exponential smoothing of two aspects of the time series parallel: (1) the interval length of zero demands and (2) the demand size in periods of demand greater zero (Croston 1972). Willemain et al. show on two sample data sets that Croston's approach overperforms single exponential smoothing (Willemain et al. 1994). Syntetos and Boylan find that Croston's approach is systematically biased. Therefore, they propose an adjustment to Croston's approach that unbiases the results through a correction factor (SBA algorithm). In a subsequent study, Teunter et al. propose to further refine the approach. They propose to predict the probability of a non-zero demand instead of inter-demand interval length. By this improvement, their so-called TSB approach can also be updated in periods of zero-demands – in contrast to Croston's approach that only updates inter-demand intervals when demands are observed. Both studies show that the SBA approach (respectively TSB) achieve better results in the studies' simulation experiments, compared with Croston's approach, standard simple moving average and single exponential smoothing (Syntetos and Boylan 2005; Teunter et al. 2011). Babai et al. refined the SBA approach especially for cases of sudden obsolescence (Babai et al. 2019). Doszyń proposes the SESAP (simple exponential smoothing for analogous sub-periods) approach for seasonal products. They find that a combination of TSB for general iTS and SESAP for seasonal iTS performs best. They also conclude that Croston's approach and SBA are outperformed by single exponential smoothing – in contrast to the previously mentioned studies (Doszyń 2019).

Willemain et al. propose a bootstrapping approach, in contrast to the approaches derived from Croston's. In their approach, they sample datapoints from the given data set distribution to forecast future demands. They find their approach outperforms Croston's approach in their study (Willemain et al. 2004). Syntetos et al. compare parametric approaches (including SBA) with bootstrapping and conclude that the bootstrapping approach yields better forecasting results in some situations; however, they find the approach more complex than SBA (Syntetos et al. 2015).

Gutierrez et al. propose the application of artificial neural networks (ANN) for iTS forecasting with good results (Gutierrez et al. 2008). Kourentzes proposes an approach based on Croston's basic approach; however, they apply ANN instead of exponential smoothing for forecasting (Kourentzes 2013). Fu et al. propose a hybrid approach of SBA and ANN based on the TS characteristics. Babai et al. compare Single Exponential Smoothing (SES), Croston's approach, SBA, Willemain's bootstrapping and Gutierrez et al. SBA inspired ANN approach in a case study on spare part demand forecasting. They report the best performance by Gutierrez et al. approach with additional Bayesian regularization (Babai et al. 2020). Kiefer et al. study Croston's approach in comparison with standard machine learning approaches. Based on their experimental design in forecasting sales of a retailer, they conclude that machine learning models (LSTM long short-term memory ANN, SVR support vector machines, XGBoost, and random forest) perform better than Croston's approach. They apply the mean average error as evaluation metric. They acknowledge that the machine learning approaches are also better than Croston's approach for intermittent TS (Kiefer et al. 2021).

Summarizing, we observe an evolution of algorithms based on Croston's approach (SBA, TSB and further). Their successors address and improve specific issues. Due to the specialization of the successors and different data sets applied, no study concludes that one approach is superior to all others in every situation. Hence, we conclude in this section that specialized algorithms exist to forecast intermittent TS which perform better than general forecasting algorithms – when the scope is on iTS.

6.2.3 Aggregation approaches for intermittent time series forecasting

To further improve forecasting of iTS, academia has studied the application of aggregation approaches. The literature describes two basic types of approaches that apply aggregation to improve iTS forecasting: (1) hierarchical aggregation of time series along a given hierarchy, also regularly referred to as top-down bottom-up (TD/BU) approach or hierarchical forecasting (HF); and, (2) temporal aggregation (TA) of periods of time of individual time series. To differentiate HF and TA, HF is also referred to as contemporaneous aggregation.

Top-down (TD) approaches aim at condensing the sparse information contained in iTS by aggregation of objects (e.g., products) along the hierarchy while keeping the granularity of time periods unchanged. TD then forecasts the aggregated time series and subsequently disaggregates the prediction down to the original level in the hierarchy. The opposite approach is bottom-up (BU) forecasting. The approach is motivated by the intuition that through the aggregation, intermittence is reduced, and thus, satisfying the assumptions of most common forecasting algorithms, makes these algorithms more suitable.

Several recent studies apply TD/BU approaches for iTS forecasting. Viswanathan et al. apply TD forecasting for highly variable iTS and BU based on Croston's approach for less variable iTS to improve overall forecasting accuracy (Viswanathan et al. 2007). In another study, Li and Lim propose a middle out approach which combines TD and BU (Li and Lim 2018). Abolghasemi et al. compare classical and machine learning approaches in iTS TD/BU scenarios. They conclude that the decision tree based algorithm XGboost (Chen and Guestrin 2016) overperforms other approaches in their study (Abolghasemi et al. 2019). In all studies, the authors show benefits of TD/BU approaches when there is an applicable object hierarchy for aggregation.

Temporal aggregation approaches follow a similar intuition as TD/BU. However, aggregation is performed on a per time series level along the timeline, i.e., aggregating buckets of time of each TS.

Nikolopoulos et al. provide a framework for TA with the aggregate-disaggregate intermittent demand approach (ADIDA). In their empirical study they show improvements in forecast applying TA and they assume the existence of an optimal aggregation level (Nikolopoulos et al. 2011). Rostami-Tabar et al. provide theoretical analytical foundation to the assumption that TA can lead to improved forecasting results when exponential smoothing is used as forecasting algorithm (Rostami-Tabar et al. 2013). The ADIDA approach is refined by the multiple aggregation prediction algorithm (MAPA), where forecasts on multiple aggregation levels are combined into one forecast (Kourentzes et al. 2014).

A comparative study between TD/BU and temporal aggregation was conducted by Lütkepohl, but with a limited scope on the VARMA forecasting model, a special type of ARMA (Lütkepohl 2010).

In this section, we conclude that aggregation is an appropriate way to improve forecasting of iTS in addition to the specialized forecasting algorithms in Section 6.2.2. Moreover, TD/BU and TA make it possible to apply forecasting algorithms that are not particularly specialized to iTS forecasting. Thus, a broader choice of forecasting algorithms becomes available to iTS forecasting.

6.2.4 Basics in the state-of-the-art in time series clustering

At this point, we provide an overview of TS clustering. One of our assumptions, as introduced in Section 6.1, is that we apply hierarchical forecasting in the absence of an externally given hierarchy. In the further course of this study, we will apply TS clustering as a substitute for this external hierarchy by clustering TS based only on their respective characteristics.

Research has evolved a broad variety of general clustering algorithms. In this study, we will not review basic clustering algorithms, but refer the reader to common surveys. We point out the k-means algorithm, which is a frequently used partitioning algorithm (Saxena et al. 2017).

More relevant to this study are time series specific clustering approaches. Generally, TS clustering follows the following steps: (1) creating a representation for the TS, (2) calculating a distance (or similarity) measure between the TS representations, (3) clustering the TS, and (4) subsequently evaluating the cluster quality (potentially followed by a reiteration of the process). Aghabozorgi et al. describe the broad variety of approaches in each step and their combination. (Aghabozorgi et al. 2015).

For this study, we concentrate on time series representation approaches. Several studies propose frameworks and packages to extract numerous features from time series. Fulcher and Jones have proposed a framework called *hctsa* of feature extraction algorithms for time series representation, stating more than 8,000 algorithms (Fulcher et al. 2013; Fulcher and Jones 2014, 2017). Hyndman et al. provide the *tsfeature* package for R that implements several TS feature extraction approaches described in a previous study in 2015 (Hyndman et al. 2015; Hyndman et al. 2022). Christ et al. describe a similar framework, called *TSFRESH*, incorporating time series feature extraction methods as well as an algorithm to select the most relevant of the features by hypothesis testing. (Christ et al. 2018). However, none of the before mentioned frameworks is specialized on iTS feature extraction. Common approaches like Fourier transformation, wavelet transformation, or dynamic time warping (amongst others) are not found to be especially suitable for iTS representation.

Montero-Manso et al. describe the *FFORMA* approach that builds upon the *tsfeature* package. They apply a meta-learner to infer from TS characteristics which forecasting algorithms are best suited for forecasting of these TS. The meta-learner is used to apply weights to the forecasts of different forecasting algorithms to calculate an ensemble prediction (Montero-Manso et al. 2020). A comparable approach is applied in this study, also referred to as meta-learner.

From this section we conclude that TS can be clustered based on their characteristics when an appropriate representation is used. Analogously to the representation approaches in the form of feature extraction described above, we will use the metrics by Syntetos et al. (see Section 6.2.1) as representation. In addition, we propose our own representation metric in Section 6.3.2. In our terminology, the representation will be a concrete metric that represents the TS characteristics.

6.2.5 Aggregation approaches for iTS incorporating clustering

In the previous sections, we summarized classical approaches to iTS forecasting and the basics of aggregation approaches and TS clustering. As the STSF approach that we propose combines aggregate forecasting with TS clustering, we reviewed literature that also addresses this combination. Table 28 provides an overview of the reviewed studies and a comparison to our STSF approach. We differentiate the studies by the following three criteria: (1) focus on intermittent TS, (2) clustering only based on time series characteristics (without additional external hierarchy), and (3) a processing sequence of clustering, aggregation, forecasting on aggregate level, and subsequent disaggregation back to the original level of aggregation.

Study	(1) Focus on iTS	(2) clustering based on TS characteristics	(3) comparable processing sequence (see text below)
Zotteri et al. 2005	no, TS in general	no	yes
Dantas and Cyrino Oliveira 2018	no, TS in general	yes	no (clustering after forecasting)
Pang et al. 2018	no, TS in general	yes	yes
Laurinec et al. 2019	no, TS in general	yes	yes, however considerably different techniques
Proposed STSF approach	yes	yes	yes

Table 28: Comparison of iTS clustering-aggregation studies with the proposed STSF approach

Zotteri et al. propose an approach in a forecasting scenario of sales and products. Their study applies a clustering algorithm to cluster stores with similar penetration rates (i.e., conversion rates) to then aggregate sales of these clusters. This study is characterized by clustering before aggregation and forecasting and by an additional external characterization criterion (store sales) – unlike STSF (see Table 28 (1) and (2)). The study compares this approach to TD and BU and concludes that best results are achieved for TD forecasting and clustering forecasting (Zotteri et al. 2005). In a study by Dantas and Cyrino Oliveira, the authors suggest an approach where exponential smoothing forecasting results on bottom levels are aggregated through bagging and clustering (partition around medoids). In contrast to the STSF, the clustering is applied after forecasting and based on variance of the forecasts, see Table 28 (3) (Dantas and Cyrino Oliveira 2018). A study by Pang et al. applies a TD/BU approach for electricity forecasting, where clustering of time series is performed depending on time series correlations. They do not focus on iTS, therefore they differ from STSF in point (1) (Pang et al. 2018). Laurinec et al. present an approach similar to our study where they improve forecasting by cluster time series by their linear regression coefficients, yielding a result that resembles a spectrogram of the TS. However, their approach does not focus on intermittent time series, hence they apply regression coefficients as representation and an ensemble bootstrapping forecasting algorithm (see Table 28 (3)). As opposed to our study, they conclude that their approach overperforms the state-of-the-art approach only for non-intermittent TS (Laurinec and Lucká 2018; Laurinec et al. 2019).

Our study differentiates from the previously mentioned literature in at least one point (see Table 28). To our best knowledge, no literature exists that applies a comparable approach to this study.

6.3 Methodology and Approach

The following Section describes the proposed approach and the experiment design that we apply to demonstrate empirically the performance of the approach. Figure 17 demonstrates the two-stage design: (1) a basic pipeline acts as foundation that can be repeatedly executed with different experiment settings to be compared to a baseline approach. On a (2) superior level, the experiment level, different experiment settings are systematically tested and compared. Each experiment is always executed on the basic pipeline level with different random seeds for randomization. A meta-learner classifier is applied on experiment level, based on the experiment results, to generate an "all-stars" forecast out of the best predictions per TS.

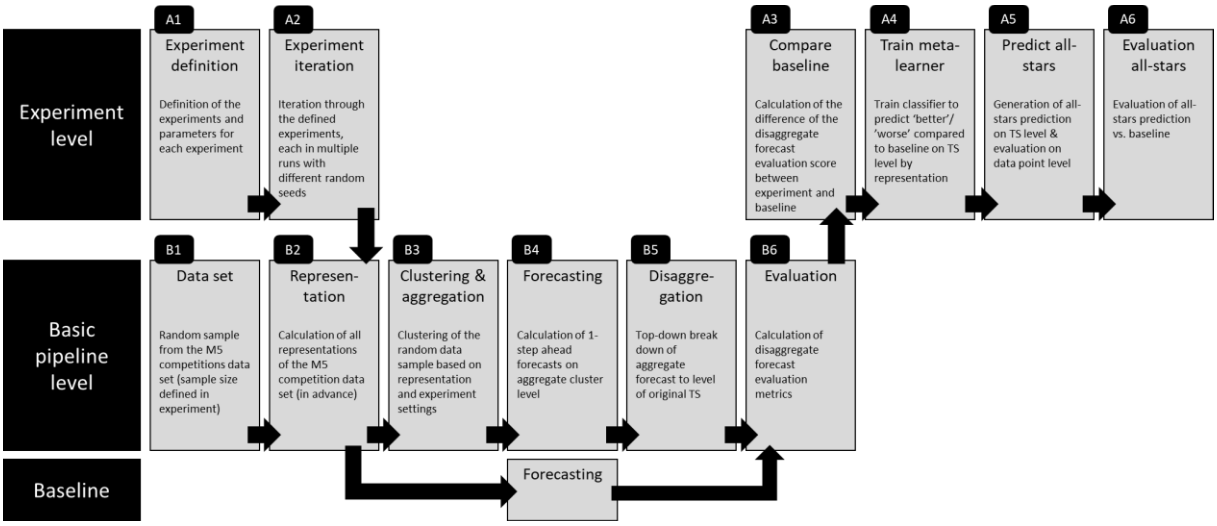

Experiment level
A1
Experiment definition
Definition of the experiments and parameters for each experiment
A2
Experiment iteration
Iteration through the defined experiments, each in multiple runs with different random seeds
A3
Compare baseline
Calculation of the difference of the disaggregate forecast evaluation score between experiment and baseline
A4
Train meta-learner
Train classifier to predict 'better'/ 'worse' compared to baseline on TS level by representation
A5
Predict all-stars
Generation of all-stars prediction on TS level & evaluation on data point level
A6
Evaluation all-stars
Evaluation of all-stars prediction vs. baseline
Basic pipeline level
B1
Data set
Random sample from the M5 competitions data set (sample size defined in experiment)
B2
Represen-tation
Calculation of all representations of the M5 competition data set (in advance)
B3
Clustering & aggregation
Clustering of the random data sample based on representation and experiment settings
B4
Forecasting
Calculation of 1-step ahead forecasts on aggregate cluster level
B5
Disaggre-gation
Top-down break down of aggregate forecast to level of original TS
B6
Evaluation
Calculation of disaggregate forecast evaluation metrics
Baseline
Forecasting

Figure 17: Overview of the experiment and basic pipeline design.

6.3.1 The studied data sets

In this study, we apply the proposed approach to two data sets. Both are publicly available for reproducibility.

The M5 data set was published in 2020 on Kaggle for download and contains Walmart sales [1]. It consists of 30,490 time series on the lowest level of aggregation and 1,941 time steps per time series (Makridakis and Spiliotis 2021).

The Kaggle store sales competitions is another time series forecasting competition, published in 2021. It consists of 1.782 time series in total, representing sales of different goods at retailing stores as well [2].

Table 29 shows an overview of the data sets' time series categorization by Syntetos et al. We observe that the majority of time series in the M5 data set is intermittent or lumpy (91%) as well as a considerable share of intermittent or lumpy time series in the Kaggle store sales data set (23%).

TS category	p	CV2	No. of time series per category and data set	
			M5 data set	Kaggle store sales
Smooth	≤ 1.32	≤ 0.49	1,908 TS (6%)	945 TS (53%)
Erratic	≤ 1.32	> 0.49	868 TS (3%)	427 TS (24%)
Intermittent	> 1.32	≤ 0.49	22,150 TS (73%)	226 TS (13%)
Lumpy	> 1.32	> 0.49	5,564 TS (18%)	184 (10%)

Table 29: Categorization of TS following the scheme of Syntetos-Boylan in the two studied data sets with number of time series in each category (absolute and percentage of total).

This study does not aim to develop a competitive forecast particularly for the two competitions. Instead, the data sets are chosen, because they are available to the public, they are extensively researched and provide a broad range of time series characteristics. Hence, the results of our study can be compared and more importantly can be reproduced by any interested researcher.

[1] Source: https://www.kaggle.com/competitions/m5-forecasting-accuracy

[2] Source: https://www.kaggle.com/competitions/store-sales-time-series-forecasting

However, many input information that are contained in the data sets are deliberately not considered in this study, e.g., the supplementary sales prices and calendar information (events, holidays, day of week, month, etc.). To comply with the initial premise of the study (i.e., “no external hierarchy is known to support the prediction”), we also omitted store and product category information. What remains are exclusively univariate time series.

6.3.2 Machine learning pipeline

The basic pipeline consists of six steps: B1 to B6.

In the first step (B1), a random sample is selected from the data set. The sample size depends on the experiment settings, as well as the random seed used for the random sampling.

In a second step (B2), the pipeline calculates representations for each TS. In addition to the representation criteria proposed by Syntetos et al., we propose a second representation metric that we found to perform comparably well: PD-statistics. It consists of two components: periodicity P and non-zero demand D. For each component (per TS), we calculate the three statistics: arithmetic mean (*_mean*), standard deviation (*_stdev*), and the slope of the ordinary least squares regression (*_trend*), see Table 30 for details.

For a better understanding of the PD-statistics, we introduce the following notation to amend to the notations in Section 6.2. Let D_i be the sequence of non-zero observations of $x_{t,i}$ of an object i for the D-statistic. Let further be P_i be a sequence of the number of non-zero observations of $x_{t,i}$ in rolling windows $w_{k,i}$ of the observations $x_{t,i}$ of the length h of an object i.

P-statistic: Periodicity	**D-statistics: Non-zero demands**
Number of periods with non-zero demands within a rolling time frame of h consecutive time steps	Values of demands in periods with non-zero demand
Be $w_{k,i} = n(x_{t,i} > 0\ t \in \{k, k+1, \dots, k+h-1\}$ with $k \in \{0,1,\dots,t_{max}-h+1\}$ rolling windows where $n(x_{t,i})$ denotes the number of observations of $x_{t,i}$ that fulfil the condition (i.e. cardinality). Be $P_i = \left\{w_{k,i}, k \in \{0,1,\dots,t_{max}-h+1\}\right\}$ the sequence of rolling windows $w_{k,i}$ of the number of non-zero observations of $x_{t,i}$ and object i .	Be $d_{t,i} = \left\{x_{t,i} > 0\right\}, t \in T$ the non-zero demand of object i at time t and $D_i = \left\{d_{t,i}, t \in T\right\}$ the sequence of non-zero demands of object i.
Arithmetic mean of the elements in P_i (**“p_mean”**): $p_{i,mean} = \frac{1}{n(P_i)} \sum_{p_i \in P_i} p_i$	Arithmetic mean of the elements in D_i (**“d_mean”**): $d_{i,mean} = \frac{1}{n(D_i)} \sum_{d_i \in D_i} d_i$
Standard deviation of the elements in P_i (**“p_stdev”**): $p_{i,stdev} = \frac{1}{n(P_i)} \sum_{p_i \in P_i} \lvert p_i - p_{i,mean} \rvert$	Standard deviation of the elements in P_i (**“d_stdev”**): $d_{i,stdev} = \frac{1}{n(D_i)} \sum_{d_i \in D_i} \lvert d_i - d_{i,mean} \rvert$
Slope of the ordinary least squares regression of the elements of P_i (**“p_trend”**): $p_{i,trend} = \frac{\sum_{k \in \{0,1,\dots,n(P_i)-1\}} (k-\overline{k})(w_{k,i} - p_{i,mean})}{\sum_{k \in \{0,1,\dots,n(P_i)-1\}} (k-\overline{k})^2}$ where $\overline{k}$ denotes the arithmetic mean of the values of $k \in \{0,1,\dots,n(P_i)-1\}$	Slope of the ordinary least squares regression of the elements of D_i (**“d_trend”**): $d_{i,trend} = \frac{\sum_{k \in \{0,1,\dots,n(D_i)-1\}} (k-\overline{k})(-d_{i,mean})}{\sum_{k \in \{0,1,\dots,n(D_i)-1\}} (k-\overline{k})^2}$ where $\overline{k}$ denotes the arithmetic mean of the values of $k \in \{0,1,\dots,n(P_i)-1\}$

Table 30: Calculation of the two components and their statistics of the PD-statistics

The PD-statistics aim to provide an additional approach to capture periodicity of non-zero demands and the value of non-zero demands alongside the Syntetos-Boylan classification. By using the *mean*, *stdev* and *trend* statistics, we aim to not only capture only static characteristics of the TS but also variations over time. The effectiveness of the PD-statistics compared to the Syntetos-Boylan approach is discussed in Sections 6.4 and 6.5.

We chose the P-statistic rolling window size $h = 7$ for all experiments to cover exactly one week, as both data sets provide daily observations. This parameter is subject to further optimization in future research, as we also discuss in Section 6.6.

In step three (B3), the pipeline clusters the TS. As a clustering algorithm, we apply the k-means clustering algorithm. The number of clusters is given in the experiment settings. As input, the cluster algorithm takes either the Syntetos-Boylan representations or the PD-statistics. All values are min-max (linear normalization to domain zero to one) scaled before clustering. Time series in identical clusters are then aggregated (summed up) for each time step.

In step four (B4), the pipeline trains a forecasting model on the aggregated time series and calculates a one-step ahead forecast for each time step of the aggregated TS. As forecasting algorithm, we apply the LGBM algorithm as implemented in the scikit-learn package (*HistGradientBoostingRegressor*). We use only the default hyperparameter setting from the package and do not apply hyperparameter optimization. Additionally, we applied the TSB approach based on Croston's approach (referred to as *Croston_TSB*) proposed by Teunter et al. at one point of the study. The results cannot compete with the results of the LGBM algorithm, however we find it a good comparison for the reader to relate to a well-known classical iTS forecasting algorithm (Teunter et al. 2011).

The forecasting model is trained with the training data of all aggregated time series simultaneously. No additional exogenous data as explanatory variables is provided. In order to split training and test data, we apply k-fold cross validation (with three folds). We refer the reader to the following articles for those not familiar to the k-fold approach in conjunction with TS cross validation (Bergmeir and Benítez 2012; Bergmeir et al. 2018).

The three-fold cross validation ensures that the forecasting algorithm is never trained on data that it is supposed to predict. During the training and forecasting, the model is iteratively trained on two thirds of the data set and predicts the outcome of the remaining third part of the data without knowledge of the time series of the third part. The model is then reset to the initial state, before it is trained on a different part of the data int the next iteration (i.e., "cold start", see Figure 18). After three iterations, the forecasting algorithm has generated predictions for all three thirds of the time series, whilst having learned always the other two thirds in each iteration. The same principle is also applied in the subsequent steps B5, A4 and A5. We emphasize that no hyperparameter tuning is done (i.e., the default parameters of the scikit-learn package are always used), so we do not draw an additional validation sample from the data.

