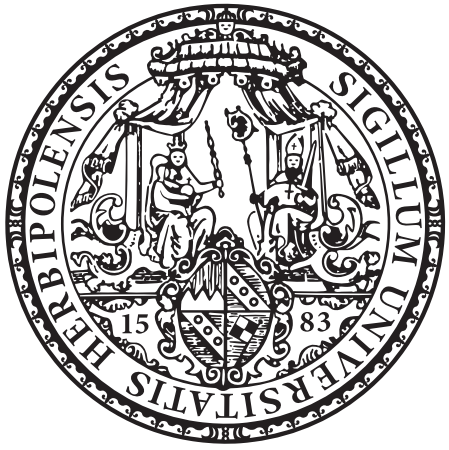
Julius-Maximilians-Universität Würzburg
Wirtschaftswissenschaftliche Fakultät



Prescriptive Analytics for Data-driven Capacity Management

Inauguraldissertation
zur Erlangung des akademischen Grades
doctor rerum politicarum (Dr. rer. pol.)

vorgelegt von
Pascal Markus Notz, M.Sc.
geboren in Frankfurt am Main

Name und Anschrift:	Pascal Markus Notz Riedeselstr. 70 64283 Darmstadt
Erstgutachter:	Prof. Dr. Richard Pibernik
Zweitgutachter:	Prof. Dr. Christoph Flath
Datum der Einreichung:	18. August 2020

Acknowledgements

This work would not have been possible without the support, encouragement, and guidance I received from many people throughout the last years. First and foremost, I would like to thank my doctoral advisor, Prof. Dr. Richard Pibernik, for his excellent supervision. I am extremely grateful for his continuous support and guidance along the journey that led to this dissertation. His insightful suggestions and constructive criticism in numerous deep and fruitful discussions enriched my work and enabled me to develop early ideas and findings into solid results.

I would also like to thank my second advisor, Prof. Dr. Christoph Flath, for the many insightful comments, useful hints, and challenging questions that not only opened up new perspectives but oftentimes led to a new avenue for further investigation that helped to improve my research.

Furthermore, I would like to thank all colleagues at the Chair of Logistics and Quantitative Methods in Business Administration and the Chair of Information Systems and Business Analytics for the stimulating discussions and the valuable feedback during our OPIM seminars, and in particular Peter Wolf for the great collaboration on our research article. Thanks for all the chats at the coffee machine, for all the fun during our retreats and the legendary Christmas celebrations, and for making this journey such a great experience.

Finally, I would like to express my gratitude to my family and especially to my wife, who always encouraged and supported me in a loving way. I am deeply grateful for her endless patience and loving care throughout this journey.

Deutschsprachige Zusammenfassung (Summary in German Language)

Digitalisierung und künstliche Intelligenz führen zu enormen Veränderungen in nahezu allen Bereichen von Wirtschaft und Gesellschaft. Grundlegend für diese Veränderungen ist die Technologie des maschinellen Lernens (ML), ermöglicht durch ein Zusammenspiel großer Datenmengen, geeigneter Algorithmen und ausreichender Rechenleistung. Ein ML-System lernt auf Basis eines großen Datensatzes von Beispielen, eine definierte Aufgabe zu lösen, beispielsweise, Nachfragewerte für ein bestimmtes Produkt auf Basis historischer Nachfragebeobachtungen vorherzusagen. Diese Technologie bildet die Basis für die Entwicklung neuartiger Ansätze zur Lösung klassischer Planungsprobleme des *Operations Research* (OR): Präskriptive Ansätze integrieren Methoden des ML und Optimierungsverfahren des OR mit dem Ziel, Lösungen für Planungsprobleme direkt aus historischen Observationen von Nachfrage und *Features* (erklärenden Variablen) abzuleiten. Dadurch kann beispielsweise direkt die optimale Bestellmenge oder Mitarbeiterkapazität bestimmt werden.

Mittels der digitalen Verarbeitung von Aufträgen können Unternehmen automatisiert Nachfragewerte und zugehörige Features (beispielsweise Datums-, Feiertags-, oder Wetterdaten) erfassen und somit große Datensätze anlegen. Eine zentrale Forschungsfrage besteht darin, wie diese Datensätze bestmöglich genutzt werden können, um bei klassischen Planungsproblemen bessere Entscheidungen zu treffen. Präskriptive Verfahren setzen bei dieser Frage an, indem sie Entscheidungen direkt auf Basis der Nachfrage- und Feature-Daten vorhersagen, anstatt, wie klassische OR-Ansätze, eine Nachfrageverteilung zu

schätzen und anhand dieser eine Entscheidung abzuleiten. Diese neuartigen Lösungsansätze bieten ein enormes Potential zur Verbesserung von Planungsentscheidungen, wie erste numerische Analysen mit historischen Daten gezeigt haben, und begründen damit ein neues Forschungsfeld innerhalb des OR.

In ersten Beiträgen zu diesem neuen Forschungsfeld wurden präskriptive Verfahren für verhältnismäßig einfache Planungsprobleme aus dem Bereich des Lagerbestandsmanagements entwickelt. Häufig weisen Planungsprobleme aber eine deutlich höhere Komplexität auf, beispielsweise durch eine mehrstufige Struktur, durch Elemente der Warteschlangentheorie oder durch mehrere voneinander abhängige Entscheidungen, die auf der Grundlage vektor- oder matrixwertiger Nachfragebeobachtungen zu treffen sind. Viele dieser komplexen Planungsprobleme gehören zum Bereich der Kapazitätsplanung. Daher ist die Entwicklung präskriptiver Ansätze zur Lösung komplexer Probleme im Kapazitätsmanagement das Ziel dieser Dissertation. In drei inhaltlich abgeschlossenen Teilen werden neuartige präskriptive Ansätze konzipiert und auf realistische Kapazitätsplanungsprobleme angewendet.