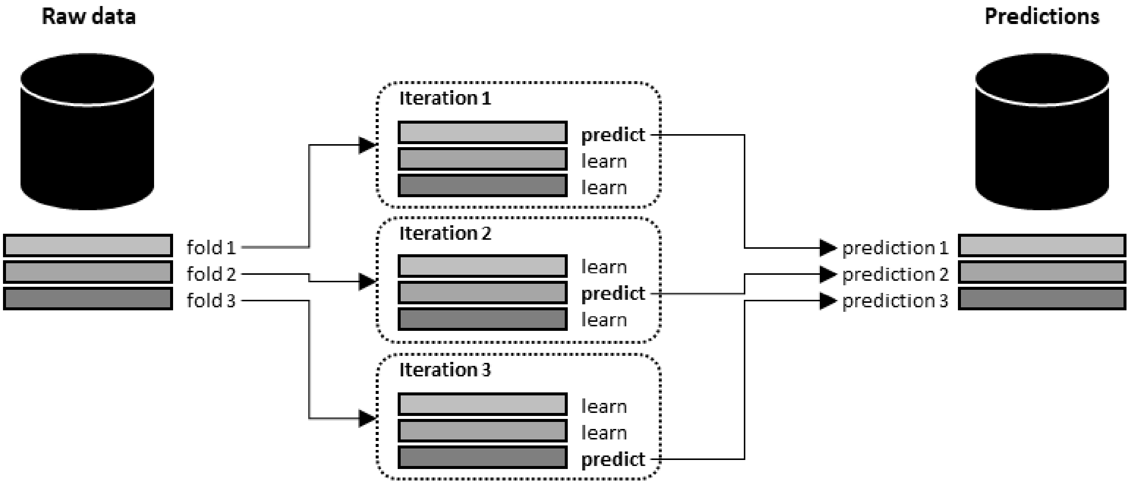


Figure 18: Scheme of the three fold k-fold cross-validation approach. The input data is split into three folds. The k-fold principle ensures that no information about the forecasted time series is leaked to the forecasting model.

In step five (B5) of the basic pipeline, we apply another LGBM model to disaggregate the aggregate forecast. The model is trained to learn the relationship between forecasts on aggregated level and disaggregated actual values on disaggregate level. Technically speaking, the model is trained with aggregate forecast values and the TS identification number as input and disaggregate actual values as target variable. The trained model then produces disaggregation forecasts, again using k-fold with three folds to separate training and test data.

In the final step (B6) of the basic pipeline, the disaggregate forecasts are compared to the actual time series values using the following evaluation metrics: coefficient of determination (R^2), rooted mean squared error (RMSE), mean absolute error (MAE) equal to mean absolute deviation (MAD), MAD by mean ration (MADmeanRatio) (Kolassa and Schütz 2007), mean absolute scaled error (MASE), and rooted mean squared scaled error (RMSSE). The pipeline applies the metrics per TS and then calculates the weighted mean, weighted by the TS share of total values.

To compare our approach with a state-of-the-art approach, we also produce one forecast without aggregation-disaggregation, which we refer to as “baseline”. We employ the same experiment settings and randomization; however, the data is not clustered, and the forecasting model is trained and produces a forecast on the original time series level. The baseline is implemented as clustering approach with cluster size equal to exactly one time series per cluster. The pipeline skips the steps B3 and B5.

The baseline predictions are forecasted by an LGBM model, which is newly trained on the not-aggregated input time series. Thereby, there is no systematic difference induced through the forecasting algorithm between the baseline and the experiments. By this design choice we aim to keep the approach agnostic to the forecasting algorithm applied.

The entire pipeline is implemented using scikit-learn for Python 3.8 (Pedregosa et al. 2011).

6.3.3 Experiment design

The basic pipeline as we describe before, can be called with different settings and randomization. Table 31 shows the relevant settings that we used for this study (A1). From the possible experiment values, we generated all possible combinations of experiments $s \in S = \{0,1,\ldots,s_{max}\}$ (cross product of the attributes and values) and executed the basic pipeline (A2). In total, we have defined 11 experiments (5 cluster sizes × 2 representation approaches + 1 baseline) and 132 experiment runs (11 experiments × 12 randomizations: 12 arbitrary random seeds with values 900 to 911).

Experiment	Sample size	Representation	No. of clusters
0	500 TS	PD-statistics	3
1			5
2			10
3			100
4			200
5		Syntetos-Boylan	3
6			5
7			10
8			100
9			200
10 (Baseline)		w/o	no clustering

Table 31: Overview of experiment design settings.

The result of all runs is a comprehensive database of aggregate and disaggregate forecasts, the corresponding baseline forecasts, and their evaluation results in step A3. We calculate the difference of evaluation results of all evaluations compared to their respective baseline on disaggregate level:

$$delta_{i,s,evaluation\ metric} = evaluation\ result(baseline_s) - evaluation\ result(experiment_s) \tag{1}$$

$$delta\_perc_{i,s,evaluation\ metric} = delta_{i,s,evaluation\ metric} / evaluation\ result(baseline_s) \tag{2}$$

From the deltas per TS and experiment, we can decide whether our approach or the state-of-the-art baseline approach yields better results in the particular instance. We label all experiment-TS combinations with “*stsf*” for delta values larger or equal zero and “*baseline*” for delta smaller zero. This means that time series are labeled “*stsf*” if the evaluation metric indicates lower or equal forecasting errors and “baseline” in the case that the evaluation metric indicates higher forecasting errors. This is the case for all evaluation metrics, except for the R² metric. Here, delta values lower or equal.

$$label_{i,s} = \begin{cases} 'stsf', & delta_{i,s,evaluation\ metric} \geq 0 \\ 'baseline', & delta_{i,s,evaluation\ metric} < 0 \end{cases} \tag{3}$$[3]

At this point, we can systematically run experiments with different settings and compare them to the baseline on a per-TS level. Therefore, we can study whether our proposed approach yields better forecasting results, which settings of the approach perform best and which TS benefit how much from the approach. To deduce whether and which TS characteristics account for better forecasting results, we introduce the last steps of the experiment design.

In step A4, we train another LGBM model to classify the labels of the time series. We call this model the meta-learner model. As additional input, we provide the model with the representation data of the time series (both Syntetos-Boylan representation and PD-statistics). In A5, the model then predicts the labels for the time series, again in a three folded k-fold cross-validation approach. Finally, based on the predicted label of the meta-learner, we have a prediction, whether the forecast of our approach or the baseline’s forecast will probably be more accurate. From this, we can construct an all-stars forecast, consisting of either our approach’s results or the baseline’s results.

To provide a evident comparison with a simple state-of-the-art approach, we also implemented the approach Teunter et al. as an additional baseline approach. We will refer to the results as *Croston_TSB* in the remainder of the study. We apply time series cross validation as implemented in the *TimeSeriesSplit*

[3] For all evaluation metrics except the R² metric. For the R² metric, labels are assigned reverse.

function of the scikit-learn package. We provide at least 50% of each univariate time series to the algorithm for training and the next consecutive time step as test.

6.3.4 Remarks on the STSF pipeline design

We point out that the STSF pipeline approach is generally agnostic to the standard algorithms employed in most steps. The essence of the STSF approach is the design depicted in Figure 17, not the particular algorithm. We therefore encourage future research to replace especially the PD-statistics and Syntetos Boyle classification as representation in (B2), the k-means algorithm in (B3) or the LGBM algorithm in steps (B4), (B5) and (A4/5) to further optimize the approach.

The influence of the LGBM algorithm as the forecasting algorithm is minimized by the baseline comparison. Both the experiments as well as the baseline apply the LGBM algorithm in the same way in steps (B4) and (B5).

Further, the pipeline presented in the previous sections implies several parameters and design choices (e.g., the selection of the clustering algorithm or forecasting algorithm, the P-statistics window size h, etc.). Important parameters, such as the number of clusters, were included in the experiment design with different values for direct comparison. However, due to limited computational power, we did not consider all parameters, but made reasonable decisions for initial values for some of them. We note in Section 6.6 that the comprehensive research of optimal parameter values is subject to future research on the STSF pipeline.

6.4 Results

This section describes the results of the approach in the experiments. We will show that the basic pipeline cannot improve results overall, and how the meta-learner can lead to overall results and which time series benefit from our approach.

6.4.1 Basic pipeline results

We first compare the total weighted evaluation results over all TS per run (experiment and randomization). As we are primarily forecasting intermittent time series, we concentrate on the RMSE evaluation metric as unbiased metric for the conditional median, as proposed by Kolassa (Kolassa 2016, 2020). Section 6.4.4 provides a comparison of the results for all evaluation metrics. Further discussion of the interpretation of the differences between the metrics results are provided in Section 6.6.

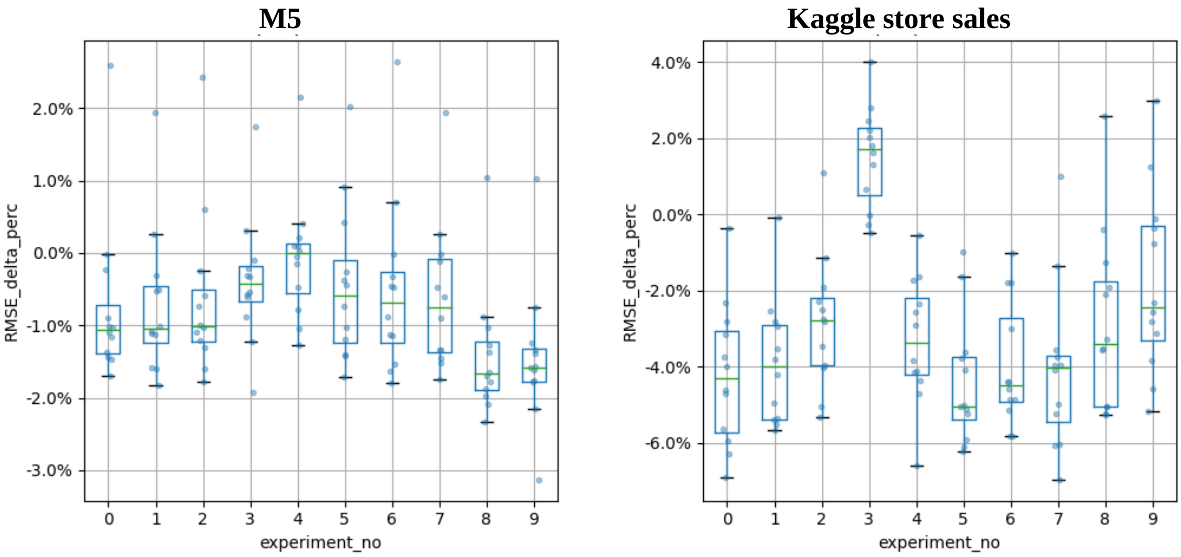


Figure 19: Boxplot of RMSE_delta_perc per experiment before application of the meta-learner (see pipeline step B6). The RMSE is calculated as weighted average of TS scores, weighted by total values of the TS. Dots indicate the RMSE_delta_perc results of each experiment and randomization. Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

Figure 19 shows the results of the experiments before application of the meta-learner as described in step B6, measured by the percentage delta of the RMSE compared to the baseline. The pairwise t-test for mean values between scores of the experiments with their respective baselines indicates statistical significance for different mean values at p-values ≪ 0.05. All significance tests are performed using the Benjamini-Hochberg correction for multiple hypothesis test and p-values are corrected according to their approach (Benjamini and Hochberg 1995). Section 6.4.4 describes how the correction was applied. We will generally consider p-values below 5% as significant throughout this study for pairwise t-tests for independent mean values under Benjamini-Hochberg correction. We note that the actual p-values are mostly considerably lower than this value in fact, unless stated differently. As the mean RMSE_delta_perc is negative for all experiments but one (Kaggle store sales experiment 3), we can deduce that the overall performance of our approach is less than the performance of the baseline (i.e., baseline score is lower than experiment score – where RMSE is a metric where smaller values are better). Consequently, our approach is up to 5% worse than the baseline, when applied to all TS.

Table 32 exhibits the mean RMSE evaluation results compared between the *LGBM* algorithm and the *Croston_TSB* algorithm after step B6 of the basic pipeline (Teunter et al. 2011). As the LGBM evaluation results dominate the TSB results in each randomization, we concentrate on the LGBM results in the further course of the study.

Mean RMSE evaluation result of the basic pipeline before meta-learner application (Step B6)												
Randomization (seed no) →	900	901	902	903	904	905	906	907	908	909	910	911
M5												
Croston_TSB	1.96	2.52	2.06	2.61	2.31	2.98	1.93	2.06	1.79	1.94	2.32	2.16
STSF (LGBM)	1.80	2.51	1.93	2.42	2.14	2.73	1.80	1.87	1.65	1.87	2.22	2.11
Baseline (LGBM)	1.77	2.49	1.91	2.39	2.12	2.79	1.77	1.85	1.63	1.86	2.20	2.11
Kaggle store sales												
Croston_TSB	485.1	385.1	461.9	390.3	515.7	589.3	450.0	395.1	378.0	402.9	562.3	534.6
STSF (LGBM)	305.6	255.8	287.1	230.3	319.9	358.6	272.4	258.8	237.6	255.2	347.6	323.6
Baseline (LGBM)	296.7	256.4	279.7	223.4	308.4	343.4	263.4	256.5	231.7	246.7	333.3	309.7

Table 32: Experiments mean RMSE evaluation results before meta-learner application. The randomizations are denoted by the used seed values. The LGBM algorithm exhibits notably lower errors than the TSB algorithm throughout all randomizations.

6.4.2 Meta-learner results

As described in the previous section, a meta-learning model is applied on TS-level to predict independently which TS benefit from our approach and to assemble an all-stars approach. The results are shown in Figure 20.

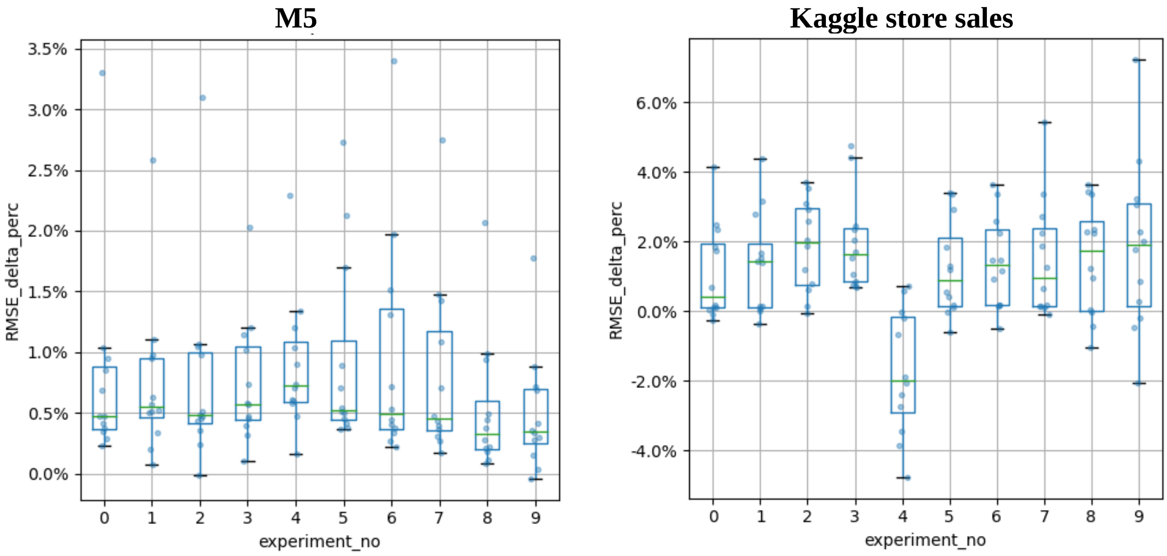


Figure 20: Boxplot of RMSE_delta_perc per experiment after application of the meta-learner (see pipeline step A6). The RMSE is calculated as weighted average of TS scores, weighted by total values of the TS. Dots indicate the RMSE_delta_perc results of each experiment and randomization. Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

In contrast to the previous results in Figure 19, the mean experiment scores with meta-learning all outperform the baseline on an experiment level, only with one exception (Kaggle store sales Experiment 4). The mean value of all RMSE_delta_perc per experiment accounts for 0.7% for M5 and 1.2% for Kaggle store sales. See Table 34 for tabular experiment results.

We observe no patterns in the direct comparison of Figure 19 and Figure 20 with respect to the relationship of the experiments amongst each other. As well, the experiment with PD-statistics approach representation (Experiments 0-4) perform comparably to the experiments with Syntetos-Boylan representation (Experiments 5-9). Observing results on the experiment and randomization level (single dots in the figures), 98% of M5 results perform better than the baseline (RMSE_delta_perc results greater or equal to zero) and 80% of Kaggle store sales results are above baseline.

Figure 21 shows the significant improvement of our approach for TS that the meta-learning algorithm predicts as *stsf*. The mean scores are up to 10% and at least 2% better than the corresponding baseline scores.

Table 33 depicts the accuracy of the meta-learner when predicting TS labels. Over all experiments, the meta-learner classifies 84% of labels correctly, with values between 68% (minimum) and 93% (maximum) per experiment and data set.

The results of all figures shown in this and the previous section exhibit p-values $\ll 0.05$ in paired t-tests for independent mean values (baseline vs. experiment evaluation scores, under Benjamini-Hochberg correction). We therefore deduce statistical significance for all statements in this section (also see Table 38).

	Experiment																			
	0		**1**		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**	
predicted label → actual label ↓	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**	**stsf**	**base line**
M5 data set																				
stsf	25%	6%	26%	4%	29%	6%	31%	17%	30%	23%	27%	6%	27%	6%	24%	5%	26%	7%	26%	8%
baseline	3%	65%	2%	68%	2%	63%	10%	42%	9%	38%	2%	65%	2%	66%	4%	67%	3%	64%	2%	64%
Kaggle store sales data set																				
stsf	38%	9%	40%	11%	41%	11%	43%	14%	49%	16%	34%	8%	34%	9%	32%	9%	20%	11%	16%	12%
baseline	7%	46%	7%	42%	8%	41%	9%	34%	9%	27%	8%	50%	7%	50%	9%	50%	11%	58%	10%	62%

Table 33: Confusion matrix of the actual (rows) and predicted (columns) ‘stsf’- and ‘baseline’-labels of the meta-learning step for each experiment (aggregated over all seeds). Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

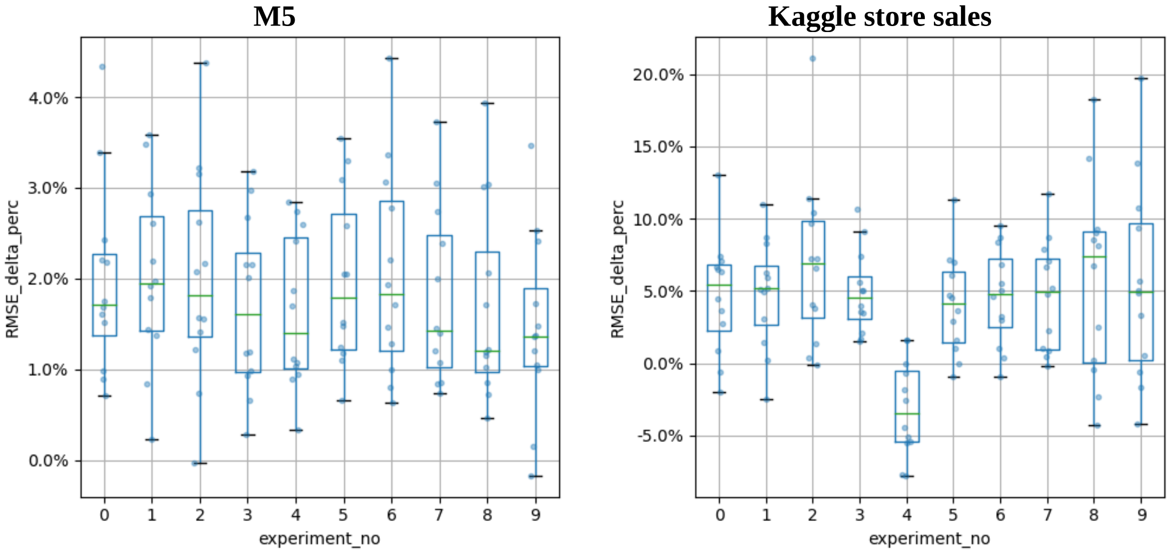


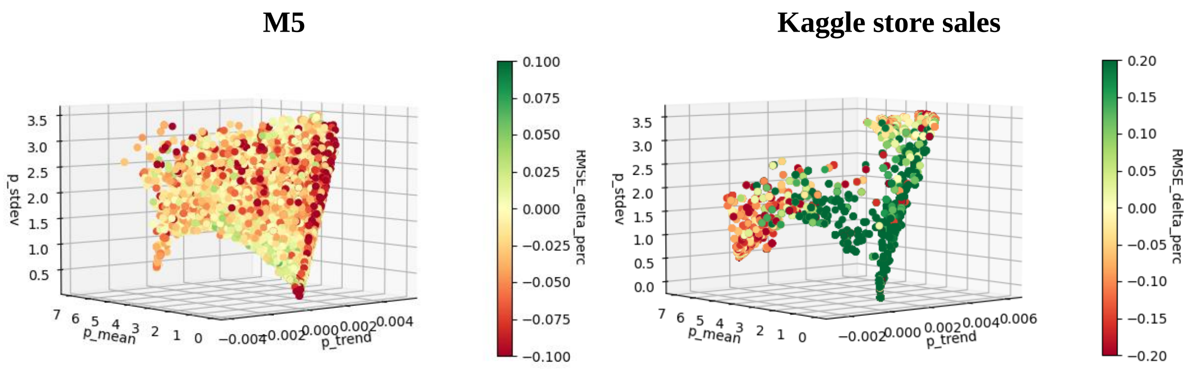
Figure 21: Boxplot of RMSE_delta_perc per experiment after application of the meta-learner (see pipeline step A6) and only for TS with predicted “a” label. The RMSE is calculated as weighted average of TS scores, weighted by total values of the TS. Dots indicate the RMSE_delta_perc results of each experiment and randomization. Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

	Experiment (measured by mean RMSE_delta_perc x 100% over all random seeds)											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
M5 data set												
Before meta-learner (B6) – see Figure 19	-0.73%	-0.70%	-0.63%	-0.39%	**-0.06%**	-0.43%	-0.51%	-0.61%	-1.41%	-1.44%	-0.73%	-0.70%
After meta-learner (A6) all TS – see Figure 20	0.79%	0.75%	0.76%	0.75%	0.89%	0.94%	**0.96%**	0.82%	0.53%	0.49%	0.79%	0.75%
After meta-learner (A6) only predicted “stsf”-labels – see Figure 21	1.97%	2.03%	2.01%	1.70%	1.63%	1.98%	**2.06%**	1.79%	1.70%	1.46%	1.97%	2.03%
Kaggle store sales data set												
Before meta-learner (B6) – see Figure 19	-4.22%	-3.91%	-2.85%	**1.50%**	-3.30%	-4.40%	-3.96%	-4.08%	-2.84%	-1.79%	-4.22%	-3.91%
After meta-learner (A6) all TS – see Figure 20	1.10%	1.36%	1.87%	**1.94%**	-1.73%	1.21%	1.40%	1.50%	1.50%	1.85%	1.10%	1.36%
After meta-learner (A6) only predicted “stsf”-labels – see Figure 21	4.67%	4.80%	**6.91%**	4.92%	-3.16%	4.07%	4.59%	4.69%	5.81%	5.54%	4.67%	4.80%

Table 34: Overview of results of the experiments, given as mean RMSE in percent from the previous Figure 19, Figure 20 and Figure 21. Bold values indicate best experiment result per table row. Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

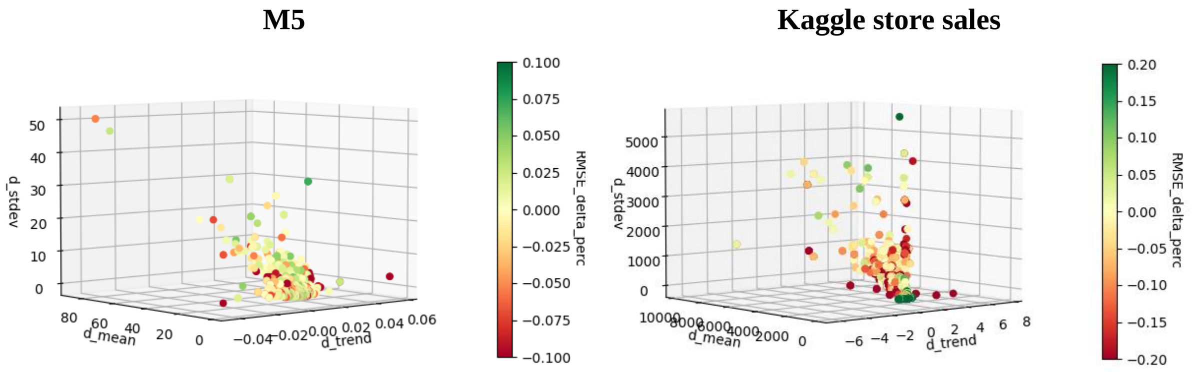
6.4.3 Classification of time series with forecast improvement

From the high classification accuracy of the meta-learner, we deduce that there are characteristics of the time series that make the identification of TS that benefit from our approach possible. From the previous results, general intuition implies that the number of clusters, the representation type, and the individual TS characteristics play a role in the degree of improvement. In this subsection, we concentrate on the TS characteristics only, which were the only inputs available to the meta-learning algorithms: PD-statistics and Syntetos-Boylan representation (irrespective of the actual representation used in clustering). Figure 22 and Figure 23 visualize the PD-statistics and evaluation results. Each point in the scatter plot denotes one TS out of the experiments 0 – 9. Points with green color show TS where the approach exhibits better evaluation results than the baseline. From the visualization, intuition indicates that accumulations of green points are identifiable – especially for low values of *p_stdev*, *p_mean*, *d_stdev*, and *d_mean*. One can also identify a cone-shaped relationship between the P-statistics in Figure 22. It is also apparent that the separation between positive and negative *RMSE_delta_perc* for the Kaggle store sales data is more distinct than for the M5 data, as well as the spread of the *RMSE_delta_perc* values themselves.


M5

Kaggle store sales

Figure 22: 3D scatter plot of P-statistics on TS level. The colors indicate the RMSE_delta_perc. Data from all experiments and seeds. Observe that the color scales are different for M5 and Kaggle store sales.


M5

Kaggle store sales

Figure 23: 3D scatter plot of D-statistics on TS level. The colors indicate the RMSE_delta_perc. Data from all experiments and seeds. Observe that the color scales are different for M5 and Kaggle store sales.

Input	Importance	
	M5	Kaggle store sales
d_mean	13%	19%
d_stdev	11%	12%
p_stdev	11%	2%
CV2	11%	17%
adi	8%	4%
p_mean	7%	3%
d_trend	7%	9%
p_trend	6%	2%
CV2_aggr	4%	6%
d_trend_aggr	4%	6%
d_stdev_aggr	4%	6%
d_mean_aggr	3%	7%
adi_aggr	3%	1%
p_mean_aggr	3%	1%
p_trend_aggr	2%	1%
p_stdev_aggr	2%	1%

Table 35: Meta-learner feature importance values. The values are generated by a random forest model (not the original LGBM model) for visualization purposes only.

	M5		Kaggle store sales	
Feature	cut off	AUC	cut off	AUC
p_stdev	1.58	0.73	3.08	0.47
d_mean	1.55	0.72	11.32	0.77
d_stdev	0.91	0.70	7.03	0.74
CV2	0.31	0.67	0.48	0.32
p_mean	1.79	0.65	4.21	0.77
d_stdev_aggr	14.60	0.60	3,482.52	0.67
d_mean_aggr	30.74	0.59	4,976.98	0.69
p_trend	0.00	0.57	0.00	0.46
d_trend_aggr	0.02	0.57	3.18	0.66
p_mean_aggr	6.96	0.55	6.95	0.65
p_trend_aggr	0.00	0.50	0.00	0.40
CV2_aggr	0.27	0.46	0.12	0.32
d_trend	0.00	0.45	0.00	0.61
p_stdev_aggr	0.42	0.45	0.43	0.39
adi_aggr	1.00	0.44	1.00	0.35
adi	2.01	0.29	1.00	0.31

Table 36: Overview of possible cut-off values and AUC values determined by the ROC-curve approach.

To determine the overall importance of the TS characteristics, we evaluated the feature importance function of the meta-learner. The scikit-learn LGBM implementation (*HistGradientBoostingRegressor*) does not provide a feature importance function, therefore we trained a random forest model instead of the LGBM meta-learner (only to obtain feature importance values). Table 35 shows the importance of each input to the meta-learner. We note that the most relevant inputs are *p_stdev*, *adi*, *p_mean*, *d_mean and CV2*. While *p_stdev* represents a feature of high importance for the M5 data set TS characterization, the input is not of high importance for the classification of the Kaggle store sales data set.

In Table 36, we determined cut-off values for each input to discriminate "*stsf*" and "*baseline*" label TS. Therefore, we chose the ROC curve approach per input. The high AUC values for *p_stdev*, *p_mean*, *d_mean*, and *CV2* again imply the relevance of these inputs to the discrimination of labels. The suggested cut-off values were determined by the threshold value of the minimum Euclidian distance of the ROC curve to the point $(false\ positive\ rate, true\ positive\ rate) = (0,1)$. In contrast to Table 35, the AUC of the input *adi* implies no relevance to discriminate the labels.

We also observe that the representation values of the aggregated clustered time series (denoted by *_aggr*) have limited relevance to discriminate between the labels.

6.4.4 Comparison of different evaluation metrics

While the previous sections concentrated on the results by the RMSE metric, all experiments were evaluated by all metrics stated in section 6.3.2. For comparison with the already discussed results, we show the results in the following tables.

Experiment No.	**delta_perc for TS with predicted “*stsf*”-label**					
	MADmeanRatio	MAE	MASE	R2	RMSE	RMSSE
M5 data set						
0	**5.29%**	**5.29%**	2.67%	-3.27%	1.97%	1.63%
1	4.94%	4.94%	2.42%	**-4.43%**	2.03%	1.51%
2	4.71%	4.71%	1.95%	-4.28%	2.01%	1.44%
3	3.08%	3.08%	1.84%	-2.66%	1.70%	1.41%
4	2.74%	2.74%	1.97%	-2.53%	1.63%	1.27%
5	4.85%	4.85%	**3.01%**	-2.25%	1.98%	1.72%
6	4.43%	4.43%	2.44%	-2.28%	**2.06%**	**1.86%**
7	3.47%	3.47%	1.80%	-1.75%	1.79%	1.52%
8	1.94%	1.94%	1.28%	-3.41%	1.70%	1.04%
9	1.93%	1.93%	1.66%	-2.93%	1.46%	0.98%
Kaggle store sales data set						
0	17.59%	17.59%	**21.38%**	**-1.64%**	4.67%	4.25%
1	15.10%	15.10%	-16.82%	-0.68%	4.80%	4.51%
2	17.25%	17.25%	1.62%	-0.79%	**6.91%**	**6.53%**
3	9.26%	9.26%	10.58%	-0.48%	4.92%	4.00%
4	7.45%	7.45%	8.34%	0.34%	-3.16%	-1.75%
5	18.38%	18.38%	19.78%	-0.54%	4.07%	4.03%
6	**19.88%**	**19.88%**	21.26%	-0.55%	4.59%	4.36%
7	16.84%	16.84%	16.50%	-0.62%	4.69%	5.06%
8	15.73%	15.73%	18.38%	-0.61%	5.81%	4.79%
9	11.02%	11.02%	8.16%	-0.62%	5.54%	4.77%

Table 37: Delta_perc values for different evaluation metrics for TS with predicted "stsf"-label. Bold values indicate the experiment with the best result per evaluation metric. Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

Table 37 shows the *delta_perc* results for different metrics (only predicted “*stsf*”-label TS). The RMSE results are identical to the previous sections. The MAE results are equal to the MADmeanRatio results, as the weighted average function has the same effect. For the R² metric, we observe predominantly negative values. However, these are expected as the R² metric indicates better forecasting accuracy with higher evaluation scores, as opposed to the other metrics. Hence, negative *delta_perc* values indicate only for R² that the experiment results exhibit a better evaluation result than the baseline. The values for MASE, RMSE and RMSSE are comparable to the MADmeanRatio values, however on an overall lower level.

From Table 37 we draw the following conclusions:

- Most of the experiments exhibit a positive *delta_perc* value, which indicates that the STSF approach performs a higher forecasting accuracy than the baseline.
- The evaluation results of the Kaggle store sales are overall higher compared with the M5 results.
- We do not observe a systematic pattern when we compare the results between the experiments.

Metric → Exper-iment ↓	p-value of paired t-test, Benjamini-Hochberg corrected						H0 hypothesis rejected?					
	MAD-mean-Ratio	MAE	MASE	R2	RMSE	RMSSE	MAD-mean-Ratio	MAE	MASE	R2	RMSE	RMSSE
M5 data set												
0	0.02%	0.21%	1.47%	0.09%	0.36%	0.00%	True	True	True	True	True	True
1	0.06%	0.04%	0.00%	0.04%	0.11%	0.00%	True	True	True	True	True	True
2	0.03%	0.30%	0.33%	0.07%	0.48%	0.00%	True	True	True	True	True	True
3	0.02%	0.02%	0.06%	0.05%	0.07%	0.05%	True	True	True	True	True	True
4	0.00%	0.01%	0.06%	0.04%	0.08%	0.02%	True	True	True	True	True	True
5	0.06%	0.02%	0.39%	0.02%	0.03%	0.19%	True	True	True	True	True	True
6	0.12%	0.02%	0.04%	0.03%	0.26%	0.39%	True	True	True	True	True	True
7	0.44%	0.26%	0.01%	0.00%	0.16%	0.19%	True	True	True	True	True	True
8	0.49%	0.23%	2.98%	0.30%	0.47%	0.12%	True	True	True	True	True	True
9	0.06%	0.75%	0.81%	0.26%	0.32%	0.55%	True	True	True	True	True	True
Kaggle store sales data set												
0	0.00%	0.00%	0.00%	*6.32%*	2.88%	1.78%	True	True	True	*False*	True	True
1	0.11%	0.11%	1.62%	1.04%	0.95%	0.98%	True	True	True	True	True	True
2	0.02%	0.00%	*93.92%*	1.03%	0.20%	0.55%	True	True	*False*	True	True	True
3	3.30%	3.22%	3.07%	0.06%	0.06%	0.02%	True	True	True	True	True	True
4	1.67%	0.69%	1.95%	1.30%	1.30%	3.92%	True	True	True	True	True	True
5	0.02%	0.00%	0.02%	0.53%	1.61%	0.69%	True	True	True	True	True	True
6	0.00%	0.00%	0.00%	0.25%	0.45%	0.21%	True	True	True	True	True	True
7	0.02%	0.00%	0.02%	0.40%	0.62%	0.45%	True	True	True	True	True	True
8	0.01%	0.00%	0.00%	*7.01%*	*5.53%*	*8.17%*	True	True	True	*False*	*False*	*False*
9	0.02%	0.00%	*23.87%*	2.09%	2.33%	2.79%	True	True	*False*	True	True	True

Table 38: Left part of the table shows p-values of the paired t-tests for independent mean values of the experiment evaluations compared to their respective baseline, corrected by Benjamini-Hochberg approach for multiple comparisons (false-detection-rate parameter: 5%). The right part of the table shows the, whether the H0 hypothesis ("experiment and baseline yield same mean value") can be rejected at a confidence level of 95%. The table directly relates to Table 37. Experiments 0-4 based on PD-statistics representation, experiments 5-9 on Syntetos-Boylan representation. Experiment cluster sizes: 3, 5, 10, 100, 200.

Table 38 shows the p-value of paired t-tests for independent mean values of the evaluation results between the experiments and the respective baselines. To prevent alpha inflation in the multiple comparison problem, the p-values are corrected by the Benjamini-Hochberg approach at a false-detection-rate of 5% (Benjamini and Hochberg 1995). The table also shows whether the H0 hypothesis ("mean values of the experiment and baseline evaluation are identical") can be rejected at a confidence level of 95%. We observe that most evaluation results on experiment and evaluation metric level are significant (114 out of 120). We conclude that the experiment evaluation results significantly deviate from the baseline evaluation results and therefore that our findings from the previous section to be highly unlikely to be the product of lucky sampling.

6.5 Summary of Results and Conclusion

In this section, we take up the research questions from Section 6.1 and draw our conclusions based on the results of the previous section.

Regarding **RQ1**, we conclude from the results that the approach that we propose can improve forecasting of time series. The results in Section 6.4.1 show that the basic pipeline (B1-B6) is not able to improve the evaluation score when the approach is applied to all time series. All overall evaluation results of the basic pipeline indicate lower forecasting performance than the baseline approach (see Figure 19, all RMSE_delta_perc values less than zero). However, the results in Section 6.4.2 show that the full pipeline, including the meta-learner yields better evaluation results than the respective baseline approach in most of the cases (see Figure 20). This empirical finding is supported by the high confidence levels of the t-test for different mean values (see Table 38). We remark that, depending on the applied evaluation metric, not all experiments yield that the STSF approach shows better results than the baseline approach (see Section 6.4.4).

The question raised in **RQ2** concerns the capability to identify TS that benefit from the approach. As the meta-learner exhibits high accuracy in classifying "*stsf*"/"*baseline*" (better or worse than baseline) labels on TS level only based on the knowledge of the TS representation itself, we can conclude that classification is possible. Furthermore, RQ2 addresses how the TS that benefit from our approach are characterized and how much the improvement is. From the results in Section 6.4.3, we conclude that these TS are mainly characterized by: (1) strong influence of a low variation in periodicity (i.e., *p_stdev*), (2) low overall demand and demand fluctuation (i.e. *d_mean*, *d_stdev*), and (3) low periodicity values (i.e. *p_mean*) (see Table 36). In the Syntetos-Boylan categorization, these are "intermittent" TS (however not exactly congruent with their definition of "intermittent"). The reader should note that the mentioned cut-off values are meant to help to understand the results, but not as a practical recommendation to select TS suitable to the approach, especially as we do not provide overarching values for the statistics for both data sets. As the overall results of the approach rely heavily on the correct labeling of TS, the meta-learner approach is advised.

With respect to **RQ3**, we conclude that TS that benefit from our approach can be identified in advance, i.e., before running the approach. Table 35 shows that the TS characteristics (PD-statistics and Syntetos-Boylan representation on TS-level) account for most of the feature importance of the meta-learner. These characteristics can be calculated before the actual approach (see Step B1). The characteristics on aggregate level account for a small proportion of the feature importance (~ 22%). Accordingly, training the meta-learner only for non-aggregated TS characteristics inputs yields comparably accurate predicted labels.

In addition to the focus on TS characteristics, we can also conclude for **RQ4** how other influences affect the approach. We could not observe systematical differences between the experiments; therefore we cannot conclude whether the number of clusters or the representation approach yield relevant differences. This poses a potential for improvement of the approach, as we will conclude in Section 6.6. From a practitioner view, this means that the STSF approach is relatively robust to the parameters and will exhibit improved forecasting accuracy even without emphasized hyper-parameter optimization.

However, as the evaluation results strongly differ between the two data sets, we can conclude that the data set characteristics influence the performance of the STSF approach.

Summarizing the results, we conclude that our approach contributes to the forecasting of iTS by providing an approach that exhibits reliable, predictable, and reproducible increase of forecasting accuracy for TS that can be characterized in advance before conducting the approach. With an overall improvement based on the RMSE metric for the experiment forecast compared to the state-of-the-art baseline approach of 2.06% (M5, Experiment 6) and 6.91% (Kaggle store sales, Experiment 2) the improvement is to be considered relevant and significant (by the t-tests conducted, see Table 37 and Table 38). Also, for other metrics (see Section 6.4.4), we find significant experiments that provide an improvement over the baseline approach. In its current state, it already provides a tool for practitioners to improve the forecasting quality of iTS. Due to the modular design of the pipeline, it also provides an extensible framework for research to enhance the pipeline further. To the best of our knowledge from the review of the literature, no comparable approach exists that applies aggregation-disaggregation to iTS only based on TS characteristics alone. As discussed in Section 6.1, this is a relevant use case for practice and academia, in cases where forecasting is supposed to be improved where no TS hierarchy is available or is not applicable.

6.6 Critical Discussion

An empirical study always bears the risk that particularities of the setup bias the study results – especially the choice of the data set or data sampling. Our study minimizes this bias by: (1) two well-known and publicly available data sets (M5 competition and Kaggle store sales data set), (2) standard implementation of well-researched libraries (mainly scikit-learn), and (3) the extensive use of reproducible randomization in all steps. The data sets are known for a realistic representation of practice relevant time series, as it comes from actual retail store sales. We also remark that corrected p-values of the t-tests imply statistical significance on high levels of confidence. Therefore, we argue that our results are unlikely to be caused by lucky sampling.

In addition, the pipeline and experiment design offer flexibility for many design choices, especially (1) the choice of a representation method, (2) different clustering, aggregation, and disaggregation algorithms, (3) various forecasting algorithms, and several more decisions. We explain our choices for the three particular points in the following. However, we point out that, in general, due to the vast number of design choices and combinations thereof, neither exhaustive empirical testing nor theoretical research could systematically lead to a best optimal approach in feasible time. Our study represents one feasible approach with reasonable and justified derivation of the applied methods. We assess the vast possibilities of design choices as potentials for future research to enhance our proposed approach.

As the Syntetos-Boylan representation method is a state-of-the-art approach to classify intermittent TS, we decided to apply this method in our approach. The PD-statistics approach that we propose is based on the basic principles of iTS forecasting: periodicity and demand size (comparable to Croston's method). Deriving primitive statistics (mean, standard deviation, and trend) follows the TS feature extraction approaches of several authors who provide time series feature extraction approaches (Fulcher and Jones 2014; Hyndman et al. 2015; Christ et al. 2018). Future research can systematically study whether one of the two representation approaches (PD-statistics or Syntetos-Boylan) yields higher forecasting accuracy.

We also tested different clustering algorithms but found little difference in results. The k-means algorithm is a state-of-the-art approach. In our study, it is useful as it is suitable for clustering when data points are in overall dense formations rather than distinctly separable clusters. However, we are aware that more advanced clustering techniques for TS exist, as shown in Section 6.2.4. For disaggregation, we chose the same forecasting algorithm as in Step B4, for consistency. We assume that advanced disaggregation poses a potential lever for improvement.

The forecasting algorithm could also be a point of discussion in the assessment of our approach. However, we deliberately chose to abstract from the choice of forecasting algorithm as much as possible by using the exact same algorithm for baseline and experiment calculation. From our tests, different algorithms did not systematically improve or impair the results. As shown in Section 6.2, the LGBM can be regarded as a state-of-the-art forecasting algorithm, especially as it has proven best results in the M5 competition. We also considered using Croston's method or SBA/TSB forecasting as baseline forecasting algorithm. However, this yields lower forecasting accuracy, which, if used as a baseline, would impair our experiment results, as it would provide an unrealistic low baseline accuracy. We point out that evaluating the effectiveness of the STSF approach is subject to further research.

We also remark that there were no additional inputs, such as date and time information (e.g., day of week, month, weekend, holiday, etc.) or other exogenous variables (e.g., weather, sales promotion, etc.). Such information could be easily incorporated into the STSF approach whenever a forecasting algorithm is implemented that accepts additional input. Generally, additional inputs are suitable to improve forecasting accuracy under the right conditions. However, we aim to only study the STSF approach and not to overall achieve a high forecast accuracy in this study. We therefore consciously chose to not employ additional inputs to achieve a simple experiment design with as little further influences as possible. We advise future studies to systematically probe how additional inputs affect the performance

In Section 6.4.4 we discussed the different results and significance levels of the results depending on the evaluation metric. Research shows that different metrics can indicate considerably different interpretations. For example, Kiefer et al. compare evaluation results of different metrics and forecasting approaches. The ranking of the approaches differs much depending on the applied evaluation metric (Kiefer et al. 2021). In our case, the overview of the metrics also indicates that the improvement by the

STSF approach is relatively high for the metrics MADmeanRatio, MAE, R² and MASE, however relatively low for the quadratic metrics RMSE and RMSSE. Generally, we note that the metrics that employ squared forecasting error measurements are more suitable for intermittent TS forecasting as they represent unbiased metrics for the conditional mean forecasting error, rather than the conditional median forecasting error (Kolassa 2016, 2020). We remark that it is not clear how particularities of the data set account to this result. We conclude that studying the systematics of the influence of the applied evaluation metrics on the results in the context of the STSF approach is subject for further research.

As an outlook for future research, we propose several directions. First, the approach should be tested in the context of more data sets to gain a deeper understanding of the relationship between the performance of the approach and the characteristics of the TS in the data sets. Second, we propose that the major levers in the STSF, namely the representation and clustering logic (pipeline steps B2, B3), the disaggregation logic (pipeline step B5) as well as the meta-learner setup (pipeline steps A4, A5) should be systematically tested by comparing different methods. From our experiments, we believe that especially more relevant TS features used for TS representation and for meta-learning can boost the performance of the STSF approach. An important step towards understanding ways to improve the representation and clustering will be to compare the hierarchy constructed by the clustering based on the TS characteristics with the already existing hierarchy of the M5 and Kaggle store sales data sets.

6.7 References

Abolghasemi, Mahdi; Hyndman, Rob J.; Tarr, Garth; Bergmeir, Christoph (2019): Machine learning applications in time series hierarchical forecasting. Available online at http://arxiv.org/pdf/1912.00370v1.

Aghabozorgi, Saeed; Shirkhorshidi, Ali Seyed; Wah, Teh Ying (2015): Time-series clustering – A decade review. In *Information Systems* 53 (12), pp. 16–38. DOI: 10.1016/j.is.2015.04.007.

Babai, M. Z.; Dallery, Y.; Boubaker, S.; Kalai, R. (2019): A new method to forecast intermittent demand in the presence of inventory obsolescence. In *International Journal of Production Economics* 209 (2), pp. 30–41. DOI: 10.1016/j.ijpe.2018.01.026.

Babai, M. Z.; Tsadiras, A.; Papadopoulos, C. (2020): On the empirical performance of some new neural network methods for forecasting intermittent demand. In *IMA Journal of Management Mathematics* 31 (3), pp. 281–305. DOI: 10.1093/imaman/dpaa003.

Benjamini, Yoav; Hochberg, Yosef (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. In *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bergmeir, Christoph; Benítez, José M. (2012): On the use of cross-validation for time series predictor evaluation. In *Information Sciences* 191, pp. 192–213. DOI: 10.1016/j.ins.2011.12.028.

Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo (2018): A note on the validity of cross-validation for evaluating autoregressive time series prediction. In *Computational Statistics & Data Analysis* 120, pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

Boylan, John; Syntetos, A. A. (2021): Intermittent demand forecasting. Context, methods and applications / John Boylan, Aris Syntetos. 1st. Hoboken: John wiley & sons.

Chen, Tianqi; Guestrin, Carlos (2016): XGBoost. In Balaji Krishnapuram, Mohak Shah, Alex Smola, Charu Aggarwal, Dou Shen, Rajeev Rastogi (Eds.): Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA, 13.08.2016. New York, NY, USA: ACM, pp. 785–794.

Christ, Maximilian; Braun, Nils; Neuffer, Julius; Kempa-Liehr, Andreas W. (2018): Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). In *Neurocomputing* 307, pp. 72–77. DOI: 10.1016/j.neucom.2018.03.067.

Croston, J. D. (1972): Forecasting and Stock Control for Intermittent Demands. In *Journal of the Operational Research Society* 23 (3), pp. 289–303. DOI: 10.1057/jors.1972.50.

Dantas, Tiago Mendes; Cyrino Oliveira, Fernando Luiz (2018): Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing. In *International Journal of Forecasting* 34 (4), pp. 748–761. DOI: 10.1016/j.ijforecast.2018.05.006.

Doszyń, Mariusz (2019): Intermittent demand forecasting in the enterprise: Empirical verification. In *Journal of Forecasting*. DOI: 10.1002/for.2575.

Eaves, Andrew Howard Charles (2002): Forecasting for the ordering and stock-holding of consumable spare parts. Thesis (Ph.D.). Lancaster University, Lancaster.