Im ersten Artikel, „Prescriptive Analytics for Flexible Capacity Management“, werden zwei präskriptive Verfahren entwickelt, und zwar *weighted Sample Average Approximation* (wSAA) und *kernelized Empirical Risk Minimization* (kERM), um ein komplexes, zweistufiges stochastisches Kapazitätsplanungsproblem zu lösen: Ein Logistikdienstleister sortiert täglich eintreffende Sendungen auf drei Sortierlinien, für die die wöchentliche Mitarbeiterkapazität geplant werden muss. Während wSAA eine Erweiterung des klassischen *Sample-Average-Approximation*-Verfahrens (SAA) darstellt, basiert kERM auf dem im ML etablierten ERM-Prinzip. Dieser Artikel ist der erste Beitrag, in dem ein kERM-Verfahren zur direkten Lösung eines komplexen, zweistufigen Planungsproblems mit matrixwertiger Nachfrage und vektorwertiger Entscheidung entwickelt, eine Obergrenze für die erwarteten Kosten für nichtlineare, kernelbasierte Funktionen abgeleitet und die *Universal Approximation Property* bei Nutzung spezieller Kernelfunktionen gezeigt wird. Die Ergebnisse der numerischen Studie demonstrieren, dass präskriptive Verfahren im Vergleich mit klassischen Lösungsverfahren (*Two-step*-Verfahren sowie SAA) zu signifikant besseren Entscheidungen führen können und ihre Entscheidungsqualität

bei Variation der exogenen Kostenparameter deutlich robuster ist.

Im zweiten Artikel, „Prescriptive Analytics for a Multi-Shift Staffing Problem“, werden wSAA und kERM auf ein Planungsproblem der klassischen Warteschlangentheorie angewendet: Ein Dienstleister erhält über den Tag verteilt Aufträge, deren Anzahl und Zeitpunkt des Eintreffens unsicher sind, und muss die Mitarbeiterkapazität für zwei Schichten planen. Dieses Planungsproblem ist aus zwei Gründen komplexer als die bisher mit präskriptiven Ansätzen gelösten Probleme: Zum einen werden Auftragseingang und Bearbeitung als Wartesystem modelliert, zum anderen folgt die Nachfrage innerhalb einer Schicht einem nicht stationären Prozess, wodurch die Zeitstruktur der Nachfrage entscheidungsrelevant wird. Diese Komplexität wird mit zwei Näherungsmethoden bewältigt, sodass das Planungsproblem mit wSAA und kERM sowie einem neu entwickelten präskriptiven Verfahren – dem *Optimization-Prediction*-Verfahren (OP) – gelöst werden kann. Numerische Analysen mit realistischen Kostenparametern sowie einer Parametervariation, die verschiedene Servicelevel induziert, zeigen, dass wSAA bei diesem Problem zu den besten Entscheidungen führt. Die in diesem Artikel entwickelte Methode legt den Grundstein zur Lösung komplexer Warteschlangenmodelle mit präskriptiven Verfahren und schafft damit eine Verbindung zwischen den „Welten“ der Warteschlangentheorie und der präskriptiven Verfahren.

Im dritten Artikel, „Explainable Subgradient Tree Boosting for Prescriptive Analytics in Operations Management“, wird ein neues präskriptives Verfahren zur Lösung der Planungsprobleme der ersten beiden Artikel entwickelt, das neben guter Entscheidungsqualität insbesondere durch die Erklärbarkeit der Entscheidungen attraktiv ist: *Subgradient Tree Boosting* (STB). Es kombiniert das erfolgreiche *Gradient-Boosting*-Verfahren aus dem ML mit SAA und verwendet Subgradienten, da die Zielfunktion von OR-Planungsproblemen häufig nicht differenzierbar ist. Der Artikel zeigt Methoden zur Ableitung eines Subgradienten für gängige OR-Probleme inklusive der Klasse der zweistufigen Planungsprobleme und führt eine umfassende numerische Analyse zum Vergleich von STB mit wSAA und kERM durch, die zeigt, dass STB zu einer vergleichbaren Entscheidungsqualität wie wSAA und kERM führen kann. Zusätzlich wird demonstriert, wie Kapazitätsentscheidungen in Beiträge einzelner

Features zerlegt und damit erklärt werden können. Das in diesem Artikel neu entwickelte STB-Verfahren ist damit nicht nur aufgrund seiner Entscheidungsqualität attraktiv für Entscheidungsträger, sondern insbesondere auch durch die inhärente Erklärbarkeit.

Die in diesen drei Artikeln präsentierten Ergebnisse zeigen, dass die Nutzung präskriptiver Verfahren, wie wSAA, kERM und STB, bei der Lösung komplexer Planungsprobleme zu deutlich besseren Ergebnissen führen kann als der Einsatz klassischer Methoden, die Feature-Daten vernachlässigen oder auf einer parametrischen Verteilungsschätzung basieren. Mit der Entwicklung der präskriptiven Verfahren für die Kapazitätsplanung, der theoretischen Analyse und den Ergebnissen der praktischen Anwendungen wird ein relevanter Beitrag zu diesem neuen Forschungsfeld innerhalb des OR geleistet.

Contents

1 Introduction

Digitization and artificial intelligence (AI) are radically changing virtually all areas across business and society (Brynjolfsson and McAfee 2014, Westerman et al. 2014). Since the mid-1990s large productivity increases have been captured by transferring simple and routine tasks to machines. For example, robots are now commonly used in car manufacturing lines, and payroll processing is often done by computer programs. Such uses of machines for jobs once done by humans is described as the first phase of the second machine age (McAfee and Brynjolfsson 2017). This phase is marked by the use of computers that are preprogrammed with sets of commands that define how to accomplish a certain task. Now we are experiencing the second phase of the second machine age as machines take over even more tasks, particularly complex cognitive tasks. Well-known examples of such machines include *AlphaGo*, an AI program that defeated the human world champion of the board game Go (Silver et al. 2017), and AI systems that are better at analyzing medical images than human experts are, for example, in predicting breast cancer based on mammograms (McKinney et al. 2020). Given the innumerable new technological opportunities that can be enabled through the use of AI, the economic impact is expected to be enormous: AI could increase the global GDP by 13 trillion USD by 2030 (Bughin et al. 2018). Most of these developments are driven by one sub-field of AI, *machine learning*, which focuses on algorithms that learn from massive amounts of data (McAfee and Brynjolfsson 2017). In contrast to the preprogrammed computers that enabled phase one of the second machine age, these machines are not explicitly programmed but learn for themselves how to accomplish a task (e.g., predicting breast cancer based on a mammogram) from large amounts of training data.