Fu, Wenhan; Chien, Chen-Fu; Lin, Zih-Hao (2018): A Hybrid Forecasting Framework with Neural Network and Time-Series Method for Intermittent Demand in Semiconductor Supply Chain. In Ilkyeong Moon, Gyu M. editor Lee, Jinwoo Park, Dimitris Kiritsis, Gregor von Cieminski (Eds.): Advances in production management systems. Part II. Smart manufacturing for Industry 4.0 : IFIP WG 5.7 International Conference, APMS 2018, Seoul, Korea, August 26-30, 2018, Proceedings / edited by Ilkyeong Moon, Gyu M. Lee, Jinwoo Park, Dimitris Kiritsis, Gregor von Cieminski, vol. 536. Cham, Switzerland: Springer (IFIP advances in information and communication technology, 1868-4238, 536), pp. 65–72.

Fulcher, Ben D.; Jones, Nick S. (2014): Highly Comparative Feature-Based Time-Series Classification. In *IEEE Trans. Knowl. Data Eng.* 26 (12), pp. 3026–3037. DOI: 10.1109/TKDE.2014.2316504.

Fulcher, Ben D.; Jones, Nick S. (2017): hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. In *Cell systems* 5 (5), 527-531.e3. DOI: 10.1016/j.cels.2017.10.001.

Fulcher, Ben D.; Little, Max A.; Jones, Nick S. (2013): Highly comparative time-series analysis: the empirical structure of time series and their methods. In *Journal of the Royal Society, Interface* 10 (83), p. 20130048. DOI: 10.1098/rsif.2013.0048.

Gutierrez, Rafael S.; Solis, Adriano O.; Mukhopadhyay, Somnath (2008): Lumpy demand forecasting using neural networks. In *International Journal of Production Economics* 111 (2), pp. 409–420. DOI: 10.1016/j.ijpe.2007.01.007.

Hyndman, Rob; Kang, Yanfei; Montero-Manso, Pablo; Talagala, Thiyanga; Wang, Earo; Yang, Yangzhuoran; O'Hara-Wild, Mitchell (2022): tsfeatures: Time Series Feature Extraction. Available online at https://pkg.robjhyndman.com/tsfeatures/.

Hyndman, Rob J.; Wang, Earo; Laptev, Nikolay (2015): Large-Scale Unusual Time Series Detection. In : 2015 IEEE International Conference on Data Mining Workshop (ICDMW). 2015 IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA, 14.11.2015 - 17.11.2015: IEEE, pp. 1616–1619.

Ke, Guolin; Meng, Qi; Finley, Thomas; Wang, Taifeng; Chen, Wei; Ma, Weidong et al. (2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.): Advances in Neural Information Processing Systems, vol. 30: Curran Associates, Inc. Available online at https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kiefer, Daniel; Bauer, Markus; Grimm, Florian (2021): Univariate Time Series Forecasting: Machine Learning Prediction of the Best Suitable Forecast Model Based on Time Series Characteristics. In Alfred Zimmermann, Robert J. Howlett, Lakhmi C. Jain, Rainer Schmidt (Eds.): Human Centred Intelligent Systems, vol. 244. Singapore: Springer Singapore (Smart Innovation, Systems and Technologies), pp. 152–162.

Kolassa, Stephan (2016): Evaluating predictive count data distributions in retail sales forecasting. In *International Journal of Forecasting* 32 (3), pp. 788–803. DOI: 10.1016/j.ijforecast.2015.12.004.

Kolassa, Stephan (2020): Why the “best” point forecast depends on the error or accuracy measure. In *International Journal of Forecasting* 36 (1), pp. 208–211. DOI: 10.1016/j.ijforecast.2019.02.017.

Kolassa, Stephan; Schütz, Wolfgang (2007): Advantages of the MAD/Mean Ratio over the MAPE. In *Foresight: The International Journal of Applied Forecasting* (6), pp. 40–43. Available online at https://ideas.repec.org/a/for/ijafaa/y2007i6p40-43.html.

Kourentzes, Nikolaos (2013): Intermittent demand forecasts with neural networks. In *International Journal of Production Economics* 143 (1), pp. 198–206. DOI: 10.1016/j.ijpe.2013.01.009.

Kourentzes, Nikolaos; Petropoulos, Fotios; Trapero, Juan R. (2014): Improving forecasting by estimating time series structural components across multiple frequencies. In *International Journal of Forecasting* 30 (2), pp. 291–302. DOI: 10.1016/j.ijforecast.2013.09.006.

Laurinec, Peter; Lóderer, Marek; Lucká, Mária; Rozinajová, Viera (2019): Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption. In *J Intell Inf Syst* 53 (2), pp. 219–239. DOI: 10.1007/s10844-019-00550-3.

Laurinec, Peter; Lucká, Mária (2018): Usefulness of Unsupervised Ensemble Learning Methods for Time Series Forecasting of Aggregated or Clustered Load. In Annalisa Appice, Corrado Loglisci, Giuseppe Manco, Elio Masciari, Zbigniew Raś (Eds.): New frontiers in mining complex patterns. 6th International Workshop, NFMCP 2017, held in conjunction with ECML-PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Revised selected papers / Annalisa Appice, Corrado Loglisci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras (eds.), vol. 10785. Cham: Springer (LNCS sublibrary. SL 7, Artificial intelligence, 10785), pp. 122–137.

Li, Chongshou; Lim, Andrew (2018): A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. In *European Journal of Operational Research* 269 (3), pp. 860–869. DOI: 10.1016/j.ejor.2018.02.029.

Lütkepohl, Helmut (2010): Forecasting Aggregated Time Series Variables. In *OECD Journal: Journal of Business Cycle Measurement and Analysis* 2010 (2), pp. 1–26. DOI: 10.1787/jbcma-2010-5km399r2jz9n.

Makridakis, Spyros; Spiliotis, Evangelos (2021): The M5 Competition and the Future of Human Expertise in Forecasting. In *Foresight: The International Journal of Applied Forecasting* (60), pp. 33–37. Available online at http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3dbsu%26AN%3d148015455%26site%3dehost-live.

Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios (2022): M5 accuracy competition: Results, findings, and conclusions. In *International Journal of Forecasting* 38 (4), pp. 1346–1364. DOI: 10.1016/j.ijforecast.2021.11.013.

Montero-Manso, Pablo; Athanasopoulos, George; Hyndman, Rob J.; Talagala, Thiyanga S. (2020): FFORMA: Feature-based forecast model averaging. In *International Journal of Forecasting* 36 (1), pp. 86–92. DOI: 10.1016/j.ijforecast.2019.02.011.

Nikolopoulos, K.; Syntetos, A. A.; Boylan, J. E.; Petropoulos, F.; Assimakopoulos, V. (2011): An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. In *Journal of the Operational Research Society* 62 (3), pp. 544–554. DOI: 10.1057/jors.2010.32.

Pang, Yue; Yao, Bo; Zhou, Xiangdong; Zhang, Yong; Xu, Yiming; Tan, Zijing (2018): Hierarchical Electricity Time Series Forecasting for Integrating Consumption Patterns Analysis and Aggregation Consistency. In Jeffrey S. Rosenschein, Jérôme Lang (Eds.): Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18}. Stockholm, Sweden, 7/13/2018 - 7/19/2018. California: International Joint Conferences on Artificial Intelligence Organization, pp. 3506–3512.

Pedregosa, Fabian; Varoquaux, Gaël; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier et al. (2011): Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* 12 (85), pp. 2825–2830. Available online at http://jmlr.org/papers/v12/pedregosa11a.html.

Rostami-Tabar, Bahman; Babai, M. Zied; Syntetos, Aris; Ducq, Yves (2013): Demand forecasting by temporal aggregation. In *Naval Research Logistics* 60 (6), pp. 479–498. DOI: 10.1002/nav.21546.

Saxena, Amit; Prasad, Mukesh; Gupta, Akshansh; Bharill, Neha; Patel, Om Prakash; Tiwari, Aruna et al. (2017): A review of clustering techniques and developments. In *Neurocomputing* 267 (10), pp. 664–681. DOI: 10.1016/j.neucom.2017.06.053.

Syntetos, A. A.; Boylan, J. E.; Croston, J. D. (2005): On the categorization of demand patterns. In *Journal of the Operational Research Society* 56 (5), pp. 495–503. DOI: 10.1057/palgrave.jors.2601841.

Syntetos, Aris A.; Babai, Zied M.; Gardner, Everette S. (2015): Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. In *Journal of Business Research* 68 (8), pp. 1746–1752. DOI: 10.1016/j.jbusres.2015.03.034.

Syntetos, Aris A.; Boylan, John E. (2005): The accuracy of intermittent demand estimates. In *International Journal of Forecasting* 21 (2), pp. 303–314. DOI: 10.1016/j.ijforecast.2004.10.001.

Teunter, Ruud H.; Syntetos, Aris A.; Zied Babai, M. (2011): Intermittent demand: Linking forecasting to inventory obsolescence. In *European Journal of Operational Research* 214 (3), pp. 606–615. DOI: 10.1016/j.ejor.2011.05.018.

Viswanathan, S.; Widiarta, H.; Piplani, R. (2007): Forecasting aggregate time series with intermittent subaggregate components: top-down versus bottom-up forecasting. In *IMA Journal of Management Mathematics* 19 (3), pp. 275–287. DOI: 10.1093/imaman/dpn001.

Willemain, Thomas R.; Smart, Charles N.; Schwarz, Henry F. (2004): A new approach to forecasting intermittent demand for service parts inventories. In *International Journal of Forecasting* 20 (3), pp. 375–387. DOI: 10.1016/S0169-2070(03)00013-X.

Willemain, Thomas R.; Smart, Charles N.; Shockor, Joseph H.; DeSautels, Philip A. (1994): Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. In *International Journal of Forecasting* 10 (4), pp. 529–538. DOI: 10.1016/0169-2070(94)90021-3.

Williams, T. M. (1984): Stock Control with Sporadic and Slow-Moving Demand. In *Journal of the Operational Research Society* 35 (10), pp. 939–948. DOI: 10.1057/jors.1984.185.

Zotteri, Giulio; Kalchschmidt, Matteo; Caniato, Federico (2005): The impact of aggregation level on forecasting performance. In *International Journal of Production Economics* 93-94 (3), pp. 479–491. DOI: 10.1016/j.ijpe.2004.06.044.

7 Generalization of the Approach's results

[*This chapter corresponds to an article submitted to the Journal of Forecasting (JoF). It is currently in first submission status in the acceptance process at JoF:* Bauer, Markus; van Dinther, Clemens; Grimm, Florian; Kiefer, Daniel (2023a): Evaluating the Similarity-based Time Series Forecasting Approach: Generalization of the Results. *– further referred to as* Bauer et al. 2023a. *The article is based on the submitted article in the previous chapter.*]

Evaluating the Similarity-Based Time Series Forecasting Approach: Generalization of the Results

Forecasting sets of time series with no additional external data is known to be difficult. In this article, we pick up the Similarity-based Time Series Forecasting (STSF) approach which demonstrates an approach that exploits statistical similarities to cluster time series and applies an aggregation-disaggregation forecasting procedure. Our study extends this approach to six additional data sets with substantially different characteristics and compares the results of the STSF approach on these data sets. We describe the extensive experiment design conducted, the evaluation based on the RMSE metric and five more metrics, and the tests for statistical significance. We conclude that the STSF approach improves the forecasting results by 1% to 5% for four data sets and 25% to over 40% for two more data sets. Furthermore, we empirically explore the potential of the approach for the data set and identify levers to improve the STSF approach, and point out future potentials for research of the STSF pipeline.

7.1 Introduction and Problem Definition

The field of time series forecasting generally studies how the prediction of future values of time series can be performed and optimized. In this context, we define the time series of sequences of time as discrete observations of a variable. Time series forecasting is relevant for both research (e.g., physics, biology, engineering, etc.) and practical applications (e.g., finance, business, healthcare, etc.). As a concrete example, we chose the customer demand for goods in the stores of a retailer – without the loss of generality. To predict future observations, researchers employ both statistical and machine learning (ML) methods.

In this study, we concentrate on specific cases of time series forecasting, which can be further specified by four characteristics:

1. The absence of external data that can be used to correlate the variable observation to other dependent variables (for example, retail store sales: the demand for ice could depend on the outside temperature).

2. The knowledge that a hierarchical relationship exists between the time series objects where objects' time series are likely to have correlated patterns of values (for example, retail store sales: "green umbrellas" and "black umbrellas" can be considered part of the general article group "umbrellas" that could be expected to exhibit similar customer demand patterns).
3. No actual knowledge of the particular hierarchy and relationships of objects (for example, retail store: no comprehensive and statistically justified product hierarchy for some hundred-thousands of products has been compiled yet – the usual situation practitioners will find in real-life situations).
4. The presence of intermittent time series, which are time series with a notable share of periods where the value of the variable equals zero (for example, retail store sales: umbrellas might not be requested by customers during periods characterized by no precipitation).

In this study, we examine an approach described earlier by Bauer et al., called Similarity-based Time Series Forecasting (STSF) which is specialized in the conditions explained previously. The approach employs an ML pipeline where the time series of one data set is first clustered based on specific time series characteristics, then forecasted in an aggregation-disaggregation approach, and finally assessed by a classifying ML algorithm to identify time series that are likely to benefit from the approach. The authors compare the results of their approach to a baseline approach that omits the clustering and aggregation-disaggregation steps. In their empirical study, they apply the STSF pipeline to two publicly available data sets, namely the M5 competition data set and the Kaggle store sales data set (see description in Section 7.3.1).

Bauer et al. conclude that the approach empirically exhibits improvements of up to 7% compared to the baseline approach for time series that are predicted to benefit from the STSF approach assumed from the applied evaluation metric (Bauer et al. 2023b). Extending their work, we address the following research questions in this study:

RQ1. Does the STSF approach yield higher forecasting accuracy rates than the baseline approach for other data sets than the *M5 competition* and *Kaggle store sales* data set?

RQ2. Are the improvements consistent and significant?

RQ3. What ranges of improvements in the forecast evaluation can be measured based on the examined data sets?

RQ4. Is the approach able to determine time series that yield higher evaluation results with the approach compared to the baseline? Which factors can be identified that contribute to higher evaluations?

7.2 Related Work

In this study, we specifically assess the study by Bauer et al. To relate the approach to the state-of-the-art of time series forecasting, we researched literature concerning general time series forecasting as well as certain subtopics that are relevant components of the proposed approach:

1. Similarity measures for time series and intermittent time series categorization
2. Time series clustering based on similarities
3. Aggregation-disaggregation forecasting

7.2.1 General approaches in time series forecasting

Time series forecasting applies general algorithms from statistics and, in recent years, also ML. Commonly used statistical methods are linear regression, exponential smoothing, and autoregression and notable examples thereof are ARIMA and ARCH and their manifold variants (de Gooijer and Hyndman 2006). In the studies in the following sections, the exponential smoothing method by Holt and Winters is commonly used (Winters 1960; Holt 2004). While many studies refer to this method as ETS (exponential smoothing), the Holt-Winters method is only one special form of ETS that can capture seasonality.

Recent developments in ML algorithms have also been widely adopted by the community of time series forecasting researchers. Such examples are genetic or evolutionary algorithms (GA, EA), support vector machines (SVM), decision tree algorithms (DT), and artificial neural networks (ANN). Again, a broad variety of specialized techniques have been developed and adopted for each of the above-mentioned algorithms (Mahalakshmi et al. 2016).

A general statement about which algorithm performs best is almost impossible to make and highly dependent on specific circumstances such as the data set, the forecasting goal, the evaluation metric, etc., as Petropoulos et al. elaborate in their study (Petropoulos et al. 2014).

In recent years, forecasting competitions have been a playground for researchers and practitioners to compare their forecasting approaches in comparable ways on publicly available data sets and under reproducible conditions. The online platform Kaggle hosts numerous such competitions for all purposes. Among the competitions that received significant attention from the research community, some are especially noteworthy here: the Makridakis competitions (1982: M1, 1993: M2, 2000: M3, 2020: M4, 2021: M5), the neural networks competitions (NN3, NN5), the KDD Cup, and the Global Energy Forecasting Competitions (Hyndman 2020). In the recent M5 competition, implementations of the LightGBM Gradient Boosting Decision Tree algorithm (Ke et al. 2017) have been ranked on top scores (Makridakis et al. 2022).

7.2.2 Time series clustering

Time series clustering is a field of research that specifically applies clustering approaches to time series. The goal is to define groups of similar time series out of a population of time series. As Aghabozorgi et al. point out, the approaches follow four steps on a most abstract level: (1) calculation of a time series representation, (2) definition of a similarity measure based on the time series representation, (3) clustering of the time series using the similarity measure, and (4) evaluating the quality of clusters of time series. Steps (3) and (4) are often combined in iterative steps until the algorithm reaches a certain level of cluster quality or until the algorithm has constructed a pre-defined number of clusters. In each of the steps (1) to (4), the study finds a multitude of approaches and variations thereof. In general, none of the approaches dominates any other, but each has its case-specific advantages (Aghabozorgi et al. 2015).

Most relevant for this study are approaches to characterize time series. i.e., time series representations. Generally, several studies propose extensive frameworks to extract time series features, for example, by Hyndman et al., `FeatuRe` by Christ et al., `TSFEL` by Barandas et al., as well as `Tslearn` by Tavenard et al. (Hyndman et al. 2015; Christ et al. 2018; Barandas et al. 2020; Tavenard et al. 2020). All studies run up to several hundred calculations to generate generic features that can be used to represent the time series. The main challenge for researchers and practitioners is then to select the truly relevant features from the plethora of generic features. With their frameworks, the authors provide several feature selection tools, but the application still needs to be adapted case specifically.

Syntetos et al. provide a categorization approach to differentiate between smooth, erratic, intermittent, and lumpy time series. Generally, intermittent time series are time series with a considerable portion of observations with zero demand. The authors propose to categorize time series by two measures: (1) average demand interval (*adi*) denotes the mean number of observations with zero demand between two non-zero demand observations. (2) *CV2* denotes the coefficient of variation of only the non-zero time series observations (i.e., excluding the zero demand observations). Based on the two measures, the authors derive four classes of time series (see Table 39, Syntetos et al. 2005). In this study, we apply the measures *adi* and *CV2* as one approach for time series categorization.

TS categorization	adi ≤ 1.32	adi > 1.32
CV2 > 0.49	"*erratic*"	"*lumpy*"
CV2 ≤ 0.49	"*smooth*"	"*intermittent*"

Table 39: Categorization scheme of intermittent time series (Syntetos et al. 2005)

7.2.3 Aggregation-disaggregation and hierarchical forecasting

We observe two types of methodologies that exploit aggregation-disaggregation for time series forecasting: (1) temporal aggregation (TA) techniques and (2) hierarchical forecasting (HF). Temporal aggregation techniques aggregate observations into greater buckets of time periods. Hierarchical forecasting exploits hierarchical relationships between the objects to be forecasted, e.g., products in a common product group.

Both in TA and HF, the literature differentiates between bottom-up (BU) approaches where forecasts of objects (lower temporal levels respectively for TA) are aggregated to yield an aggregate object group forecast (higher temporal levels respectively for TA) or, vice versa, aggregate object level forecasts are disaggregated to the object level to yield object level forecasts (top-down approach – TD) (Babai et al. 2022). Early studies have shown that time series forecasting at low levels can benefit from aggregate level forecasts and vice versa (Zotteri et al. 2005; Athanasopoulos et al. 2009). Recent studies have further developed this field of research by focusing on coherent BU-TD forecasts, where BU forecasts and TD forecasts consistently add up between levels of aggregation. Notable foundational research has been conducted by Hyndman et al., known as "GLS reconciliation" or "optimal combination" and Wickramasuriya et al., known as "minimum trace reconciliation (minT)." Their studies show that an integrated reconciled forecast that is consistent across all levels of aggregation can be practically estimated (Hyndman et al. 2011; Wickramasuriya et al. 2019). Recently, Hollyman et al. surveyed and consolidated the numerous works in the field of coherent TD-BU approaches into one approach (Hollyman et al. 2021). Pennings and van Dalen propose an approach that simultaneously forecasts all levels of a given product hierarchy and conclude better forecasting accuracy than with traditional BU or TD approaches (Pennings and van Dalen 2017). Along with several experimental studies, the literature shows that the approaches not only produce coherent forecasts over aggregation levels but also overperform traditional TD or BU approaches (Eckert et al. 2021; Athanasopoulos et al. 2009).

The research also covers several approaches where temporal hierarchies and TA are analogously applied to HF to improve forecasting (Athanasopoulos et al. 2017). Nystrup et al. propose an approach where the time series of different levels of a temporal hierarchy are assessed by their autocorrelation and forecasted simultaneously. In their experimental study on load forecasting, they also find that their approach is coherent over the temporal levels and outperforms TD or BU approaches (Nystrup et al. 2020).

The studies in this section all assume the knowledge of an externally given hierarchy of objects (e.g., by region, product group, customer group, week, month, etc.). In contrast to this, the study by Bauer et al.

makes the premise that such a hierarchy may exist but is not available for the forecast algorithm (see Section 6.1, points 2 and 3).

7.2.4 Time series similarity forecasting

Numerous studies combine the previously described elements of time series similarity measures, clustering, and/or aggregation-disaggregation.

In a first step, we find that several studies use time series similarity measures to identify similar time series and then cluster the time series. In many cases, the studies primarily use this clustering to segment the time series for forecasting. They either choose specific forecasting algorithms for each cluster of time series (cluster-specific algorithm) (for example, Gür et al. 2015) or train different models of the same algorithm separately and independently on each cluster of time series (cluster-specific learning) (for example, Venkatesh et al. 2014; Bandara et al. 2020; Hartomo and Nataliani 2021). Other studies use time series similarities to sample from clusters of similar time series to obtain forecasts (Martinez Alvarez et al. 2011).

Laurinec et al. present a comparable approach to Bauer et al. to forecast electricity demand. They normalized electricity demand time series and calculated representations of the time series which are the coefficients of multiple linear regressions of each time series. The time series are subsequently clustered (k-means and DBSCAN) by their representations and aggregated. Different forms of classical statistical, bootstrap, and ensemble forecasting methods are then applied to generate forecasts. In contrast to the study by Bauer et al., the aggregated forecast is not disaggregated back to the original level. The study concludes that several algorithms produced better forecasting outputs when the clustering and aggregation approach was applied (Laurinec et al. 2019; Laurinec et al. 2016).

Pang et al. also propose a hierarchical forecasting approach to predict electricity demand. Their approach uses the x-means algorithm (Pelleg and Moore 2002) to construct a hierarchy of consumption timelines with similar patterns. The study then applies different state-of-the-art coherent HF techniques (GLS reconciliation and minT) as well as their extended approach using regularization (called CHF, CHF-reg) for forecasting on all levels of the constructed hierarchy. However, the exact clustering procedure is not explained in detail. The authors conclude that their CHF-reg approach performs with a higher forecasting accuracy than GLS or minT (Pang et al. 2018) although the study does not evaluate whether the clustering approach yielded better results than forecasting the original time series separately without the constructed hierarchy.

Pipeline		Application	Representation and similarity	Clustering and aggregation	Forecasting	Disaggregation
Laurinec et al. 2019 and Laurinec et al. 2016	1. Representation 2. Clustering 3. Aggregation 4. Forecasting	Electricity demand forecast	a) Robust Linear Model, b) Generalized Additive Model c) Holt-Winters Exponential Smoothing and d) Median daily profile	k-means and k-means ++	Several different approaches from statistics, machine learning and ensemble methods	No disaggregation
Pang et al. 2018	1. Clustering 2. Coherent hierarchical forecasting	Electricity demand forecast	Not described	x-means (Pelleg and Moore 2002)	ETS	
Bauer et al. 2023b	1. Representation 2. Clustering 3. Aggregation 4. Forecasting 5. Disaggregation 6. Best forecast prediction	General intermittent time series	PD-statistics, Syntetos-Boylan	k-means euclidiean distance	LGBM regression	LGBM regression

Table 40: Comparison of time similarity forecasting approaches.

7.3 Methodology and Approach

Similar to the previous study by Bauer et al., we evaluate the STSF approach empirically by employing a systematic scheme of experiments. We thus chose six data sets, ran a series of experiments and randomizations, and then compared the STSF results with the baseline approach. The following sections depict the data sets, experiments, evaluation results, and the analysis of the evaluation.

For the formulas used below, we introduce the following terminologies. A forecast is an estimation $\hat{y}_{i,t+1}$ of an actual future observation $y_{i,t+1}$ where $i \in I = \{0,1,\dots,i_{max}\}$ is an object (e.g., product) from a set of objects at a time step $t \in T = \{0,1,\dots,t_{max}\}$ from a sequence of time steps. The forecast is based on an approximation of the function $y_{i,t+1} = f(X_i,\Phi) + \epsilon_{i,t+1}$, where $X_i = \{x_{i,t}\}, t \in \{0,1,\dots,t\}$ are known past observations of the object i, Φ denotes a set of observations of stochastic influences, and $\epsilon_{i,t+1}$ a random stochastic variable that may depend on X_i and Φ. The mean value of the observations of the object i is denoted as $\overline{y}_i$.

7.3.1 Overview of data sets

While Bauer et al. only evaluated their approach with the *M5 competition* and *Kaggle store sales* data sets, this study considers six additional data sets (see Table 41 and Figure 24). Except for the *OEM article* data set, all other data sets are publicly available as forecasting competitions on Kaggle and hence the results of the experiments are reproducible and verifiable by other researchers. The data sets *Kaggle store sales*, *M4 monthly*, *M5 competition*, *OEM article*, *Rossmann*, and *Walmart* represent different forms of sales of goods. *COVID4* represents the number of confirmed COVID19 cases while *Website traffic* represents the traffic of Wikipedia articles.

	COVID4	Kaggle store sales	M4 monthly	M5 competition	OEM article	Rossmann	Walmart	Website traffic
Description	Confirmed COVID cases per country	Kaggle competition on store sales	Makridakis M4 accuracy competition monthly level	Makridakis M5 accuracy competition	After market article sales of an automotive OEM	Rossmann store sales forecasting	Walmart store sales forecasting	Wikipedia web traffic per page
Source	Kaggle[4]	Kaggle[5]	Kaggle[6]	Kaggle[7]	Non-public	Kaggle[8]	Kaggle[9]	Kaggle[10]
No. of time series[11]	313	1,782	48,000	30,490	89,329	934	3,291	145,035
No of time steps	115	1,684	2,794[12]	1,941	284	942	143	804
Share of zero time periods	34.7%	31.3%	0.0%	68.0%	90.2%	16.8%	10.5%	7.4%
TS classification[13]								
- smooth	133 (42.5%)	945 (53.0%)	45,563 (94.9%)	1,908 (6.3%)	17,457 (19.5%)	932 (99.8%)	2,536 (77.1%)	26,985 (18.6%)
- erratic	180 (57.5%)	427 (24.0%)	2,437 (5.1%)	868 (2.8%)	1,300 (1.5%)	0	397 (12.1%)	111,156 (76.6%)
- intermittent	0	226 (12.7%)	0	22,150 (72.6%)	44,427 (49.7%)	2 (0.2%)	68 (2.1%)	581 (0.4%)
- lumpy	0	184 (10.3%)	0	5,564 (18.2%)	26,145 (29.3%)	0	290 (8.8%)	6,313 (4.4%)

Table 41: Overview of data sets applied in this study, including an outline of their time series' characteristics. The composition of the data sets varies strongly with respect to size and TS classification.

Table 41 shows that the composition of the data sets concerning time series characteristics differs considerably. The experiments are therefore supposed to represent a broader average of typical forecasting data sets than the original data sets alone, to allow for more general statements in this study.

[4] https://www.kaggle.com/competitions/covid19-global-forecasting-week-4

[5] https://www.kaggle.com/competitions/store-sales-time-series-forecasting

[6] https://www.kaggle.com/datasets/yogesh94/m4-forecasting-competition-dataset

[7] https://www.kaggle.com/competitions/m5-forecasting-accuracy

[8] https://www.kaggle.com/competitions/rossmann-store-sales

[9] https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting

[10] https://www.kaggle.com/competitions/web-traffic-time-series-forecasting

[11] Only considering complete time series within the data set containing data (even if zero or NaN) for all time steps

[12] Describes the maximum value per time series; not all time series in M4 monthly data set exhibit full length

[13] According to the classification of Syntetos et al. (Syntetos et al. 2005)

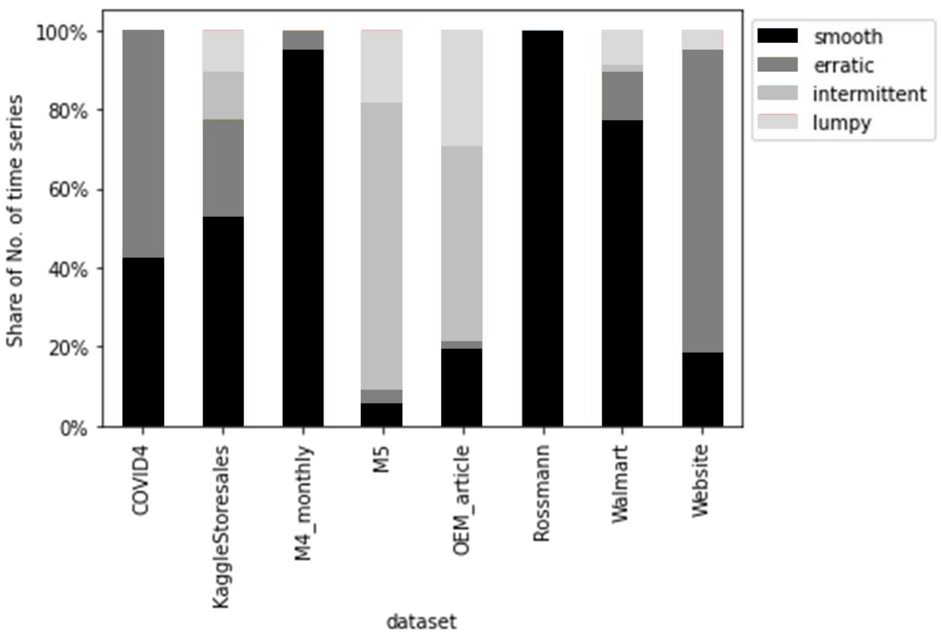


Figure 24: Visualization of the proportions of the data sets' time series characterizations.

Analogously to the original study by Bauer et al., we only extracted the univariate time series from the data sets and do not provide additional data such as date and time information (e.g., day of the week, month, season, holidays, etc.) or exogenous information (e.g., weather, sales promotions, product information, store ID, etc.). We deliberately chose to only provide the univariate time series to reduce the complexity of the experiment comparison in the following sections. We emphasize that the target of this study is to deduct general statements concerning the STSF approach's performance, and not to achieve a competitive result in the forecasting competitions (which would indisputably require several of the additional forecasting inputs mentioned).

7.3.2 STSF pipeline

The STSF approach applies a basic pipeline and an experiment layer (see Figure 25). In the following, we describe the STSF as proposed in Bauer et al. and highlight the deviations from the original approach.

The basic pipeline takes a data set as input (B1), applies a representation algorithm on the time series to quantify their time series characteristics (B2), clusters the time series according to their representation, and aggregates the time series values per time step for each cluster (B3), conducts forecasting on the aggregate time series (B4), and disaggregates the result using another forecasting algorithm (B5) before it finally evaluates the forecast result on a disaggregate level with the original data (B6). To be in line with the original study by Bauer et al., we employ the same algorithms in the pipeline: k-means clustering (B3), Histogram-based Gradient Boosting Regression Tree for aggregate forecasting (B4), and disaggregation (B5). Both algorithms are used from the scikit-learn implementation (Pedregosa et al. 2011). The forecast (B4, B5, and the baseline forecasting) is provided with a rolling window of 20 past time steps as input to perform a prediction of the next consecutive time step (i.e., one step ahead prediction). The evaluation (B6) is based on the *root mean square error* (*RMSE*). In addition to the *RMSE*, we will also report further metrics: *MADmeanRatio* (Kolassa et al. 2007), *mean absolute error* (*MAE*), *mean absolute scaled error* (*MASE*), *root mean square scaled error* (*RMSSE*) (Hyndman and Koehler 2006), and *coefficient of correlation* (*R²*). The result per time series is subsequently weighted by the mean value of the time series as a share of the whole data set to yield an evaluation result per data set. We chose the *RMSE* metric as the main evaluation metric as it yields an unbiased estimation of the unbiased conditional mean error as opposed to absolute errors that aim for the conditional median error. As Kolassa shows, the latter is unfavorable for intermittent time series forecasting (Kolassa 2016, 2020).

As time series representation approaches (see B2), we apply the Syntetos-Boylan approach as well as the PD-statistics approach, as proposed by Bauer et al. The PD-statistics calculate representations of each time series based on the periodicity (P-statistic, i.e., the sequence of non-zero demands in rolling windows of the time series of length h) and the non-zero demand values (D-statistics, i.e., sequence of values of the demand observations from the time series that are greater than zero). Based on the periodicity sequence and the non-zero demand sequence, three statistical features are calculated: the arithmetic mean (*_mean*), the standard deviation (*_stdev*), and the slope of the ordinary least square regression of the sequence values along the time axis (*_trend*). The result comprises six features derived from each time series: *p_mean*, *p_stdev*, *p_trend*, *d_mean*, *d_stdev*, and *d_trend*.

On the experimental level, a set of experiments is defined, specifying the parameters for the basic pipeline (A1) and these experiments are then run iteratively by the basic pipeline with randomization (A2). The results of the randomized experiment runs are then compared to the baseline results (A3), both on a time series level as well as on an experiment level. Subsequently, each time series of each experiment is labeled "*stsf*" if the *RMSE* evaluation result of the time series is smaller than the respective corresponding baseline's time series' *RMSE* evaluation result or labeled "*baseline*" otherwise. The meta-learner classifier is then trained to predict "*stsf*" or "*baseline*" labels for each time series per experiment (A4). The inputs for the meta-learner are the time series representation data (see B2) as well as the respective aggregate's time series' representation data (see B3). The result of the meta-learner's prediction represents predicted "*stsf*" and "*baseline*" labels for each time series per experiment. In step A5, a predicted optimal set is compiled: the prediction of the time series from the basic pipeline aggregation-disaggregation forecast for predicted "*stsf*"-label time series and the baseline forecast for predicted "*baseline*"-label time series (A5) and finally evaluated (A6).

The baseline comparison in step A3 is calculated as the difference between the *RMSE* of the baseline minus the *RMSE* of the experiment divided by the baseline *RMSE* (see formula F1). We refer to this metric as *RMSE_delta_perc*. Section 7.3.3 describes the details of the experiments conducted in this study.

F1. $$RMSE\_delta\_perc = \frac{RMSE_{baseline} - RMSE_{experiment}}{RMSE_{baseline}}$$

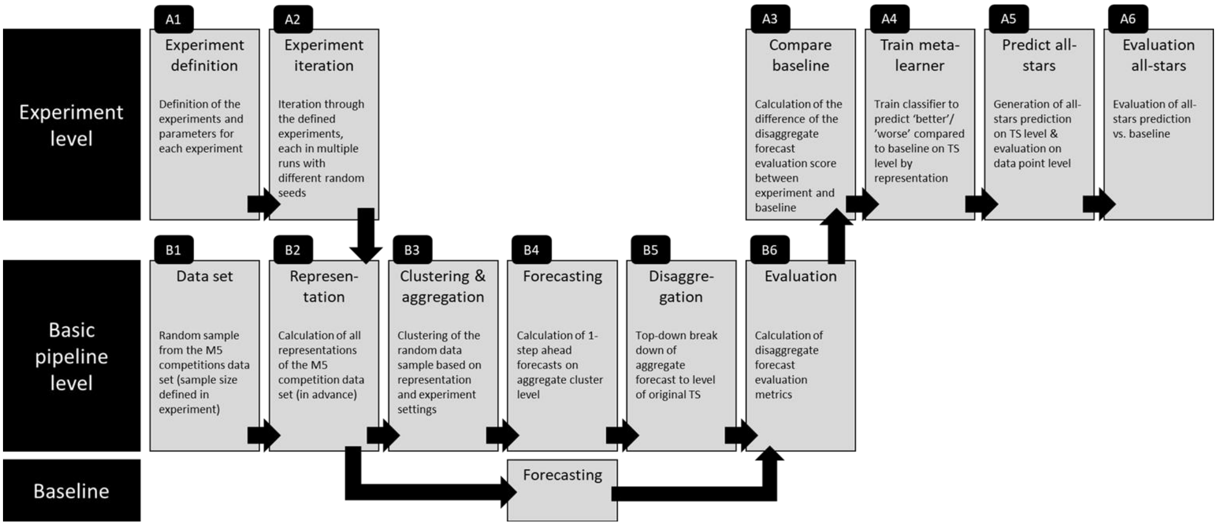

Experiment level
A1
Experiment definition
Definition of the experiments and parameters for each experiment
A2
Experiment iteration
Iteration through the defined experiments, each in multiple runs with different random seeds
A3
Compare baseline
Calculation of the difference of the disaggregate forecast evaluation score between experiment and baseline
A4
Train meta-learner
Train classifier to predict 'better'/ 'worse' compared to baseline on TS level by representation
A5
Predict all-stars
Generation of all-stars prediction on TS level & evaluation on data point level
A6
Evaluation all-stars
Evaluation of all-stars prediction vs. baseline
Basic pipeline level
B1
Data set
Random sample from the M5 competitions data set (sample size defined in experiment)
B2
Represen-tation
Calculation of all representations of the M5 competition data set (in advance)
B3
Clustering & aggregation
Clustering of the random data sample based on representation and experiment settings
B4
Forecasting
Calculation of 1-step ahead forecasts on aggregate cluster level
B5
Disaggre-gation
Top-down break down of aggregate forecast to level of original TS
B6
Evaluation
Calculation of disaggregate forecast evaluation metrics
Baseline
Forecasting

Figure 25: The forecasting and experiment pipeline of the STSF approach (Bauer et al. 2023b).

All forecasts (B4, B5, A4/A5) are performed using k-fold cross validation with three folds to split the data into train and test data portions. Therefore, the predictions are independent of the training data, i.e., the model always predicted the forecasts for portions of the data that it has not been trained on (Bergmeir and Benítez 2012).

We highlight one relevant difference to the pipeline described by Bauer et al. in step B5. While the original study only provides a window of the last two historical values of the target variable to the ML disaggregation algorithm, we provide a window of as many historical values as provided in step B4 or also to the forecasting algorithm in the baseline step (i.e., the last 20 time steps in all cases consistently). Providing fewer data in step B5 is (1) unnecessary because it does not provide more information to the STSF algorithm than to the baseline approach algorithm. Rather, providing a shorter window of historical values (2) poses an unjustified disadvantage to the STSF approach as the disaggregation loses information on the disaggregated time series behavior compared to the baseline approach.

7.3.3 Experiments

The experiments conducted for each data set are listed in Table 42. The parameter *sample size* describes the number of time series randomly selected from the data sets for the experiment (without replacement). For each experiment, 12 randomized iterations of the basic pipeline were performed (see Figure 25 step A3).

The experiments mainly differ in the representation scheme applied (see Figure 25 step B2) and the number of clusters (see Figure 25 step B3).

	COVID4	**Kaggle store sales**	**M4 monthly**	**M5 competition**	**OEM article**	**Rossmann**	**Walmart**	**Website traffic**
Sample size	300 TS	500 TS	500 TS	500 TS	500 TS	500 TS	500 TS	500 TS
P-statistics window size h	7	7	12	7	7	7	7	7
Randomizations	12	12	12	12	24	12	12	12

Represen-tation	No of clusters	Experiment No.							
PD-statistics	3	0	0	0	0	0	0	0	0
	5	1	1	1	1	1	1	1	1
	10	2	2	2	2	2	2	2	2
	100	3	3	3	3	3	3	3	3
	200	4	4	4	4	4	4	4	4
Syntetos-Boylan	3	5	5	5	5	5	5	5	5
	5	6	6	6	6	6	6	6	6
	10	7	7	7	7	7	7	7	7
	100	8	8	8	8	8	8	8	8
	200	9	9	9	9	9	9	9	9
Baseline	No clustering								

Table 42: Overview of the experiments conducted and the experimental settings per data set.

The experiment design is directly comparable to the original study of Bauer et al. We can therefore directly relate the experiment results with their study and deduce the influence of the data set characteristics on the approach's performance.

We point out two deviations in the experiment design:

- The sample size for COVID4 was reduced to 300 time series per randomization (compared to 500 time series for the other data sets) because the total number of time series in the COVID4 data set is only 313 (i.e., it is impossible to randomly draw 500 time series without replacement).
- We used 24 randomization runs for the experiments that used the OEM article data set to increase the mass of statistical evidence (see Section 7.4.4).

7.4 Experiment Results

This section describes the evaluation results of the experiments and is divided into five subsections:

1. **Before meta-learner:** the evaluation results of the basic pipeline (i.e., step B6), before the meta-learner is applied.
2. **Scope all:** the evaluation results of the entire experiment's sample of time series (both predicted "*stsf*"- and "*baseline*"-label time series, see Section 7.3.2).
3. **Scope predict "*stsf*":** the evaluation result of only predicted "*stsf*"-label time series.
4. **Comparison of evaluation and statistical significance:** detailed overview of all experiments, comparison of the calculated metrics, and significance test results.
5. **Meta-learner classification accuracy:** an overview of the classification accuracy of the meta-leaner.

Subsections 7.4.1 to 7.4.3 describe the evaluation results using the *RMSE* metric. The structure is identical in these three subsections.

7.4.1 Before meta-learner

Figure 26 and 4 and Table 43 show the results of the experiments after the completion of the basic pipeline step B6. This means that no meta-learn classification prediction has taken place up to this point and we are not looking at the "*all-stars*" predictions.

The systematic comparison of all results follows in Section 7.5. In this section, we briefly outline some remarkable observations.

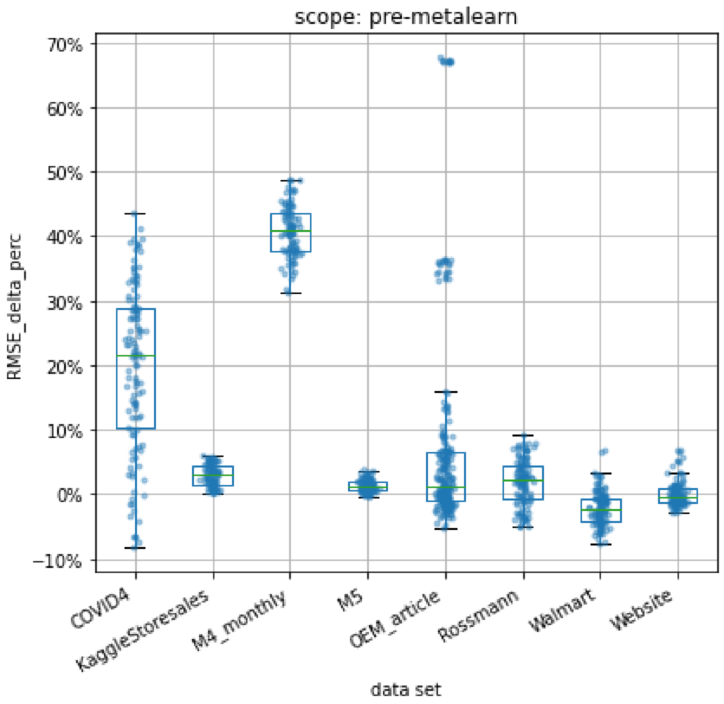


Figure 26: Overview of evaluation results per data set – before meta-learner

Mean RMSE_delta_perc per data set	
COVID4	19.67%
Kaggle store sales	2.92%
M4 monthly	40.77%
M5 competition	1.32%
OEM article	7.38%
Rossmann	1.92%
Walmart	-2.07%
Website traffic	0.10%

Table 43: Overview of mean RMSE_delta_perc per data set – before meta-learner

On the data set level, we observe three main particularities:

- The mean *RMSE_delta_perc* values per data set vary from -2% to 41%, and the mean *RMSE_delta_perc* values per experiment and randomization from -8% to 68%.
- Of all experiments and randomizations, we observe that in 70% of the cases, the *RMSE_delta_perc* is greater than zero, i.e., the mean result of the experiment is better when the STSF basic pipeline is applied.
- The results of the COVID4 data set show the greatest deviation by far compared to the other data sets – whereas especially the data sets *Kaggle store sales* and *M5* show little variation in the evaluation results.
- As a major difference from the results reported by Bauer et al., we observe that the mean evaluation results of all data sets (except *Walmart*) are positive. This means that, on average, the basic pipeline results of the STSF approach are better than their corresponding baselines. We conclude that the reason is the change of the increased time frame history of the target variable provided to the disaggregation algorithm in step B5 (as described in Section 7.3.2).
- In addition, we observe that all experiments and randomizations of the data sets *Kaggle store sales* and *M4 monthly* exhibit *RMSE_delta_perc* values greater than or equal to zero. Hence, the STSF approach appears to be better than the baseline in all experiments even before the application of the meta-learner.
- Both the mean and deviation of the evaluation results vary strongly between the data sets.

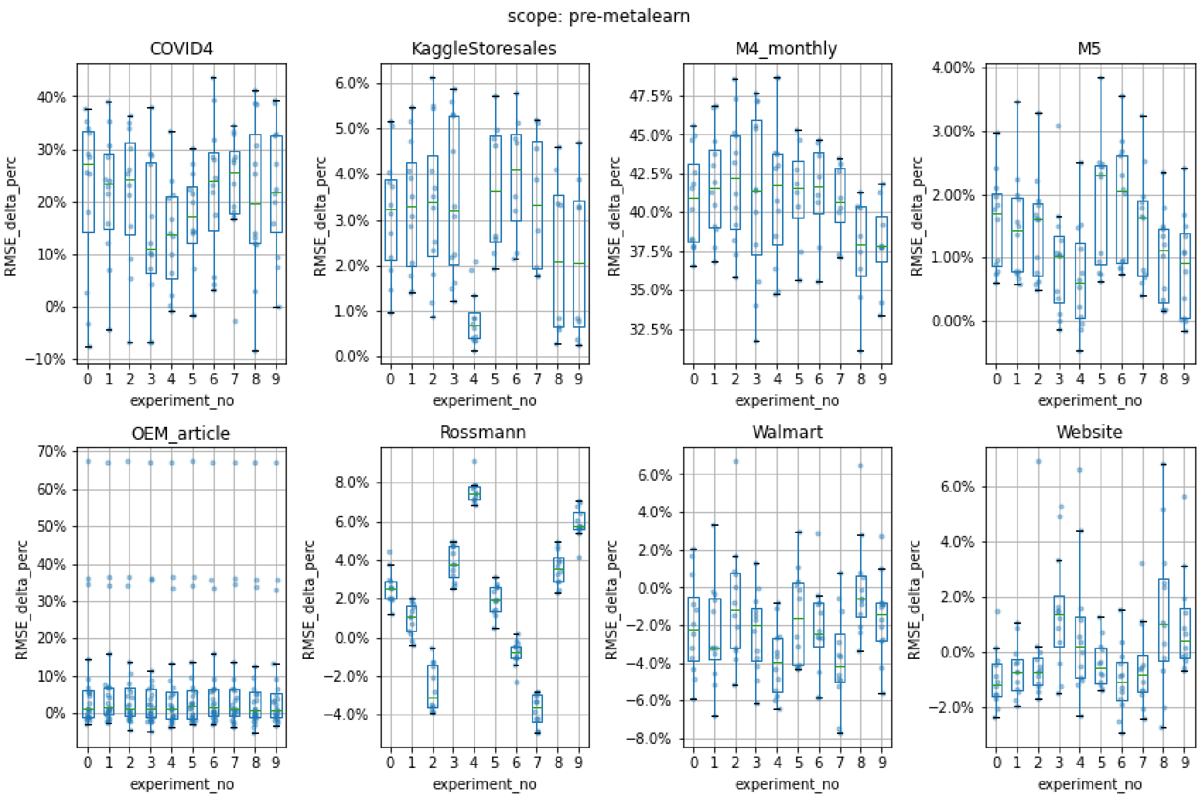


Figure 27: Overview of the evaluation results per data set and experiment – before meta-learner

While we cannot identify patterns in the results of many data sets when comparing the evaluation results on an experimental level, we find the following particularities in the *M5 competition*, *OEM article*, and *Rossmann* data sets:

- The evaluation results of the *M5 competition* exhibit descending *RMSE_delta_perc* values with an increasing number of clusters. We remember that experiments 0–4 and 5–9 exhibit the cluster number pattern: 3, 5, 10, 100, and 200. Hence, we can suspect a dependency between the number of clusters and evaluation results for the data set *M5 competition*.
- The *Rossmann* evaluation results also seem to exhibit a pattern with respect to the number of clusters. However, in this case, we observe a decrease from the cluster numbers 3 to 10 and an increase from 10 to 200.
- The previous results appear to indicate that an optimal sweet spot exists depending on the number of clusters. While we report the number of clusters here, we note that the actual dependency can rather be assumed in the number of elements per cluster, rather than the absolute number of clusters.
- We also observe that the results of the *OEM article* data set exhibit two randomizations where the evaluation results are considerably higher than the remaining randomizations.