The theoretical foundations of today's machine learning approaches date back to the early work of Vapnik and Chervonenkis (1968) on uniform conver-

gence, which formed a cornerstone of statistical learning theory (Vapnik 1998, Schölkopf et al. 2013). Vapnik and Chervonenkis developed the well-known principle of *Empirical Risk Minimization* (ERM), which states that a prediction function can be learned by minimizing the empirical counterpart of the expected loss (commonly termed *true risk*) over a data set of observed realizations, which is necessary because the underlying true distribution is often unknown. Although statistical learning theory and many machine learning algorithms already existed in the late 1990s (e.g., support vector machines and artificial neural networks, including Hochreiter and Schmidhuber's (1997) well-known Long Short-Term Memory architecture), phase two of the second machine age started only about a decade ago, driven by the coming together of statistical learning theory, sufficient computational power, and the required amounts of training data to constitute the new age of AI (McAfee and Brynjolfsson 2017). In particular, while the advent of cloud technologies has made computational power widely available, data is a key asset in facilitating new business opportunities (Otto et al. 2016, MIT Technology Review Custom 2016, Agrawal et al. 2018).

One area in which the amount of available data has been growing significantly in the last decade is that of operations management (OM): observations of demand (e.g., customer orders of a product or service) are often recorded automatically by the computer systems that process the request (Choi et al. 2018). Examples from our project partners include a mail logistics provider that automatically stores the number of mail items sorted each day, and an aviation maintenance service provider that automatically records a timestamp for each part that arrives for maintenance. Consequently, the opportunities for the field of decision-making in OM under uncertainty through machine learning methods are tremendous, and a completely new field of research is arising around the question of how best to exploit these data sources to derive better decisions in OM.

1.1 A New Field of Research in Operations Management: Prescriptive Analytics

One of the main areas of machine learning is learning from historical observations of a quantity of interest (e.g., historical demand for a certain product) to predict the future value of this quantity (e.g., tomorrow's demand for the product). From the perspective of traditional OM, predicting demand is one of the key ingredients for decision-making (Agrawal et al. 2018), so straightforward improvement in OM decision-making by means of machine learning follows the two-step approach of first predicting demand using machine learning methods and then solving the OM planning problem to derive the optimal decision. Because this approach relies on data (historical observations of demand) and machine learning algorithms, it may already improve decisions over those that rely on the judgment of experienced or expert humans, so it may provide an important advantage for companies (McAfee and Brynjolfsson 2017). However, such two-step approaches "can be problematic because demand model specification is difficult in higher dimensions, and errors in the first step will amplify in the optimization" (Ban and Rudin 2019, p. 90). In contrast, prescriptive analytics approaches integrate prediction and optimization into a single prescription step, so they learn from historical observations of demand and a set of features (co-variates) and provide a model that directly prescribes future decisions. This combination of optimization techniques—which have long been the field of study of Operations Research (OR)—machine learning, and large amounts of historical data is considered one of the largest opportunities for OR (Bertsimas 2017), and first case reports suggest that using prescriptive analytics can reduce costs by 24 percent (Ban and Rudin 2019) or even 88 percent (Bertsimas and Kallus 2020). Consequently, these new prescriptive analytics approaches to solving classical OM problems like determining the optimal inventory of a certain product or the staff capacity needed to fulfill customer demand constitute a new field of research in OM. The need for research in this new field is evident from, for example, a call for papers for a *Management Science* special issue on "Data-Driven Prescriptive Analytics"

(Giesecke et al. 2018) and the formation of the new INFORMS *Journal on Optimization*, which follows the vision of "science that starts with data and builds models to derive optimal decisions that add value" (Bertsimas 2017, p. 14)—the key idea of prescriptive analytics.

1.2 Prescriptive Analytics for Data-driven Capacity Management

Prescriptive analytics constitutes a new field of research in OM, and first works have studied new approaches to solving comparatively simple planning problems in the area of inventory management (e.g., Ban and Rudin 2019, Bertsimas and Kallus 2020). However, common OM planning problems often have a more complex structure, such as two-stage stochastic problems with recourse and problems of queuing theory, and may require making multiple interdependent decisions depending on vector-valued or even matrix-valued observations of demand. Many of these complex planning problems are within the domain of capacity planning, so this dissertation focuses on developing new prescriptive analytics approaches for complex capacity management problems. In the most general terms, capacity can be defined as a "measure of processing abilities and limitations" (Van Mieghem 2003, p. 269), and decisions with regards to capacity must often be made under uncertainty of demand. Therefore, the guiding research question of this dissertation is:

Guiding Research Question. *How can prescriptive analytics approaches combine methods of machine learning with OR optimization and exploit available demand and feature data to prescribe better decisions for complex capacity planning problems?*

Capacity planning under demand uncertainty is a classical OR problem and Van Mieghem (2003) provides a comprehensive review of traditional approaches to capacity management, distinguishing between two classes of capacity models: newsvendor-type and queuing-type models. Newsvendor-type models are "typically set in discrete-time and focus on the impact of multi-

variate demand uncertainty" (Van Mieghem 2003, p. 281), often leading to decisions that consist of a forecasted mean demand and an uncertainty hedge in the form of a safety buffer that is structurally similar to the newsvendor solution. Therefore, the first research question to be addressed by this dissertation can be stated as:

Research Question 1. *How can prescriptive analytics approaches be used to make better decisions for a complex newsvendor-type capacity planning problem with uncertain demand?*

In contrast, typical queuing models use continuous time (Van Mieghem 2003) and allow the decision-maker to define flow-related objectives like waiting time or queue length, which often increases model complexity. The respective research question for this type of capacity planning problem can be stated as:

Research Question 2. *How can prescriptive analytics approaches be used to make better decisions for a complex queuing-type capacity planning problem with uncertain demand?*

The primary reason for using prescriptive analytics for decision-making has been to make better decisions (e.g., Ban and Rudin 2019, Bertsimas and Kallus 2020), but recently an additional aspect has emerged: the explainability of prescribed decisions. While many decision-making tasks can be fully automated using prescriptive analytics, some still require human input or collaboration between human and machine to achieve the best performance (Agrawal et al. 2018). When humans need to collaborate with machines that prescribe decisions, the machine should provide explanations for its prescriptions so the human can understand the underlying causality and build trust in the model's prescriptions, both of which are common motivations for research on explainable AI (Lipton 2018). The growing research interest in explainable AI and explainable prescriptive analytics for OM is also evident in Marsden et al.'s (2020) recent call for papers. The third research question to be addressed by this dissertation, then, focuses on explainable prescriptive analytics.