7.4.2 Scope all

Figure 28, Figure 29 and Table 44 show the results of the complete STSF pipeline, including the application of the meta-learner. In this subsection, we consider the weighted *RMSE_delta_perc* results of all time series per data set – in contrast to the next subsection, where we concentrate on the time series predicted by the meta-learner under the label “stsf” (i.e., assumed to be better with STSF than baseline approach).

The main particularities on the data set evaluation level include:

- The mean *RMSE_delta_perc* values per data set vary from 1% to 43%, and the mean *RMSE_delta_perc* values per experiment and randomization from -3% to 51%.
- Of all experiments and randomizations, we observe that in 94% of the cases, the *RMSE_delta_perc* is greater than zero, i.e., the mean result of the experiment is better when the STSF approach is applied.
- All data sets' mean evaluation results have improved (i.e., are higher) compared to the pre-meta learner results in the previous subsection.
- The variations in the mean value patterns are similar to the previous subsection's results (on a visual level).
- The data sets with negative single experiment and randomization evaluation results in the previous subsection also exhibit negative values in the scope of all evaluations.

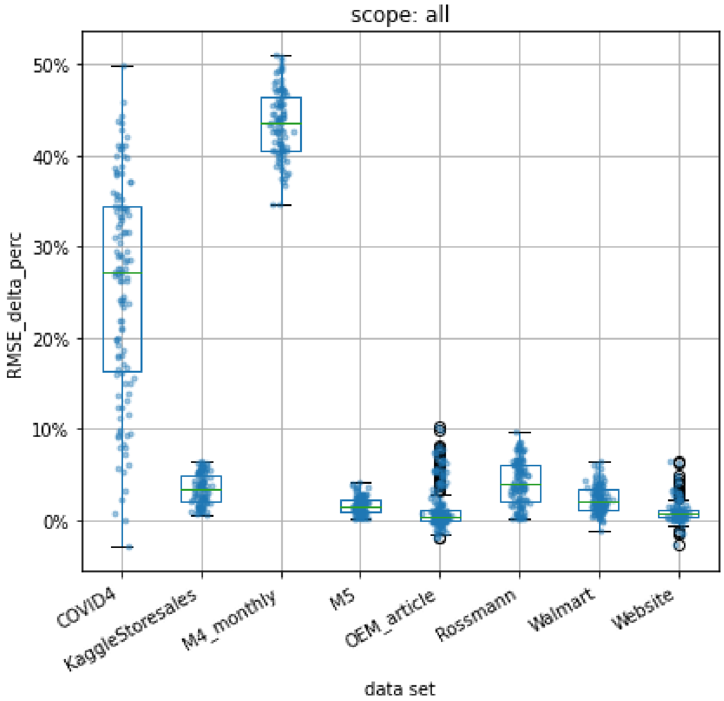


Figure 28: Overview of evaluation results per data set – scope all

Mean RMSE_delta_perc per data set	
COVID4	25.54%
Kaggle store sales	3.42%
M4 monthly	43.47%
M5 competition	1.60%
OEM article	1.31%
Rossmann	4.18%
Walmart	2.33%
Website traffic	1.02%

Table 44: Overview of mean RMSE_delta_perc per data set – scope all

On the experiment comparison level, we observe the following particularities compared to the previous subsection:

- The evaluation results of the *M5* and *Rossmann* data sets exhibit the same patterns that suggest a dependency between the number of clusters and the results.
- The OEM article's evaluation results do not exhibit particularly high values as seen in the pre-meta learner results.

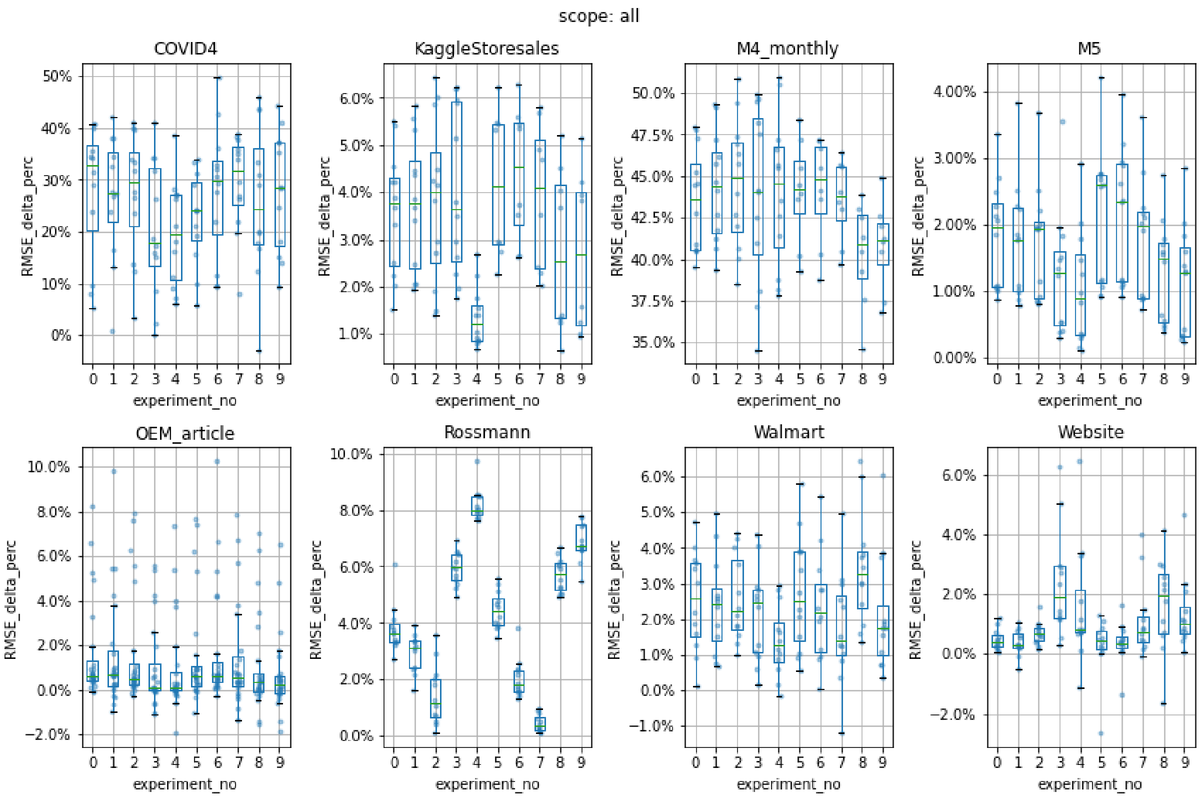


Figure 29: Overview of the evaluation results per data set and experiment – scope all

In addition to the usual evaluation results that we report in this study, Table 45 shows the *RMSE* evaluation results (as opposed to *RMSE_delta_perc*) to illustrate the absolute error levels of the predictions. It is important to note that the values are not suitable for comparison to other studies' results of the competitions, as this study does not aim to compete with the challenges but only to evaluate the STSF performance.

Mean RMSE evaluation result per data set for scope all, after application of the meta-leaner (step A6)								
	COVID4	Kaggle store sales	M4 monthly	M5 competition	OEM article	Rossmann	Walmart	Website traffic
STSF	1,284.2	274.0	1,082.2	2.0	30.9	1,115.4	5,118.5	3,875.9
Baseline	1,759.2	283.2	1,913.9	2.1	31.5	1,164.1	5,235.7	3,910.5

Table 45: Overview of the RMSE evaluation results per data set for scope "all evaluations" (as opposed to the RMSE_delta_perc values that are reported in the remainder of this study).

7.4.3 Scope predict "stsf"

This subsection concentrates on the *RMSE_delta_perc* evaluations of time series that the meta-learner model predicts to perform better under the STSF approach than the baseline approach. Figuratively speaking, this *"stsf"* prediction evaluates a subset of time series of the scope *"all"* evaluation.

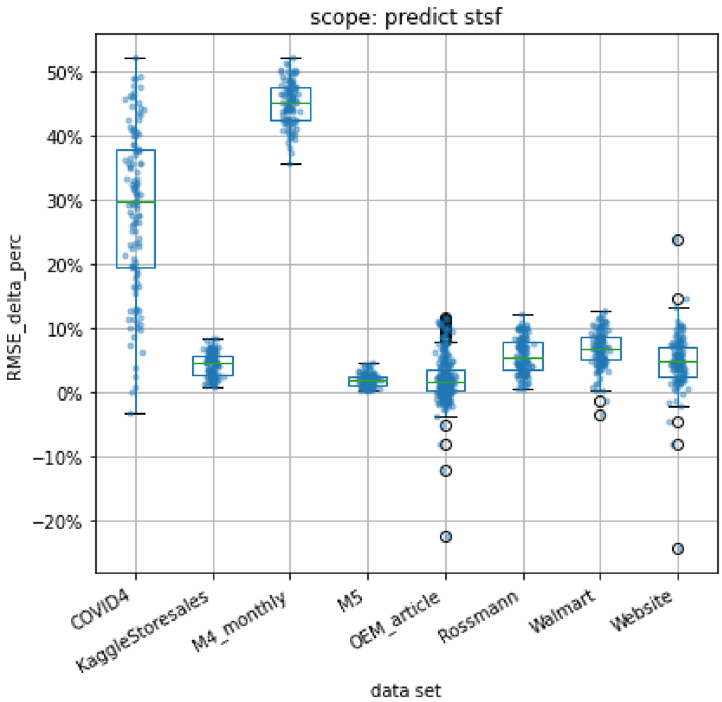


Figure 30: Overview of evaluation results per data set – scope predict "stsf"

Mean RMSE_delta_perc per data set	
COVID4	28.51%
Kaggle store sales	4.34%
M4 monthly	45.16%
M5 competition	1.79%
OEM article	2.13%
Rossmann	5.74%
Walmart	6.81%
Website traffic	4.75%

Table 46: Overview of mean RMSE_delta_perc per data set – scope "stsf"

On the data set evaluation level, we observe the following:

- The mean *RMSE_delta_perc* values per data set vary from 1% to 45%, and the mean *RMSE_delta_perc* values per experiment and randomization from -24% to 52%.
- Of all experiments and randomizations, we observe that in 94% of the cases, the *RMSE_delta_perc* is greater than zero, i.e., the mean result of the experiment is better when the STSF approach is applied.
- As expected, all data sets' mean evaluation results have improved (i.e., are higher) compared to the pre-meta learner results in the previous subsection (we study the subset of predicted "best" time series).
- The variations in the mean value patterns are similar to the previous subsections' results (on a visual level).
- The data sets with negative single experiment and randomization evaluation results in the previous subsection also exhibit negative values in the all evaluation scope.
- We observe a single experiment and randomization of the *Website* data set with a particularly low value.

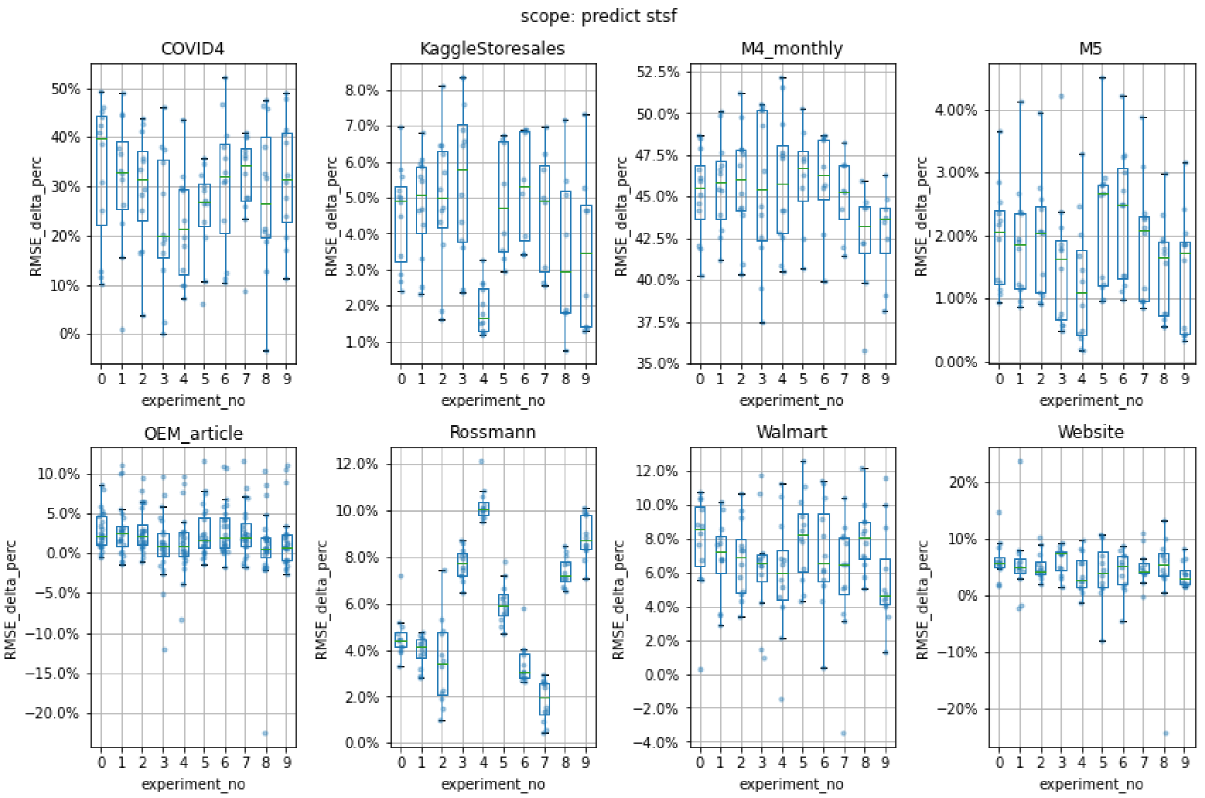


Figure 31: Overview of the evaluation results per data set and experiment – scope predict “stsf”

7.4.4 Comparison of the evaluation metrics and statistical significance

This section describes the overview of all experimental results evaluated with all six evaluation metrics. All results can be found in **Table 50** (see Section **7.8** “**Appendix**”). In addition, this section contains the description of the statistical significance tests to assess whether the STSF experiment results differ significantly from the baseline results. **Table 51** shows all results (see Section **7.8** “**Appendix**”).

We applied multiple hypothesis testing to show which experimental results differ significantly from the baseline results. Our H0 hypothesis is that the mean values of the experiments’ evaluation results do not differ from the corresponding baseline results. If the H0 hypothesis can be rejected at a high level of confidence, we conclude that the experiments and their baselines do differ significantly in their mean values.

We apply the pairwise t-test for mean values to each evaluation result pair of experiment and corresponding baseline whereby we are aware that multiple hypothesis testing results in alpha-inflation if not appropriately corrected. Therefore, we apply the *Benjamini-Hochberg procedure* to correct the p-values obtained from the pairwise t-test (Benjamini and Hochberg 1995). We require a *confidence level* of 95% and set the *false-detection rate* of the Benjamini-Hochberg procedure to 5%.

Share of rejected H0 hypothesis per data set and evaluation metric						
	MADmeanRatio	MAE	MASE	R2	RMSE	RMSSE
COVID4	100%	100%	70%	100%	100%	100%
Kaggle store sales	100%	100%	100%	100%	100%	100%
M4 monthly	100%	100%	100%	100%	100%	100%
M5 competition	100%	100%	100%	100%	100%	100%
OEM article	90%	90%	90%	70%	20%	60%
Rossmann	100%	100%	100%	100%	100%	100%
Walmart	100%	100%	100%	100%	100%	100%
Website	100%	80%	100%	90%	80%	70%

Table 47: Summary of the H0 hypothesis rejection rate based on the number of experiments rejected per data set and evaluation metric. The rejection is calculated under Benjamini-Hochberg corrected pairwise t-tests, with a 95% confidence level and 5% false-discovery rate.

Table 51 shows the detailed results of the hypothesis testing procedure whereby the entries in **Table 51** directly correspond to the entries in **Table 50** (both see Section **7.8** "**Appendix**"). Table 47 summarizes the results. We observe that almost 95% of the H0 hypotheses can be rejected under the defined parameters for p-value confidence level and false-discovery rate (i.e., the mean STSF approach evaluation result value is assumed to be significantly different from the baseline approach). This is especially remarkable as we compare 420 experiments with their respective baselines (i.e., 7 data sets x 10 experiments per data set x 6 evaluation metrics), and therefore the influence of the Benjamini-Hochberg correction is considerably high.

However, we remark that the H0 hypothesis for the *OEM article* data set cannot be rejected for the *RMSE* metric for all experiments except Experiments 1 and 6 (20% of experiments rejected) and similarly for some experiments evaluated under *RMSSE* (60%) and *R2* (70%). We suspect that apart from the obviously relatively low *RMSE_delta_perc values*, the comparably low accuracy rates of the meta-learner (see Section 7.4.5) account for the outcome of the significance tests. At this point, it is worth noting that we ran 24 randomizations for the *OEM article* data set, compared to the 12 randomizations used for the other data set. We chose this adjustment in the number of randomization runs to increase the number of observations to generate more evidence for the statistical tests.

We conclude that, except for the *OEM article* dataset, the results of all *RMSE_delta_perc* experiment evaluations can be regarded as significant under the selected premises.

7.4.5 Meta-learner classification accuracy

RQ4 asks whether it is possible to identify whether a time series is likely to benefit from the STSF approach (i.e., we observe a better evaluation result when STSF is applied compared to the baseline result). The meta-learner is trained to answer this question and thus we observe the classification accuracy of the meta-learner.

We define the *accuracy of the meta-learner* as the proportion of correctly identified labels (true-"*stsf*" and true-"*baseline*") of the total number of labels. Figure 32 shows the accuracy evaluation of all experiments and randomizations and the mean accuracy evaluations per data set are given in Table 48. We recall that separate meta-learner models are trained for each data set, so there is no overarching training of the meta-learner for all data sets.

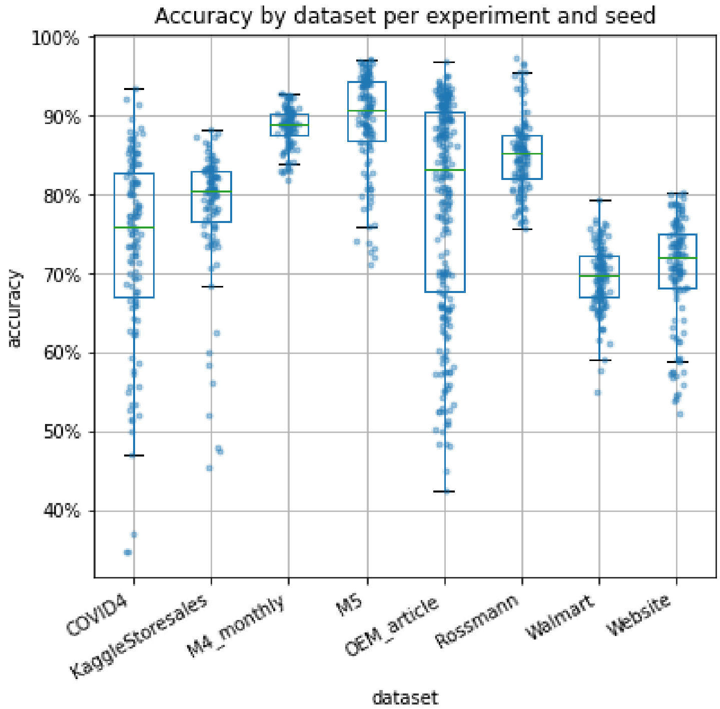


Figure 32: Accuracy of the meta-learner classification per data set denoted as a proportion of correctly identified time series of all time series. Each dot represents one experiment and randomization.

Mean meta-learner classification accuracy per data set	
COVID4	74%
Kaggle store sales	78%
M4 monthly	89%
M5 competition	89%
OEM article	79%
Rossmann	85%
Walmart	70%
Website traffic	70%

Table 48: Overview of the mean accuracy of the meta-learner classification per data set denoted as the proportion of correctly identified time series of all time series.

We can see that the average accuracy for all data sets is 70% or higher and we observe a mean accuracy score of 79% over all data sets. Especially *M4 monthly*, *M5 competition*, and *Rossmann* exhibit accuracy values of 89% and 85% respectively. However, Figure 32 shows that, apart from the mean value, the variance within each data set's accuracy results differs. Here, *COVID4* and *OEM article* show considerably high variance compared to the other data sets.

To further understand the accuracy of the meta-learner, we introduce Figure 33. It compares the *true-stsf* rate and *true-baseline* rate. The *true-stsf* rate denotes the share of rightfully predicted "*stsf*" time series of all actual "*stsf*" time series of one experiment and randomization (*true-baseline* rate analogously).

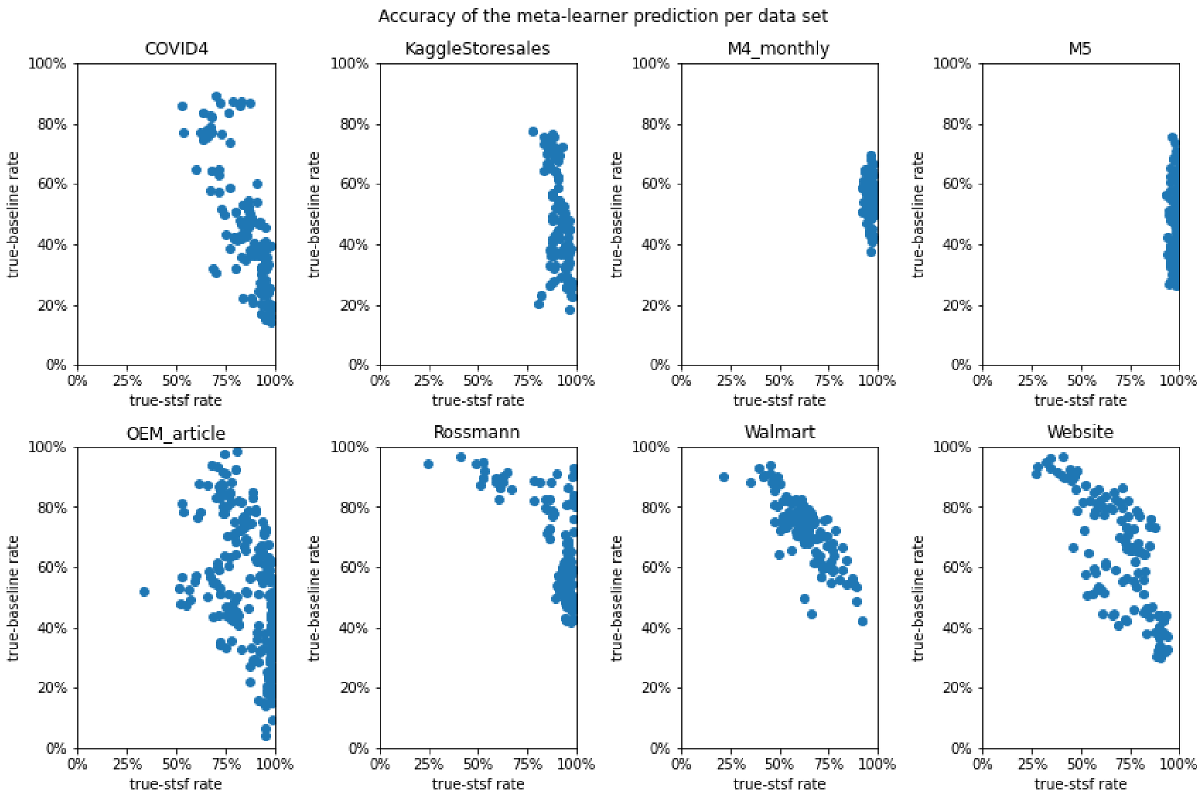


Figure 33: Detailed overview of the accuracy of the meta-learner per data set on an experiment and randomization level. "True-stsf rate" denotes the share of correctly true-stsf labeled time series of the actual stsf time series ("true-baseline rate" analogously for baseline labels).