Research Question 3. *How can prescriptive analytics approaches to solving complex capacity planning problems be used to derive explanations so decision-makers can understand the reasons for the prescribed decisions?*

The next section presents how this dissertation addresses the three research questions.

1.3 Structure of the Dissertation

This dissertation consists of three independent articles that follow the Guiding Research Question in an effort to contribute to the research on prescriptive analytics for complex capacity planning problems.

The first article, "Prescriptive Analytics for Flexible Capacity Management"[1] (Chapter 2), addresses Research Question 1 by developing two prescriptive analytics approaches, weighted sample average approximation (wSAA) and kernelized empirical risk minimization (kERM), to solve a complex two-stage capacity planning problem that has been studied extensively in the literature (Netessine et al. 2002, Bassok et al. 1999). In this problem, a logistics service provider sorts daily incoming mail items on three service lines that must be staffed on a weekly basis. This article is the first to develop a kERM approach—which applies Vapnik's ERM principle and uses the kernel trick to incorporate non-linear function spaces—to solve a complex two-stage stochastic capacity planning problem with matrix-valued observations of demand and vector-valued decisions. The article compares wSAA and kERM analytically by building on statistical learning theory, develops out-of-sample performance guarantees for kERM and various kernels, and shows the universal approximation property when using a universal kernel. A comprehensive numerical study is conducted using historical demand data from the case company, realistic cost parameters, and a variation of cost parameters to induce various service levels. The results of the numerical study suggest that prescriptive analytics approaches may lead to significant improvements in performance compared to traditional two-step approaches or SAA and that their performance is more

[1] This article is co-authored by Richard Pibernik.

robust to variations in the exogenous cost parameters.

The second article, "Prescriptive Analytics for a Multi-Shift Staffing Problem"[2] (Chapter 3), addresses Research Question 2 in using prescriptive analytics approaches to solve the (queuing-type) multi-shift staffing problem (MSSP) of an aviation maintenance provider that receives customer requests of uncertain number and at uncertain arrival times throughout each day and plans staff capacity for two shifts. This planning problem is particularly complex because the order inflow and processing are modelled as a queuing system, and the demand in each day is non-stationary, which makes the time structure of demand arrivals an important factor in determining the optimal staff capacity. The article addresses this complexity by deriving an approximation of the MSSP's queuing model formulation that enables the planning problem to be solved using wSAA and kERM. In addition, the article proposes a novel prescriptive analytics approach that is termed the *Optimization Prediction* (OP) approach. A numerical evaluation using realistic cost parameters and historical demand data from a case company demonstrates the applicability of prescriptive analytics to this type of queuing problem. Additional numerical experiments using a variety of cost parameters suggest that, while the OP approach is an attractive choice because of its simplicity, wSAA may lead to better performance, particularly for high or low service levels. The solution method developed in this article builds a foundation for solving queuing-type planning problems using prescriptive analytics approaches, so it bridges the "worlds" of queuing theory and prescriptive analytics.

The third article, "Explainable Subgradient Tree Boosting for Prescriptive Analytics in Operations Management" (Chapter 4), addresses Research Question 3 by proposing a novel prescriptive analytics approach, termed *Subgradient Tree Boosting* (STB), that allows decision-makers to derive explanations for prescribed decisions. STB combines the machine learning method Gradient Boosting with SAA and relies on subgradients because the cost function of OM planning problems often cannot be differentiated. The methods with which to derive subgradients for common OM problems that the article proposes in-

[2]This article is co-authored by Peter K. Wolf and Richard Pibernik.

clude the class of two-stage stochastic problems. A comprehensive numerical analysis is conducted that uses STB to solve the two capacity planning problems studied in the first and second articles; the results suggest that STB can lead to a prescription performance that is comparable to that of wSAA and kERM. The explainability of STB prescriptions is demonstrated by breaking exemplary decisions down into the impacts of individual features. The novel STB approach is an attractive choice not only because of its prescription performance, but also because of the explainability that helps decision-makers understand the causality behind the prescriptions.

An overview of each of the three articles' scientific contributions is presented in Table 1.1. The proofs of the theoretical results presented in the main part of this dissertation, along with additional theoretical and numerical results, can be found in Appendices A, B, and C. A summary of the results and insights is presented in Chapter 5, together with several avenues for future research.

Table 1.1: Overview of scientific contributions.

Article	*Methodological contribution*	*Practical contribution*	*Conceptual findings*
Prescriptive Analytics for Flexible Capacity Management (Chapter 2)	• Development of kERM to solve a complex two-stage problem • Derivation of performance guarantees and the universal approximation property for kERM • Analytical comparison of kERM and wSAA	• Case study of a mail logistics provider • Comparison of approaches' numerical performance, including cost parameter variations	• Prescriptive approaches can lead to significant improvements in performance over those of traditional approaches. • Prescriptive approaches' performance is more robust to variations in cost parameters than traditional approaches are.
Prescriptive Analytics for a Multi-Shift Staffing Problem (Chapter 3)	• Approximation of a complex queuing model (MSSP) • Solution of the AMSSP using prescriptive analytics approaches • Development of the OP approach	• Case study of a maintenance service provider • Numerical comparison, including comprehensive variations in cost parameters	• wSAA may lead to better performance than kERM or OP. • kERM may be prone to a regularization effect, and OP may be prone to a service-level effect.
Explainable Subgradient Tree Boosting for Prescriptive Analytics in Operations Management (Chapter 4)	• Development of the STB approach • Derivation of subgradients for common OM problems, including complex two-stage problems • Analytical comparison of STB and kERM	• Application of STB to solve two case studies (mail logistics, aviation maintenance) • Derivation of explanations for exemplary prescriptions	• STB can lead to performance similar to those of wSAA or kERM. • Explanations of STB prescriptions can support understanding of the underlying causality.