We observe that the *true-stsf rate* is comparably high for most data sets. However, the *true-baseline rate* is rather low for most data sets. One should note that the "*stsf"* and "*baseline*" labels are imbalanced in the data sets. Across all data sets, 64% of time series are actual "*stsf*" and 36% are actual "*baseline.*" We will refer to this in the discussion in Section 6.6.

7.5 Summary of the Results

Here, we summarize the results from Section 7.4 and refer to the research questions from Section 6.1.

In **RQ1**, we raised the question of whether the STSF approach by Bauer et al. 2023b can improve forecasting results in different data sets than the two presented in the original study. We conclude that the results from Sections 7.4.1 to 7.4.3 indicate that the STSF approach improves the results in the additional six data sets in a range from 1% to 44% (mean *RMSE_delta_perc*, all time series) and hence we can confirm **RQ1**.

RQ2 posed the question of whether the results are consistent and statistically significant. From the results of Sections 7.4.1–7.4.3, we conclude that the results are consistent in the way that the STSF approach's mean evaluation results are always better than the forecast without the approach (i.e., baseline) on the mean data set evaluation level and in 94% of cases when evaluated per experiment and randomization. Nevertheless, the results indicate that not every experiment and randomization is always better than the baseline and that the variance differs between data sets and experiments. Thus, overall, we conclude that the results are consistent in the vast majority of cases.

Considering the statistical significance of the results, we conclude that the hypothesis testing indicates that 95% of the experiment results significantly differ from the baseline results at high levels of confidence. However, especially for the OEM article data set, several tests cannot be rejected.

Overall, for **RQ2** we conclude that the majority of experiments are consistent and statistically significant and are therefore suitable to support the findings from **RQ1**.

In **RQ3**, we asked in what ranges we can expect improvements from the STSF approach. Table 49 shows the summary of the results from the previous sections. We observe that most (6 out of 8) data sets exhibit mean *RMSE_delta_perc* values between approximately 1% and 5% whereby two data sets (*COVID4* and *M4 monthly*) exhibit fundamentally higher values (25% to over 40% for *scope all*). Hence, we conclude that one can usually expect improvements of 1%–5% in regular cases and higher rates of improvements under certain conditions. It is important to note that these numbers concern all time series in the respective random samples (i.e., scope all). When we refer to the time series predicted by the meta-learner to exhibit better performance when STSF is applied (i.e., scope predict-“*stsf*”), relative improvements for these time series are higher—typically 2% to 7% in the data sets in this study. However, this study does not conclude what conditions determine the relative improvement.

Summary of mean *RMSE_delta_perc* results per data set			
	Before meta-learner (step B6, Section 7.4.1)	Scope all (step A6, Section 7.4.2)	Scope predict “*stsf*”(step A6, Section 7.4.3)
COVID4	19.67%	25.54%	28.51%
Kaggle store sales	2.92%	3.42%	4.34%
M4 monthly	40.77%	43.47%	45.16%
M5 competition	1.32%	1.60%	1.79%
OEM article	7.38%	1.31%	2.13%
Rossmann	1.92%	4.18%	5.74%
Walmart	-2.07%	2.33%	6.81%
Website traffic	0.10%	1.02%	4.75%

Table 49: Summary of the mean RMSE_delta_perc values from the previous section.

Finally, **RQ4** discusses if we can predict whether a time series performs better when STSF is applied than when the approach is not applied (i.e., label “*stsf*” or “*baseline*”). As the accuracy scores in Section 7.4.5 show, the meta-learner can classify the labels “*stsf*” or “*baseline*” based on the time series characteristics at a mean accuracy of 79%, which is considerably above random guessing. We therefore can confirm the conclusion by Bauer et al. 2023b that the time series characteristics contain relevant information to deduce whether a time series benefits from the application of STSF. Further, we conclude that the representation approaches PD-statistics and Syntetos-Boylan statistics are suitable to capture and represent this information.

In addition to the conclusions directly related to the research questions, we summarize two additional findings from Section 7.4:

- Based on the pattern that we observe when comparing the experiments of the data sets M5 competition and Rossmann, we suspect that an optimal cluster size exists. However, we cannot confirm this assumption for the other data sets. Decreasing the increment between the number of clusters in the experiment design could help to confirm this assumption.
- In contrast to the original study by Bauer et al., we do not find a strong indication that especially intermittent time series benefit from the approach. From the findings, we cannot deduct a strong relationship between intermittency characteristics and the evaluation results.

7.6 Critical Discussion

In this section, we discuss the limitations of this study and highlight open questions for future research.

As described in the original study on STSF by Bauer et al., one should observe the general limitations of empirical studies which can only use a limited amount of data and experiments. This study contributes to broadening the data foundation of the original study and especially the time series characteristics used in this study are more diverse than the data in the original study. Nevertheless, other data sets could behave differently and help to gain new insights into the STSF.

We observe two directions for further studies on the STSF pipeline.

First, the proposed modular pipeline design allows researchers to flexibly replace algorithms in each step. We deliberately kept these algorithms and most of the parameters identical to the original study. However, we could observe that by adjusting the history window parameter of the disaggregation algorithm, the basic pipeline results could already be improved considerably. As a positive side effect, this helps to decrease the importance of the meta-learner algorithm to improve the results after the basic pipeline whereby this is especially helpful when the accuracy of the meta-learner is low for a particular data set for any reason.

In the context of improving the pipeline's algorithms and parameters, we observed that especially the time series representation approach is of great importance for the overall pipeline's performance. We suggest that experimenting with the time feature series extraction tools introduced in Section 7.2.2 can help to improve the approach. Extracting relevant features not only helps to produce clusters that aggregate information from similar generating processes in step B3, but also in training the meta-learner in step A4.

Second, a better understanding of the relationship between the data set characteristics, the forecasting accuracy, and the meta-learner accuracy is a promising focus for future research.

7.7 Contribution and Conclusions

In this study, we substantiate the empirical evidence that the STSF approach is suitable for improving time series forecasting when a hierarchy of the time series is assumed but not available or known. In six data sets in addition to the original data sets of Bauer et al., we show that we can predict what time series benefit from using the STSF approach and that the STSF approach yields improvements in the evaluation metric results of between 2% to 7% for these predicted time series. The chosen data sets are publicly available and the results reproducible. Moreover, by exhibiting a variety of time series characteristics, the data sets are well suited to show that the STSF approach applies to different types of time series. Therefore, we contribute to the research of time series forecasting by affirming that the STSFT approach helps to improve forecasting in various scenarios.

Moreover, this study shows promising approaches for future research on the STSF pipeline.

7.8 Appendix

7.8.1 Detailed evaluation results on the experiment level for all data sets and all evaluation metrics

Experiment No.	delta_perc for TS with predicted “*stsf*”-label					
	MADmeanRatio	MAE	MASE	R2	RMSE	RMSSE
COVID4						
0	20.5%	20.5%	8.0%	-0.5%	33.1%	17.9%
1	22.2%	22.2%	23.2%	-0.4%	31.2%	24.5%
2	22.2%	22.2%	23.5%	-0.4%	29.3%	23.2%
3	20.2%	20.2%	22.7%	-0.3%	22.9%	22.1%
4	16.8%	16.8%	15.9%	-0.3%	22.1%	20.0%
5	19.9%	19.9%	14.9%	-0.3%	25.1%	19.4%
6	22.1%	22.1%	9.8%	-0.4%	30.5%	20.7%
7	21.1%	21.1%	5.2%	-0.4%	31.4%	22.5%
8	22.5%	22.5%	24.4%	-0.4%	27.7%	25.7%
9	23.8%	23.8%	25.1%	-0.4%	31.7%	28.3%
Kaggle store sales						
0	3.0%	3.0%	3.1%	-0.5%	4.5%	4.0%
1	3.2%	3.2%	3.2%	-0.6%	4.8%	4.1%
2	3.2%	3.2%	3.2%	-0.6%	4.9%	4.1%
3	4.1%	4.1%	4.0%	-0.6%	5.5%	4.5%
4	1.8%	1.8%	1.7%	-0.2%	1.9%	1.8%
5	2.9%	2.9%	2.9%	-0.5%	4.9%	4.6%
6	3.1%	3.1%	3.0%	-0.6%	5.3%	4.8%
7	2.7%	2.7%	2.6%	-0.5%	4.6%	4.2%
8	2.6%	2.6%	2.6%	-0.4%	3.5%	3.1%
9	2.4%	2.4%	2.4%	-0.4%	3.5%	3.0%
M4 monthly						
0	53.3%	53.3%	59.9%	-37.7%	45.2%	53.9%
1	53.9%	53.9%	60.4%	-38.1%	45.8%	54.4%
2	54.3%	54.3%	61.0%	-38.2%	45.8%	54.2%
3	54.4%	54.4%	61.4%	-37.9%	45.6%	53.1%
4	55.0%	55.0%	62.3%	-38.3%	45.9%	54.0%
5	53.6%	53.6%	59.5%	-37.8%	46.1%	53.1%
6	53.6%	53.6%	59.2%	-37.8%	45.8%	52.6%
7	52.9%	52.9%	58.6%	-37.3%	45.0%	51.2%
8	50.0%	50.0%	56.0%	-35.7%	42.4%	47.7%
9	50.4%	50.4%	56.2%	-36.1%	42.8%	48.0%
M5 competition						
0	1.8%	1.8%	1.6%	-1.7%	2.0%	1.8%
1	1.8%	1.8%	1.7%	-1.7%	1.9%	1.7%
2	1.9%	1.9%	1.7%	-1.8%	1.9%	1.8%
3	2.2%	2.2%	2.2%	-1.5%	1.6%	1.4%
4	2.1%	2.1%	2.0%	-1.1%	1.2%	1.1%
5	1.8%	1.8%	1.7%	-1.9%	2.2%	2.1%
6	1.7%	1.7%	1.7%	-2.0%	2.3%	2.2%
7	1.4%	1.4%	1.4%	-1.7%	1.9%	1.8%
8	1.0%	1.0%	1.0%	-1.3%	1.5%	1.4%
9	1.0%	1.0%	0.9%	-1.4%	1.4%	1.3%
OEM article						
0	6.7%	6.7%	7.1%	-31.5%	2.9%	2.8%
1	7.0%	7.0%	7.6%	30.1%	2.9%	2.6%
2	6.4%	6.4%	6.7%	-34.0%	2.8%	2.8%
3	5.8%	5.8%	6.3%	78.0%	0.8%	0.4%
4	4.0%	4.0%	4.8%	5.0%	1.3%	0.7%
5	7.5%	7.5%	8.1%	-13.2%	2.6%	2.6%
6	7.4%	7.4%	8.1%	-5.1%	3.1%	2.8%
7	6.6%	6.6%	6.9%	-19.4%	2.6%	2.6%
8	6.3%	6.3%	7.5%	-197.5%	0.6%	0.6%
9	3.8%	3.8%	4.2%	-2.3%	1.6%	0.8%
Rossmann						
0	4.3%	4.3%	4.2%	-0.9%	4.6%	4.6%
1	3.8%	3.8%	3.8%	-0.8%	4.0%	3.9%
2	3.7%	3.7%	3.5%	-0.7%	3.5%	3.4%
3	8.0%	8.0%	8.0%	-1.4%	7.7%	7.9%
4	9.8%	9.8%	9.6%	-1.8%	10.2%	10.1%
5	4.8%	4.8%	4.8%	-1.1%	6.0%	6.0%

6	3.0%	3.0%	3.0%	-0.6%	3.4%	3.3%
7	2.4%	2.4%	2.4%	-0.4%	1.8%	1.8%
8	7.6%	7.6%	7.5%	-1.3%	7.4%	7.3%
9	8.3%	8.3%	8.2%	-1.6%	8.9%	8.7%
Walmart						
0	4.8%	4.8%	4.4%	-1.3%	7.8%	6.5%
1	4.4%	4.4%	4.3%	-0.9%	6.8%	5.8%
2	4.4%	4.4%	4.7%	-0.6%	6.8%	6.8%
3	4.7%	4.7%	5.3%	-0.5%	6.3%	6.4%
4	4.0%	4.0%	4.4%	-0.5%	5.9%	5.7%
5	5.0%	5.0%	4.7%	-1.2%	8.1%	7.0%
6	4.7%	4.7%	4.3%	-1.1%	7.1%	6.0%
7	3.9%	3.9%	4.0%	-1.1%	5.8%	6.0%
8	5.1%	5.1%	5.8%	-0.6%	8.0%	8.0%
9	4.5%	4.5%	5.1%	-0.4%	5.6%	6.5%
Website traffic						
0	3.2%	3.2%	3.3%	-15.5%	6.2%	7.4%
1	5.3%	5.3%	4.3%	-19.4%	5.7%	7.1%
2	3.3%	3.3%	3.5%	-17.5%	5.1%	7.4%
3	4.2%	4.2%	5.4%	-13.6%	6.1%	7.0%
4	3.2%	3.2%	3.7%	-5.3%	3.9%	5.0%
5	5.7%	5.7%	4.9%	-17.0%	4.1%	4.6%
6	4.1%	4.1%	3.2%	4.2%	4.2%	6.0%
7	4.0%	4.0%	3.6%	-14.8%	5.0%	6.0%
8	4.4%	4.4%	4.3%	-3.2%	3.6%	0.9%
9	3.4%	3.4%	3.6%	-7.9%	3.6%	4.8%

Table 50: Evaluation results of each experiment by all evaluation metrics denoted as _delta_perc. Recap: Experiments 0–4 with PD-statistics representation, experiments 5–9 with Syntetos-Boylan representation. Cluster sizes: 3, 5, 10, 100, and 200.

7.8.2 Overview of the result significance test

	p-value of paired t-test, Benjamini-Hochberg corrected						**H0 hypothesis rejected?**					
Metric → Exper-iment ↓	**MAD-mean-Ratio**	**MAE**	**MASE**	**R2**	**RMSE**	**RMSSE**	**MAD-mean-Ratio**	**MAE**	**MASE**	**R2**	**RMSE**	**RMSSE**
COVID4												
0	3.04E-05	5.17E-04	***2.24E-01***	1.44E-04	1.74E-04	1.25E-03	True	True	***False***	True	True	True
1	2.07E-05	1.29E-05	8.94E-03	2.08E-04	9.33E-05	8.58E-05	True	True	True	True	True	True
2	8.52E-06	2.07E-06	1.07E-02	1.16E-04	9.99E-05	4.02E-05	True	True	True	True	True	True
3	6.84E-05	2.42E-05	7.52E-03	1.65E-03	7.70E-04	2.08E-04	True	True	True	True	True	True
4	2.35E-05	1.70E-05	4.98E-02	1.85E-04	1.26E-04	2.56E-05	True	True	True	True	True	True
5	1.11E-06	4.35E-08	6.20E-03	7.77E-05	5.38E-06	4.59E-05	True	True	True	True	True	True
6	5.08E-05	1.69E-05	***1.54E-01***	1.91E-04	3.95E-05	2.78E-03	True	True	***False***	True	True	True
7	1.42E-07	1.25E-08	***2.66E-01***	2.56E-05	1.67E-05	1.67E-05	True	True	***False***	True	True	True
8	3.00E-05	1.70E-05	1.61E-02	6.85E-04	3.85E-04	9.04E-05	True	True	True	True	True	True
9	4.04E-06	2.39E-06	8.97E-03	7.53E-05	4.02E-05	1.70E-05	True	True	True	True	True	True
Kaggle store sales												
0	7.91E-07	8.58E-08	2.06E-06	1.56E-06	1.04E-07	4.12E-08	True	True	True	True	True	True
1	2.43E-07	3.14E-08	7.53E-07	2.18E-06	1.28E-07	2.55E-08	True	True	True	True	True	True
2	1.54E-06	2.11E-07	6.83E-06	1.07E-05	1.41E-06	6.53E-07	True	True	True	True	True	True
3	9.60E-08	7.90E-09	3.94E-07	1.03E-05	7.97E-07	3.44E-07	True	True	True	True	True	True
4	5.14E-07	1.02E-06	2.51E-06	3.46E-06	1.25E-06	4.88E-07	True	True	True	True	True	True
5	1.47E-05	3.94E-07	1.70E-05	1.31E-04	5.35E-06	4.08E-06	True	True	True	True	True	True
6	1.08E-05	2.50E-07	6.87E-06	6.51E-05	1.46E-06	5.46E-07	True	True	True	True	True	True
7	9.04E-05	1.40E-05	7.94E-05	2.72E-04	2.63E-05	2.02E-05	True	True	True	True	True	True
8	2.05E-04	4.58E-05	2.22E-04	4.12E-03	1.61E-03	1.12E-03	True	True	True	True	True	True
9	8.50E-04	2.81E-04	6.69E-04	3.46E-03	1.49E-03	9.72E-04	True	True	True	True	True	True
M4 monthly												
0	4.60E-13	2.08E-12	6.44E-10	3.53E-12	8.03E-13	7.27E-12	True	True	True	True	True	True
1	7.27E-13	2.66E-12	2.92E-10	5.53E-12	2.32E-12	8.56E-12	True	True	True	True	True	True
2	2.32E-12	7.60E-12	6.44E-10	7.93E-12	6.17E-12	5.37E-11	True	True	True	True	True	True
3	7.27E-12	3.88E-11	4.46E-10	2.52E-11	4.74E-11	7.57E-11	True	True	True	True	True	True
4	5.16E-12	2.29E-11	6.85E-10	7.70E-12	9.00E-12	2.52E-11	True	True	True	True	True	True
5	7.66E-09	1.54E-08	4.79E-07	2.30E-08	5.77E-09	1.04E-07	True	True	True	True	True	True
6	7.95E-09	1.59E-08	4.79E-07	3.08E-08	7.97E-09	1.14E-07	True	True	True	True	True	True
7	3.77E-09	1.23E-08	5.00E-07	1.12E-08	1.06E-09	5.74E-08	True	True	True	True	True	True
8	2.27E-08	5.24E-08	4.57E-07	2.90E-08	1.53E-08	1.02E-07	True	True	True	True	True	True
9	2.59E-09	8.56E-09	3.80E-07	1.76E-08	4.65E-09	1.20E-07	True	True	True	True	True	True

M5 competition												
0	2.51E-08	9.08E-06	6.92E-07	4.47E-06	1.46E-04	4.49E-05	True	True	True	True	True	True
1	1.25E-08	2.68E-05	4.98E-07	2.14E-05	4.14E-04	7.14E-05	True	True	True	True	True	True
2	2.51E-08	5.14E-06	6.53E-07	1.97E-05	3.11E-04	7.13E-05	True	True	True	True	True	True
3	2.45E-06	4.79E-07	2.53E-05	1.85E-04	3.22E-03	1.15E-03	True	True	True	True	True	True
4	7.16E-06	2.63E-07	7.05E-05	9.21E-04	4.86E-03	2.53E-03	True	True	True	True	True	True
5	3.71E-10	8.69E-06	1.24E-09	2.07E-05	2.98E-04	8.58E-05	True	True	True	True	True	True
6	6.85E-10	6.79E-06	1.25E-08	6.33E-06	2.01E-04	4.65E-05	True	True	True	True	True	True
7	1.95E-08	1.08E-05	7.22E-08	1.96E-05	3.42E-04	8.70E-05	True	True	True	True	True	True
8	1.67E-05	2.47E-04	4.15E-05	2.64E-05	5.66E-04	1.93E-04	True	True	True	True	True	True
9	1.21E-03	2.53E-03	8.48E-03	2.72E-04	1.96E-03	6.88E-04	True	True	True	True	True	True
OEM article												
0	1.32E-03	3.96E-03	2.03E-03	1.69E-05	***8.43E-02***	2.83E-03	True	True	True	True	***False***	True
1	2.70E-03	3.03E-03	6.73E-03	1.20E-03	4.76E-02	1.69E-02	True	True	True	True	True	True
2	3.37E-03	3.25E-03	5.69E-03	8.84E-05	***6.96E-02***	1.08E-02	True	True	True	True	***False***	True
3	6.64E-03	6.31E-03	9.68E-03	***4.64E-01***	***9.27E-02***	***2.65E-01***	True	True	True	***False***	***False***	***False***
4	***6.14E-02***	***1.24E-01***	***5.02E-02***	***1.46E-01***	***7.16E-02***	***1.20E-01***	***False***	***False***	***False***	***False***	***False***	***False***
5	7.97E-04	8.58E-05	1.33E-03	1.19E-03	***1.46E-01***	4.88E-03	True	True	True	True	***False***	True
6	3.96E-03	7.79E-04	5.69E-03	4.05E-04	3.35E-02	3.29E-03	True	True	True	True	True	True
7	4.83E-03	2.14E-03	8.50E-03	1.74E-03	***5.31E-02***	1.30E-02	True	True	True	True	***False***	True
8	2.63E-03	3.91E-03	3.87E-03	***7.93E-01***	***1.59E-01***	***2.23E-01***	True	True	True	***False***	***False***	***False***
9	2.73E-02	4.19E-02	4.69E-02	4.98E-02	***1.13E-01***	***2.73E-01***	True	True	True	True	***False***	***False***
Rossmann												
0	5.22E-08	5.94E-08	5.70E-08	2.24E-08	1.39E-08	1.25E-08	True	True	True	True	True	True
1	1.46E-09	1.53E-09	2.34E-09	2.34E-09	3.57E-09	2.34E-09	True	True	True	True	True	True
2	7.77E-05	7.79E-05	1.04E-04	8.17E-05	6.75E-05	7.50E-05	True	True	True	True	True	True
3	3.53E-13	3.65E-13	3.53E-13	7.27E-12	5.94E-12	3.14E-12	True	True	True	True	True	True
4	4.60E-13	7.27E-13	3.88E-13	1.05E-12	2.57E-12	2.32E-12	True	True	True	True	True	True
5	1.76E-09	2.07E-09	2.23E-09	7.13E-10	8.24E-10	8.24E-10	True	True	True	True	True	True
6	1.65E-06	2.29E-06	1.82E-06	2.20E-07	9.46E-08	1.20E-07	True	True	True	True	True	True
7	2.76E-06	2.26E-06	5.99E-06	4.49E-05	2.50E-05	2.71E-05	True	True	True	True	True	True
8	2.32E-12	2.32E-12	2.15E-12	7.70E-12	5.31E-12	5.81E-12	True	True	True	True	True	True
9	2.86E-12	2.86E-12	3.19E-12	2.29E-11	2.50E-11	8.54E-12	True	True	True	True	True	True
Walmart												
0	1.21E-06	2.85E-07	7.00E-07	4.15E-05	2.06E-05	6.05E-07	True	True	True	True	True	True
1	1.15E-05	2.07E-06	7.68E-06	8.30E-05	2.85E-06	6.82E-07	True	True	True	True	True	True
2	5.38E-06	8.86E-07	2.71E-06	4.89E-05	4.81E-06	8.02E-08	True	True	True	True	True	True
3	1.20E-07	1.87E-07	3.11E-07	5.66E-04	1.09E-04	2.29E-08	True	True	True	True	True	True
4	3.33E-06	4.34E-06	3.75E-06	8.50E-04	2.72E-04	1.36E-05	True	True	True	True	True	True
5	5.36E-07	1.42E-07	4.04E-07	6.49E-05	8.99E-06	4.17E-08	True	True	True	True	True	True
6	7.97E-07	2.51E-08	1.06E-06	6.41E-04	4.98E-05	8.40E-06	True	True	True	True	True	True
7	1.69E-05	9.84E-06	7.16E-06	1.37E-03	4.36E-04	4.25E-06	True	True	True	True	True	True
8	2.43E-08	3.28E-08	7.95E-08	4.34E-06	3.98E-07	1.22E-08	True	True	True	True	True	True
9	4.35E-08	1.28E-07	1.36E-07	4.59E-04	2.93E-04	8.04E-08	True	True	True	True	True	True
Website												
0	3.92E-04	3.76E-03	4.84E-04	9.95E-05	7.53E-05	1.15E-02	True	True	True	True	True	True
1	1.31E-02	***1.93E-01***	1.69E-02	4.30E-02	2.82E-02	***1.10E-01***	True	***False***	True	True	True	***False***
2	2.92E-03	2.38E-03	3.35E-03	1.99E-04	1.44E-04	4.24E-03	True	True	True	True	True	True
3	1.46E-04	1.04E-03	1.33E-05	3.59E-05	4.14E-05	2.94E-05	True	True	True	True	True	True
4	2.06E-04	4.29E-04	5.41E-06	5.10E-03	1.84E-02	4.55E-04	True	True	True	True	True	True
5	7.56E-03	***1.74E-01***	6.94E-03	6.88E-03	***1.55E-01***	***6.48E-02***	True	***False***	True	True	***False***	***False***
6	1.07E-03	2.24E-04	2.36E-03	3.05E-03	2.43E-02	1.51E-02	True	True	True	True	True	True
7	3.01E-03	7.06E-03	3.34E-03	7.19E-04	1.08E-04	1.00E-06	True	True	True	True	True	True
8	1.20E-05	6.90E-04	1.67E-05	***5.02E-01***	***3.17E-01***	***9.61E-01***	True	True	True	***False***	***False***	***False***
9	3.58E-04	3.05E-03	1.93E-02	1.36E-04	3.40E-04	2.72E-04	True	True	True	True	True	True

Table 51: Hypothesis testing of the statistical significance of different means of experiments and baselines. The table shows the Benjamini-Hochberg corrected p-values and H0 rejection und Benjamini-Hochberg procedure at a confidence level of 95% and false-detection rate of 5%. Values in bold and italics indicate that the H0 hypothesis cannot be rejected at the given confidence levels. The table directly corresponds to Table 50. Recap: Experiments 0–4 with PD-statistics representation, experiments 5–9 with Syntetos-Boylan representation. Cluster sizes: 3, 5, 10, 100, and 200.