2 Prescriptive Analytics for Flexible Capacity Management

Motivated by the real-world problem of a logistics company, this paper proposes a novel distribution-free prescriptive analytics approach—termed *kernelized ERM*—to solve a complex two-stage capacity planning problem with multivariate demand and vector-valued capacity decisions and compares this approach both theoretically and numerically to an extension of the well-known sample average approximation (SAA) approach termed *weighted SAA*. Both approaches use integrated machine learning algorithms to prescribe capacities directly from historical demand and numerous features (co-variates) without having to make assumptions about the underlying multivariate demand distribution. We provide extensive analytical insights into both approaches. Most important, we prove the universal approximation property for the kernelized ERM approach when using a universal (data-independent) kernel and show how out-of-sample guarantees can be derived for various kernels.

We demonstrate the applicability of both approaches to a real-world planning problem and evaluate their performance relative to traditional parametric approaches that first estimate a multivariate demand distribution and then solve a stochastic optimization problem, and a non-parametric approach (SAA). Our results suggest that the two prescriptive analytics approaches can result in substantial performance improvements of up to 58 percent compared to these traditional approaches. Additional numerical analyses shed light on the behavior and performance drivers of the various approaches and demonstrate that, in our case, the prescriptive approaches are much more robust to variations of exogenous cost parameters than traditional approaches are.[3]

[3]This paper is co-authored by Richard Pibernik and has been published in *Management Science* (Notz and Pibernik 2021, `https://doi.org/10.1287/mnsc.2020.3867`). Republished with permission of INFORMS from "Prescriptive Analytics for Flexible Capacity Management", Pascal M. Notz, Richard Pibernik, Management Science, Articles In Advance, Copyright 2021; permission conveyed through Copyright Clearance Center, Inc.

2.1 Introduction

In many manufacturing and service industries, companies have some flexibility in using their resources to meet their customers' uncertain demand for various products or services. For example, car rental companies can use mid-sized cars to meet unexpectedly high demand for compact cars, service technicians with high skill levels can be employed to meet demand for tasks that require a lower level of expertise (Netessine et al. 2002), and a logistics service provider that has semi-automated sorting lines may use the workforce that operates these sorting lines when there are more than the expected number of shipments that require manual processing. In these and many other instances, customers may be "upgraded" to a higher level of service—as in the car rental example—or a more expensive resource that delivers the same level of service without the customer's incurring additional costs. Clearly, manufacturers and service providers can benefit from this "upgrading flexibility" when demand is not perfectly (positively) correlated. However, determining the right capacities for the various resources is difficult, especially because both the individual capacity decisions and the (uncertain) demand for products and services are interrelated. Moreover, capacity decisions are made for extended time periods, such as a week or a month, while the allocation of demand to available resources is carried out for shorter time periods (e.g., daily or hourly). From a mathematical perspective, the company has to solve a complex two-stage stochastic optimization problem with recourse where the vector-valued decisions are interrelated and uncertain demand follows some (known or unknown) multivariate distribution. Researchers in operations management (OM) studied variants of this problem extensively almost two decades ago and developed solution approaches, assuming that the true multivariate demand distribution is known (e.g., Bassok et al. 1999, Netessine et al. 2002).

Motivated by the real-world capacity management problem of a logistics service provider, this paper proposes and studies new data-driven, prescriptive analytics approaches to the aforementioned capacity planning problem with upgrading. These approaches use integrated machine learning algorithms to prescribe capacities directly from historical demand and numerous *features*

(independent/explanatory variables, co-variates), without having to make assumptions about the multivariate demand distribution. Our approaches contrast with the traditional two-step approach of first estimating a (multivariate) demand distribution based on some time-series and/or causal model and then solving a stochastic optimization problem to determine the optimal capacities. As Ban and Rudin (2019, p. 90; BR hereafter) pointed out, such two-step processes "can be problematic because demand model specification is difficult in higher dimensions, and errors in the first step will amplify in the optimization." In our setting, demand model specification is particularly difficult because: (i) the first and second moments of the (multiple) demand distributions may depend on some or all of the features (e.g., the particular week during a year or month, demand in the previous week, public holidays), so demand may be non-stationary and heteroscedastic; (ii) the demands for the various services may be correlated within individual periods, but correlations can be feature-dependent, so they may not be constant across time; (iii) the number of relevant historical observations is small, while the number of features may be large, so we are likely to face the typical problems associated with *high-dimensional* data (see Hastie et al. 2009, Chapter 18). We use data from a case company to illustrate the practical relevance of the first two issues, which make it particularly attractive to employ tailored distribution-free approaches that prescribe (vector-valued) capacity decisions for a two-stage problem directly from historical demand observations and available feature data, even when the number of features is large.

The research presented in this paper draws on and extends a recent stream of work that proposes and studies prescriptive analytics approaches for solving problems in operations research and management science (OR/MS). Bertsimas and Kallus (2020) (BK hereafter) developed a comprehensive framework for prescriptive analytics in OR/MS, proposing a set of local learning methods that rely on an intuitive integration of well-established predictive machine learning methods (e.g., random forests) and traditional techniques for data-driven optimization (i.e., sample average approximation). Our first prescriptive analytics approach for solving the capacity planning problem with upgrading—termed weighted sample average approximation (wSAA)—is

based on BK's local learning methods. We demonstrate how their methods can be applied to our capacity planning problem and show that important properties of BK's approaches (especially asymptotic optimality) are preserved in our problem setting.

In addition to the wSAA approaches, BK considered prescriptive analytics approaches that are based on the well-grounded machine learning principle of empirical risk minimization (ERM) that Vapnik (1991) introduced, but they argued in favor of wSAA approaches especially because ERM-based approaches may lead to infeasible prescriptions and do not enjoy the universal guarantees of asymptotic optimality. Despite BK's arguments in favor of local learning methods, we show that our second approach, termed kernelized ERM (kERM hereafter), has properties that could make it an attractive choice in the context of our capacity planning problem. We formulate the kERM approach to solving our capacity planning problem and provide solution techniques for non-linear function spaces defined by a kernel function that allow us to solve the problem efficiently over a reproducing kernel Hilbert space. We explain why this approach does not suffer from the limitations BK stated (Section 6 in BK) and develop strong theoretical results regarding our approach's performance, including a universal approximation property and out-of-sample performance guarantees for various kernels. To the best of our knowledge, this study is the first to employ a kERM approach for solving directly a complex (two-stage) OM problem with vector-valued decisions and multivariate demand and to apply this approach to a real-world problem. It draws on and extends the theoretical results of BK by providing out-of-sample performance guarantees for various kernels and by proving a universal approximation property for kERM when employing a universal kernel. This is the main theoretical contribution of our paper.

We cannot conclude based on our theoretical results how the prescriptive analytics approaches perform in a practical setting with limited amounts of historical data. To explore the performance of our prescriptive analytics approaches in a real-world application, we conduct a comprehensive case study and compare these approaches' performance to that of various (traditional) benchmark approaches, including time-series and causal models and SAA.

This case study is our main practical contribution. We find that both wSAA and kERM lead to substantial improvements in out-of-sample performance compared to traditional two-step methods, as for realistic cost parameters they achieve cost reductions of up to 58 percent. We interpret the approaches' outcomes and extend our insights by varying the exogenous cost parameters of the capacity planning problem. Both prescriptive approaches appear to be much more robust to variations of the exogenous parameters than their traditional counterparts are, with wSAA performing better in regimes with high optimal service levels, and kERM performing better when service levels are in the medium range. We use our numerical results to provide intuition about the underlying dynamics that drive the behavior of the two approaches.

2.2 Literature Review

This paper proposes new prescriptive analytics approaches to solve a well-known capacity planning problem. Our work builds directly on the capacity planning problem with upgrading that Netessine et al. (2002) described, which is based on Bassok et al. (1999), who studied a single-period multi-product inventory management problem with downward substitution. Bassok et al. (1999) formulated a two-stage stochastic profit maximization problem, characterized the optimal policy, and—under the assumption of a known joint distribution of demand—provided an algorithm to solve the two-product problem. Netessine et al. (2002) addressed a two-stage stochastic optimization problem that is similar to that of Bassok et al. (1999) "in mathematical structure but is different in interpretation" (Netessine et al. 2002, p. 377), as they study a multi-service capacity planning problem with upgrading, as described in Section 2.1. Restricting their model to single-level upgrading and assuming a known multivariate demand distribution, Netessine et al. (2002) provided analytical bounds on the optimal capacities based on newsvendor quantities and developed an intuitive algorithm for determining these optimal capacities. They also showed analytically how demand correlation affects the optimal solution when demand follows a multivariate normal distribution. A number of

contributions have extended Netessine et al.'s (2002) model, including Shumsky and Zhang (2009) and Yu et al. (2015). Our work extends Bassok et al.'s (1999) and Netessine et al.'s (2002) models to a horizon of T periods with a decay of unused capacity in each period. This choice is motivated by the real-world capacity planning problem of a logistics service provider that we detail in Section 2.6.1.

Apart from the obvious structural similarities, the aforementioned contributions have in common that they require knowledge of the underlying multivariate demand distribution before optimal capacity decisions can be made. While this assumption has been common in OM research, it is widely acknowledged that the decision-maker typically does not know the true demand distribution (see, e.g., Liyanage and Shanthikumar 2005, Akcay et al. 2011, and Klabjan et al. 2013). The typical (parametric) approach is to assume that the true demand distribution belongs to a parametric family of distributions and to estimate the unknown parameters (Liyanage and Shanthikumar 2005).

Research on inventory management has demonstrated that the wrong choice of a distribution and/or misspecification of its parameters can lead to sub-optimal outcomes, even in cases of univariate demand distributions (see Liyanage and Shanthikumar 2005 and Klabjan et al. 2013). We use our real-world application to demonstrate that this problem can be particularly pronounced in capacity planning with multivariate demand distributions and vector-valued capacity decisions. In light of these problems, a number of researchers have proposed alternatives that rely, for example, on Bayesian updating or robust optimization (see BR and Liyanage and Shanthikumar 2005 for detailed reviews).

Distribution-free, data-driven approaches—to which we refer as non-parametric approaches—to solving stochastic optimization problems in OM have gained increasing attention. The traditional non-parametric method is SAA, where the true distribution is replaced by the empirical one (Shapiro and Kleywegt 2002, Shapiro 2003). SAA has been widely applied in single- and multi-period inventory control (see, e.g., Levi et al. 2015, Shi et al. 2016, Cheung and Simchi-Levi 2019, and Ban 2020). SAA has traditionally been used to solve two types of problems: those that are either difficult to solve

analytically, although the underlying distribution is known, and those whose objective functions would be easy to evaluate if the distribution were known, but it is not (Levi et al. 2015). Our capacity planning problem exhibits both properties. We use traditional SAA as a performance benchmark and adopt an extension, wSAA, that was initially proposed by BK.

Our prescriptive approaches are non-parametric so they are related to the stream of literature on SAA and other non-parametric approaches to solving OM problems (see Bertsimas et al. 2018 and the discussion and references provided in BK). The distinguishing aspect of our research is the direct incorporation of a (potentially) large set of auxiliary data (features). We propose two prescriptive analytics approaches that integrate machine learning and optimization techniques to prescribe vector-valued capacity decisions directly, so our work is closely related to BK's and BR's recent contributions. BK proposed a set of prescriptive analytics approaches that combine local predictive machine learning methods and traditional techniques for data-driven optimization (i.e., SAA). All of these approaches are based on deriving weights from the features by means of local predictive machine learning methods and "optimizing the decision [...] against a reweighting of the data" (BK, p. 1030). We term these approaches "weighted SAA" (wSAA). BK proposed a set of alternative weight functions and showed the tractability and asymptotic optimality of their approach for most of these weight functions. We draw on this work to propose a wSAA approach for our capacity planning problem. We demonstrate that, in our problem setting, the property of asymptotic optimality is preserved, that the rate of convergence to the full-information optimum may decline exponentially in the number of features and that, at least for scalar-valued decisions and a convex loss function, the approach interpolates between decisions that would have been optimal in the past. While the focus of BK's paper lay on wSAA, they also proposed an alternative ERM-based approach and provided a very general formulation with a linear function space. BK explained that "the linear decision rule can be generalized [...] by embedding in a reproducing kernel Hilbert space" (Section EC.1 in BK). We extend this line of thought, develop an ERM-based approach specifically for our complex capacity planning problem, and demonstrate how it can be solved over