7.9 References

Aghabozorgi, Saeed; Seyed Shirkhorshidi, Ali; Ying Wah, Teh (2015): Time-series clustering – A decade review. In *Information Systems* 53, pp. 16–38. DOI: 10.1016/j.is.2015.04.007.

Athanasopoulos, George; Ahmed, Roman A.; Hyndman, Rob J. (2009): Hierarchical forecasts for Australian domestic tourism. In *International Journal of Forecasting* 25 (1), pp. 146–166. DOI: 10.1016/j.ijforecast.2008.07.004.

Athanasopoulos, George; Hyndman, Rob J.; Kourentzes, Nikolaos; Petropoulos, Fotios (2017): Forecasting with temporal hierarchies. In *European Journal of Operational Research* 262 (1), pp. 60–74. DOI: 10.1016/j.ejor.2017.02.046.

Babai, M. Zied; Boylan, John E.; Rostami-Tabar, Bahman (2022): Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. In *International Journal of Production Research* 60 (1), pp. 324–348. DOI: 10.1080/00207543.2021.2005268.

Bandara, Kasun; Bergmeir, Christoph; Smyl, Slawek (2020): Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. In *Expert Systems with Applications* 140, p. 112896. DOI: 10.1016/j.eswa.2019.112896.

Barandas, Marília; Folgado, Duarte; Fernandes, Letícia; Santos, Sara; Abreu, Mariana; Bota, Patrícia et al. (2020): TSFEL: Time Series Feature Extraction Library. In *SoftwareX* 11, p. 100456. DOI: 10.1016/j.softx.2020.100456.

Bauer, Markus; van Dinther, Clemens; Grimm, Florian; Kiefer, Daniel (2023a): Evaluating the Similarity-based Time Series Forecasting Approach: Generalization of the Results.

Bauer, Markus; van Dinther, Clemens; Kiefer, Daniel; Grimm, Florian (2023b): Forecasting Intermittent Demand with no given External Hierarchy: An Aggregation-Disaggregation Approach based on Clustering of Time Series Characteristics Similarity.

Benjamini, Yoav; Hochberg, Yosef (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. In *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bergmeir, Christoph; Benítez, José M. (2012): On the use of cross-validation for time series predictor evaluation. In *Information Sciences* 191, pp. 192–213. DOI: 10.1016/j.ins.2011.12.028.

Christ, Maximilian; Braun, Nils; Neuffer, Julius; Kempa-Liehr, Andreas W. (2018): Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). In *Neurocomputing* 307, pp. 72–77. DOI: 10.1016/j.neucom.2018.03.067.

de Gooijer, Jan G.; Hyndman, Rob J. (2006): 25 years of time series forecasting. In *International Journal of Forecasting* 22 (3), pp. 443–473. DOI: 10.1016/j.ijforecast.2006.01.001.

Eckert, Florian; Hyndman, Rob J.; Panagiotelis, Anastasios (2021): Forecasting Swiss exports using Bayesian forecast reconciliation. In *European Journal of Operational Research* 291 (2), pp. 693–710. DOI: 10.1016/j.ejor.2020.09.046.

Gür, İzzeddin; Güvercin, Mehmet; Ferhatosmanoglu, Hakan (2015): Scaling forecasting algorithms using clustered modeling. In *The VLDB Journal* 24 (1), pp. 51–65. DOI: 10.1007/s00778-014-0363-0.

Hartomo, Kristoko Dwi; Nataliani, Yessica (2021): A new model for learning-based forecasting procedure by combining k-means clustering and time series forecasting algorithms. In *PeerJ. Computer science* 7, e534. DOI: 10.7717/peerj-cs.534.

Hollyman, Ross; Petropoulos, Fotios; Tipping, Michael E. (2021): Understanding forecast reconciliation. In *European Journal of Operational Research* 294 (1), pp. 149–160. DOI: 10.1016/j.ejor.2021.01.017.

Holt, Charles C. (2004): Forecasting seasonals and trends by exponentially weighted moving averages. In *International Journal of Forecasting* 20 (1), pp. 5–10. DOI: 10.1016/j.ijforecast.2003.09.015.

Hyndman, Rob J. (2020): A brief history of forecasting competitions. In *International Journal of Forecasting* 36 (1), pp. 7–14. DOI: 10.1016/j.ijforecast.2019.03.015.

Hyndman, Rob J.; Ahmed, Roman A.; Athanasopoulos, George; Shang, Han Lin (2011): Optimal combination forecasts for hierarchical time series. In *Computational Statistics & Data Analysis* 55 (9), pp. 2579–2589. DOI: 10.1016/j.csda.2011.03.006.

Hyndman, Rob J.; Koehler, Anne B. (2006): Another look at measures of forecast accuracy. In *International Journal of Forecasting* 22 (4), pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Hyndman, Rob J.; Wang, Earo; Laptev, Nikolay Pavlovich (2015): Large-Scale Unusual Time Series Detection. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1616–1619.

Ke, Guolin; Meng, Qi; Finley, Thomas; Wang, Taifeng; Chen, Wei; Ma, Weidong et al. (2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.): Advances in Neural Information Processing Systems, vol. 30: Curran Associates, Inc. Available online at https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kolassa, Stephan (2016): Evaluating predictive count data distributions in retail sales forecasting. In *International Journal of Forecasting* 32 (3), pp. 788–803. DOI: 10.1016/j.ijforecast.2015.12.004.

Kolassa, Stephan (2020): Why the “best” point forecast depends on the error or accuracy measure. In *International Journal of Forecasting* 36 (1), pp. 208–211. DOI: 10.1016/j.ijforecast.2019.02.017.

Kolassa, Stephan; Schütz, Wolfgang; others (2007): Advantages of the MAD/MEAN ratio over the MAPE. In *Foresight: The International Journal of Applied Forecasting* (6), pp. 40–43.

Laurinec, Peter; Loderer, Marek; Vrablecova, Petra; Lucka, Maria; Rozinajova, Viera; Ezzeddine, Anna Bou (2016): Adaptive Time Series Forecasting of Energy Consumption Using Optimized Cluster Analysis. In : 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). Barcelona, Spain, 12.12.2016 - 15.12.2016: IEEE, pp. 398–405.

Laurinec, Peter; Lóderer, Marek; Lucká, Mária; Rozinajová, Viera (2019): Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption. In *J Intell Inf Syst* 53 (2), pp. 219–239. DOI: 10.1007/s10844-019-00550-3.

Mahalakshmi, G.; Sridevi, S.; Rajaram, S. (2016): A survey on forecasting of time series data. In : 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE). Kovilpatti, India, 07.01.2016 - 09.01.2016: IEEE, pp. 1–8.

Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios (2022): M5 accuracy competition: Results, findings, and conclusions. In *International Journal of Forecasting* 38 (4), pp. 1346–1364. DOI: 10.1016/j.ijforecast.2021.11.013.

Martinez Alvarez, Francisco; Troncoso, Alicia; Riquelme, Jose C.; Aguilar Ruiz, Jesus S. (2011): Energy Time Series Forecasting Based on Pattern Sequence Similarity. In *IEEE Trans. Knowl. Data Eng.* 23 (8), pp. 1230–1243. DOI: 10.1109/TKDE.2010.227.

Nystrup, Peter; Lindström, Erik; Pinson, Pierre; Madsen, Henrik (2020): Temporal hierarchies with autocorrelation for load forecasting. In *European Journal of Operational Research* 280 (3), pp. 876–888. DOI: 10.1016/j.ejor.2019.07.061.

Pang, Yue; Yao, Bo; Zhou, Xiangdong; Zhang, Yong; Xu, Yiming; Tan, Zijing (2018): Hierarchical Electricity Time Series Forecasting for Integrating Consumption Patterns Analysis and Aggregation Consistency. In : Proceedings of the 27th International Joint Conference on Artificial Intelligence: AAAI Press (IJCAI'18), pp. 3506–3512.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. et al. (2011): Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pelleg, Dan; Moore, Andrew (2002): X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Machine Learning, p.*

Pennings, Clint L.P.; van Dalen, Jan (2017): Integrated hierarchical forecasting. In *European Journal of Operational Research* 263 (2), pp. 412–418. DOI: 10.1016/j.ejor.2017.04.047.

Petropoulos, Fotios; Makridakis, Spyros; Assimakopoulos, Vassilios; Nikolopoulos, Konstantinos (2014): 'Horses for Courses' in demand forecasting. In *European Journal of Operational Research* 237 (1), pp. 152–163. DOI: 10.1016/j.ejor.2014.02.036.

Syntetos, A. A.; Boylan, J. E.; Croston, J. D. (2005): On the categorization of demand patterns. In *Journal of the Operational Research Society* 56 (5), pp. 495–503. DOI: 10.1057/pal-grave.jors.2601841.

Tavenard, Romain; Faouzi, Johann; Vandewiele, Gilles; Divo, Felix; Androz, Guillaume; Holtz, Chester et al. (2020): Tslearn, a machine learning toolkit for time series data. In *J. Mach. Learn. Res.* 21 (118), pp. 1–6.

Venkatesh, Kamini; Ravi, Vadlamani; Prinzie, Anita; van den Poel, Dirk (2014): Cash demand forecasting in ATMs by clustering and neural networks. In *European Journal of Operational Research* 232 (2), pp. 383–392. DOI: 10.1016/j.ejor.2013.07.027.

Wickramasuriya, Shanika L.; Athanasopoulos, George; Hyndman, Rob J. (2019): Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. In *Journal of the American Statistical Association* 114 (526), pp. 804–819. DOI: 10.1080/01621459.2018.1448825.

Winters, Peter R. (1960): Forecasting Sales by Exponentially Weighted Moving Averages. In *Management Science* 6 (3), pp. 324–342. DOI: 10.1287/mnsc.6.3.324.

Zotteri, Giulio; Kalchschmidt, Matteo; Caniato, Federico (2005): The impact of aggregation level on forecasting performance. In *International Journal of Production Economics* 93-94, pp. 479–491. DOI: 10.1016/j.ijpe.2004.06.044.

8 Conclusion

In this thesis, I (1) introduced to the topic of demand forecasting, (2) systematically studied the current state-of-the-art in research and practice as well as current open issues, (3) demonstrated the application of demand forecasting in general and intermittent time series forecasting in particular, and (4) introduced a novel approach to forecasting demand for unknown hierarchical structures of time series.

This chapter concludes the thesis by summarizing the contributions of the previous chapters, answering the research questions posed in Chapter 1.1 and providing an outlook for future research.

8.1 Summary of the results

In Chapter 1.1, I raised the following research questions as guidelines for my studies:

RQ I. What is the current state of demand forecasting in research, what are open issues discussed in literature and how far has adoption of the current state of the art proceeded in companies?

RQ II. How can demand forecasting studies be structured to increase the comparability between studies?

RQ III. What is the forecasting performance of machine learning methods compared to classical approaches for intermittent time series and how can approaches be selected depending on the time series characteristics?

RQ IV. How can intermittent and hierarchical demand forecasting be improved?

Chapter 2 *(On the Industry Need for Machine Learning and Demand Forecasting*) and **Chapter 3** (*How the Demand Forecasting Literature and Applications can Benefit from Better Comparability*) together contribute to research question **RQ I.** I showed in Chapter 2 that machine learning approaches in general are well researched, yet the prevalence of the implementation of these approaches in companies is relatively low. In this chapter, I showed that the success of the adoption of machine learning approaches is connected to the maturity of the companies in the dimensions (a) machine learning know how, (b) available personnel capacities, (c) availability of data, (d) acceptance of machine learning technologies by the stakeholders and (e) capability for interdisciplinary cooperation. In Chapter 3, I showed in detail the methods applied in demand forecasting literature–comparing the major fields of application in demand forecasting and at the same time along the generalized process of demand forecasting. I could thereby show that the research in the different fields of applications could benefit from adopting of approaches of the respective other fields as well as from adopting approaches from other disciplines in forecasting. Moreover, Chapter 3 summarized the open issues stated in the literature, concerning data availability, comparability of studies mainly in input data and evaluation methods, and development of new forecasting algorithms, especially leveraging hybridization of existing approaches.

The findings concerning RQI. from these chapters are thus relevant for this dissertation as well as for research. I was able to show that there is great interest in applications for machine learning and demand forecasting in practice - for which many approaches have already been explored in research. However, I was able to show that further influencing factors are needed to introduce these into practice.

Furthermore, I could show that similar open research questions exist in the application areas in research. Nevertheless, the exchange and comparison between the application fields in research is still too low and progress could concretely be made by a better exchange between the research fields.

Chapter 3 in addition contributes to research question **RQ II.** The chapter introduces the systematic approach to describe demand forecasting studies in a structured way. As shown with the 116 studies examined, this allows studies to be described uniformly in the process steps of (1) forecasting goal definition, (2) data preparation, (3) feature engineering, and (4) model design, making them comparable for other researchers. The systematic thereby includes the description of the data input, the prediction goal, the approach and methods used, and the evaluation results. The introduction of this new systematic represents a substantial new contribution to research. Because the systematic makes it possible, as described, to create comparative studies that directly address the results of other studies without empirically reproducing them. This advantage enables the scientific community to gain new insights and make faster progress. Thus, the finding to research question RQ II. represents one of the most substantive contributions of this dissertation.

One of the most frequently cited open research questions in the literature in Chapter 3 was how to better use external influences to improve demand forecasting. In **Chapter 4** (*Developing an Understanding of External Factors Influencing Demand Forecasting Models using a Case Example*), I provided an example of how demand forecasting can be applied and the impact of training with external influences. The chapter contributed to research by showing how demand forecasting can be used to predict the impact of specific events and how models must be trained to relate past events to the impact on the target variable to find analogies in future patterns. In my literature search, I did not find any comparable work that similarly systematically explores the relationship of external influences, exceptional events, and machine learning training parameters on demand forecasts.

Research question **RQ III.** is answered in **Chapter 5** (*How Time Series Characteristics Affect the Forecast Quality in State-of-the-Art Algorithms for Intermittent Demands*). The three studies together show that the tested state of the art machine learning algorithms do not systematically produce better prediction results than the comparable classical algorithms for intermittent time series forecasting. Thus, the chapter already contributes by systematically comparing the state-of-the-art algorithms. In addition, the chapter highlights a relevant new aspect. Based on the properties of time series, the best method can be systematically chosen from a selection of forecasting methods using classification algorithms. Thus, the chapter shows not only that a relationship between time series properties must exist, but also how this relationship can be predicted. The findings from this chapter thus contribute to research by enabling researchers and practitioners to systematically select the best of the available methods for demand forecasting, depending on what types of time series are available. This can reduce the time and resources required.

Chapter 6 (*A New Approach in Hierarchical Demand Forecasting*) introduced to the special sub-discipline of hierarchical and intermittent demand forecasting. I showed in the chapter that so far there is no research on forecasting of hierarchical organized demands when the hierarchy structure is unknown–especially not in combination with intermittent time series characteristics. I closed the gap in research by introducing the novel similarity-based time series forecasting (STSF) approach. I showed in Chapter 6 that the approach is able to improve forecasting accuracy by 2-7% (based on the RMSE evaluation metric) in the empirical study incorporating two publicly available data sets. Thus, this chapter makes an important contribution to the research by providing a modular framework ready to apply and agnostic to the particular algorithms used for the empirical study.

Together with the findings from **Chapter 7** *(Generalization of the Approach's results)*, the two chapters give an answer to research question **RQ IV.** The STSF approach is an appropriate approach to improve hierarchical demand forecasting. The extended foundation of six additional data sets in Chapter 7 substantiates the empirical findings from the study in Chapter 6. Moreover, I conclude from the findings in Chapter 7 that the approach is beneficial to be applied to all sorts of time series, not only in particular to intermittent time series. In addition, the chapter contributes to the research by providing indications what kind of data sets benefit most from the STSF approach and how future research can further develop the framework.

Thus, the findings related to research question RQ IV. from Chapters 6 and 7 represent one of the most significant contributions of this dissertation to the state of the art. In summary, the STSF framework can reliably (for more than 94% of the time series tested) improve demand forecasts relevantly (between 1-40% vs. baseline depending on the dataset) for all types of time series tested in the experiments. The statements from the two chapters are supported by the confirmations of the statistical significance tests.

8.2 Critical discussion

The previous chapters contained critical discussion sections where the respective studies were scrutinized in detail. In this final chapter, this thesis summarizes the main aspects of the critical discussions into a general view and discuss major critical points of the thesis. These points are finally the foundation for Chapter 8.3 (*Outlook*).

Chapter 3 (*How the Demand Forecasting Literature and Applications can Benefit from Better Comparability*) has summarized the open issues in demand forecasting literature, e.g., new challenges of processing input data (especially big data and data quality), development of new (hybrid) forecasting algorithms and the issue of missing comparability of studies in terms of data used, methods applied and the standardization of evaluation of results. Apparently, this thesis cannot provide answers to all the questions of the research, wherefore I concentrated mainly on diving into two aspects: The development of a framework to standardize comparability of demand forecasting studies and the development of a new approach to improve demand forecasting in the absence of known time series hierarchies.

In this critical review, it should be noted, that Chapter 3 describes a suitable, comprehensive, and novel framework to systematically compare demand forecasting studies. However, at this point of time, it comes with two shortcomings. First, the novel approach is not yet applied by studies to ensure comparability to other studies. Therefore, it is so far only a descriptive framework for the demand forecasting process and a synthesis of the structures described in the research literature–it has yet to be implemented in research. Second, because the framework has not yet been applied, this thesis could only provide a qualitative overview of the state-of-the-art–as opposed to a systematic quantitative meta-analysis that compares studies' evaluation results based on the input data and methods applied.

We also critically review the findings of the novel STSF approach, described in the chapters 6 and 7. With the study of Chapter 6 alone, one would rightfully criticize that an empiric study based on only two data sets does not provide broad foundation and could be the result of lucky coincidence or carefully chosen particular data sets. However, Chapter 7 confirms the findings on additional six data sets, making this objection unlikely. Moreover, the data sets are commonly used in demand forecasting literature and not particularly known for any unusual characteristics. In addition, the experiment design was chosen such that lucky sampling within the data sets is unlikely, as shown by the statistical tests.

However, the studies remain empirical research. The findings from the empirical experiments strongly support the hypothesis that the STSF approach considerably improves demand forecasting under the given conditions (time series with a hierarchical structure that is not known). However, I am aware that the thesis does not provide the theoretical foundation to make general proven statements.

Also, the two studies in Chapters 6 and 7 support the assumption that there is a definable dependency between the characteristics of the input data, the applied STSF pipeline including its parameters and the improvement by the approach. The meta-learner that is part of the STSF pipeline showed high accuracy scores in classifying time series that benefit from the STSF approach. It can be concluded that there is exactly the dependence between time series characteristics and forecasting accuracy. However, this dependence is only defined implicitly in the meta-learner–the thesis does not make the dependence explicit.

8.3 Outlook

This final Sub-chapter of the thesis provides an outlook for future developments in demand forecasting and especially opportunities for further studies based on the proposed approaches. This outlook is oriented on the previous critical discussion.

Chapter 2 showed that there is a gap between the research of machine learning methods and their application in companies. I showed what success factors provide chances for companies to implement machine learning, amongst which especially demand forecasting applications. I conclude that the research community can support the advancement of applications by providing use-case specific frameworks and approaches that are easy to implement. In turn, it is safe to presume that an accelerated implementation of approaches in practice will lead to beneficial circumstances for research (e.g., availability of data, private and public funding, new use cases for research, etc.).

The open issues in demand forecasting summarized and derived from the literature review in Chapter 3 all pose opportunities for future research, especially as also named in the previous Sub-chapter: new challenges of processing input data (especially big data and data quality), development of new (hybrid) forecasting algorithms and the issue of missing comparability of studies in terms of data used, methods applied and the standardization of evaluation of results. Literature has unanimously identified these open issues as of high concern for the future development of demand forecasting.

In addition, the framework proposed in Chapter 3 to describe and compare demand forecasting studies provides two additional approaches for further research. First, the research community can benefit from applying the framework to structure their studies to make their results easily comparable for other researchers. Second, based on more comparable studies, systematic quantitative meta-studies can analyze what differences in the approaches have which effects on the results.

I also suggest potential future research based on the proposed STSF approach. As discussed in the critical review in Chapter 8.2, I derive two major open points from the empirical study. First, developing a solid theoretical foundation of the empirical findings would further substantiate the generality of the statements. Second and forward looking, systematic research of the mechanisms between the data input characteristics, the pipeline setup, and the forecasting result accuracy will enable to focus effectively on how to further optimize the pipeline. The existing empirical study setup can be further developed to systematically run experiments, altering the parameters and algorithms and measuring the improvements on the forecasting accuracy.