a reproducing kernel Hilbert space (Section 2.4.2). Drawing on BK's generalization of the standard function-space complexity theory and out-of-sample guarantees to multivariate uncertainty and decisions (see Section EC.1 in BK), we derive out-of-sample performance guarantees for the kERM approach using data-independent kernels or the random forest kernel, using a sample-splitting approach. We extend these results by proving a universal approximation property for kERM when employing a universal kernel (Section 2.5).

BR developed two prescriptive analytics approaches to solving what they termed the "Big Data Newsvendor", that is, a newsvendor problem in which the decision-maker "has access to a potentially large amount of relevant information, such as customer demographics, weather forecasts, seasonality (e.g., day of the week, month of the year, and season), and economic indicators (e.g., the consumer price index) as well as past demands to inform the [decision-maker's] ordering decisions" (BR, p. 90). BR first proposed a linear ERM-based approach (with and without regularization) to solving the newsvendor problem, which is "equivalent to high-dimensional quantile regression" (p. 91), and derived corresponding out-of-sample guarantees under the assumption of a linear demand model. They further quantified the "value of feature information" (BR, p. 97) by comparing the ERM decisions with those of SAA for two exemplary demand models and also indicated how their ERM approach could be solved for non-linear function spaces by using kernels (see Appendix A in BR). However, BR did not develop out-of-sample guarantees for non-linear decision rules or provide numerical evidence for how a non-linear ERM approach performs relative to a linear approach. We provide a kernelized (non-linear) solution to our more complex, two-stage capacity planning problem with multivariate uncertainty and vector-valued decisions and provide out-of-sample guarantees for various kernels, including a non-standard random forest kernel, using a sample-splitting approach. BR's second approach, the "Kernel Optimization Method", is equivalent to wSAA in that it has a weight function based on kernel methods. In their numerical evaluation, BR used a Gaussian kernel and found that the wSAA approach outperforms the linear ERM approach. This finding highlights the need for a comparison of the wSAA and kERM approaches because, in general, "there is no reason to expect

that optimal solutions will have a linear structure" (BK, p. ec2, e-companion). We pursue such a comparison of the two approaches using the same random forest-based similarity measure for the weight and kernel functions, and we study how these prescriptive approaches' underlying mechanics contrast those of traditional two-step and data-driven approaches.

2.3 Problem Statement and Model Formulation

This section provides a formal characterization of the capacity planning problem we address based on Netessine et al. (2002) and Bassok et al. (1999). Consider a company that offers I services at a per-unit price of p_i $(i = 1, ..., I)$. Demand for each service i in each planning period $t \in \{1, ..., T\}$ (e.g., every day of a week) is uncertain, and we model it as a random variable denoted by D_i^t. The $I \times T$ distributions of D_i^t are unknown, but the company has access to a set of data $S_N = \{(\boldsymbol{d}^1, \vec{x}^1), ..., (\boldsymbol{d}^N, \vec{x}^N)\}$ that contains the most recent historical observations of demand $\boldsymbol{d}^n = (\vec{d}^{n,1}, ..., \vec{d}^{n,T}) \in \mathcal{D} \subset \mathbb{R}^{I \times T}$, with daily demand $\vec{d}^{n,t} = (d_1^{n,t}, ..., d_I^{n,t})$, and corresponding observations of features represented by vectors $\vec{x}^n \in \mathcal{X} \subseteq \mathbb{R}^p$. The features describe, for example, seasonality (day, month, week, season), weather conditions, and other independent observable variables that may be predictive of demand. To provide the various services i, the company employs distinct resources $j = 1, ..., I$, to which we refer as *service lines*.[4] Service i is delivered by service line $j = i$, but it can also be delivered by any service line $j \leq i$, although at a lower profit per unit. Hence, there is a hierarchy of service lines, with $j = 1$ the most flexible service line, as it can deliver all services $i = 1, ..., I$. The company has to determine the service lines' capacities prior to the first period of the T-period planning horizon, when demand for the services is unknown. In our context, the capacities are the staffing levels for one week (i.e., $T = 5$), which are constant for every day t of the week. We denote by $\vec{q} = (q_1, ..., q_I)$ the constant capacities available in each period $t \in \{1, ..., T\}$ of the planning horizon, and by f_j the fixed cost per unit of capacity j. In addition to this fixed

[4]The time for delivering one service equals one period.

cost, we assume a variable cost v_j that is incurred with the use of one unit of capacity j. In each period t the company first observes the demand in the period and then allocates the realized demands $\vec{d}^t = (d_1^t, ..., d_I^t)$ to the given capacities $\vec{q}$. The company incurs a penalty cost c_i per unit for not fulfilling the demand for service i and realizes a per-unit contribution margin of $a_{i,j}$ when service line j is used to fulfill demand for service i. The contribution margin is defined as $a_{i,j} = p_i - v_j + c_i$, and we assume $a_{i,i} \geq a_{i,j}$ for $j < i$—that is, it is more profitable to fulfill demand of type i with service line $j = i$ and less profitable if demand is upgraded and fulfilled with service line $j < i$. The company's objective is to determine the optimal capacities for the T-period planning horizon to maximize expected profit under the assumption that realized demands $\vec{d}^t$ are allocated optimally to service lines j in each period t. Let $y_{i,j}$ denote the amount of service i fulfilled by service line j once demand is observed. The company's capacity planning problem can be represented by the following two-stage stochastic optimization problem:

$$
\begin{aligned}
\text{Stage 1: } \max_{\vec{q}, q_j \geq 0} \Pi(\vec{q}) &= \max_{\vec{q}, q_j \geq 0} \left(\sum_{t=1}^{T} \mathbb{E}\left(\pi(\vec{D}^t, \vec{q})\right) - \sum_j f_j q_j \right) \\
\text{Stage 2: } \pi(\vec{d}, \vec{q}) &= \max_{\{y_{ij}\}} \sum_{i,j} a_{ij} y_{ij} - \sum_i c_i d_i \\
&\quad \text{s.t. } \sum_j y_{ij} \leq d_i \ \forall i \\
&\quad\quad \sum_i y_{ij} \leq q_j \ \forall j \\
&\quad\quad y_{ij} \geq 0 \ \forall i, j \\
&\quad\quad y_{ij} = 0 \text{ if } i < j.
\end{aligned}
\tag{2.1}
$$

This formulation differs in one minor aspect from that which Netessine et al. (2002) proposed: While for reasons of tractability they assumed a single-period allocation problem on stage 2, we model multiple allocation decisions that are made independently in each period $t \in \{1, ..., T\}$, given the capacities $\vec{q}$ determined on stage 1 and the demand realizations $\vec{d}^t$.[5] Because the

[5] Because the allocation decisions on stage 2 are independent across the T periods, we can assume that they are taken simultaneously after all demands $\vec{d}^t$ are known.

consecutive allocation decisions on stage 2 are independent, we can also show that the profit function $\Pi(\vec{q})$ is concave.

Proposition 2.1. *The profit function $\Pi(\vec{q})$, as defined in (2.1), is jointly concave in $\vec{q}$.*[6]

The distribution of the non-stationary, feature-dependent demand $\boldsymbol{D}$, given $\vec{X} = \vec{x}$ is unknown in most practical settings, and a traditional approach would be to first estimate the conditional distribution and to then solve Problem 2.1.

2.4 Prescriptive Analytics Approaches

In contrast to the traditional approach of first estimating a feature-dependent, multivariate distribution of the demand $\boldsymbol{D}$ and then solving Problem 2.1, a prescriptive analytics approach directly prescribes the optimal decision $\vec{q}(\vec{x})$ that minimizes the loss when given a new feature vector $\vec{x}$.

In the case of our capacity planning problem, the loss function can be defined as

$$L(\vec{q}, \boldsymbol{d}) = \Pi^*(\boldsymbol{d}) - \Pi(\vec{q}, \boldsymbol{d}), \tag{2.2}$$

where $\Pi(\vec{q}, \boldsymbol{d})$ represents the profit associated with the capacity decision $\vec{q}$ under demand realization $\boldsymbol{d}$, and $\Pi^*(\boldsymbol{d}) = \max_{\vec{q}} \Pi(\vec{q}, \boldsymbol{d})$ is the ex-post optimal profit. This definition ensures that $L(\vec{q}, \boldsymbol{d}) \geq 0$.

Proposition 2.2. *The loss function $L(\vec{q}, \boldsymbol{d})$ is jointly convex in $\vec{q}$.*

Two approaches have been proposed to solve the prescriptive analytics problem of determining $\vec{q}(\vec{x})$. The first seeks to minimize the true risk $R(\vec{q}(\cdot))$, which is defined as the expected loss over the joint distribution of $\vec{X} \times \boldsymbol{D}$, by selecting from a function space $\mathcal{F}$, which we assume to be a Banach space such that a norm is defined and the minimum in Equation 2.3 exists, a function $\vec{q}(\cdot) : \mathcal{X} \to \mathcal{Q}$ that maps from the feature space $\mathcal{X}$ to a decision space $\mathcal{Q}$:

$$\min_{\vec{q}(\cdot) \in \mathcal{F}} R(\vec{q}(\cdot)) := \min_{\vec{q}(\cdot) \in \mathcal{F}} \mathbb{E}_{\vec{X} \times \boldsymbol{D}} \left[L\left(\vec{q}(\vec{X}), \boldsymbol{D}\right) \right]. \tag{2.3}$$

[6] All proofs can be found in Appendix A.1.

Knowing such a function $\vec{q}(\cdot)$ allows one to determine a capacity decision $\vec{q}(\vec{x}) \in \mathcal{Q}$ for each new observation of a feature vector $\vec{x} \in \mathcal{X}$. The second approach defines $\vec{q}(\vec{x})$ point-wise, without the need to define a specific function space, by solving

$$\min_{\vec{q} \in \mathcal{Q}} \mathbb{E}_{\boldsymbol{D}} \left[L\left(\vec{q}, \boldsymbol{D}\right) \middle| \vec{X} = \vec{x} \right], \tag{2.4}$$

for each new $\vec{x}$.

Because neither the joint probability distribution of $\vec{X} \times \boldsymbol{D}$ nor the conditional probability distribution of $\boldsymbol{D}$, given $\vec{X} = \vec{x}$ is known, (2.3) and (2.4) cannot be solved directly. However, prescriptive analytics approaches to (2.3) and (2.4) can be derived for a decision-maker who has access to a data set S_N that consists of historical observations of demand and features. The well-established machine learning principle of ERM proposes that one can prescribe capacities using the function $\vec{q}^{\,\mathrm{ERM}}(\cdot)$ that minimizes the empirical risk $R_N(\vec{q}(\cdot))$ instead of the true risk $R(\vec{q}(\cdot))$ (see BK and BR):

$$\min_{\vec{q}(\cdot) \in \mathcal{F}} R_N(\vec{q}(\cdot)) := \min_{\vec{q}(\cdot) \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} L(\vec{q}(\vec{x}^n), \boldsymbol{d}^n). \tag{2.5}$$

BK proposed a number of alternatives to an ERM approach based on local learning techniques that take the common form of deriving some weights $w_n(\vec{x}) \in [0, 1]$ from the features and "optimizing the decision $[\vec{q}]$ against a reweighting of the data" (BK, p. 1030), as expressed in (2.6):

$$\vec{q}^{\,\mathrm{wSAA}}(\vec{x}) = \arg\min_{\vec{q} \in \mathcal{Q}} \sum_{n=1}^{N} w_n(\vec{x}) L(\vec{q}, \boldsymbol{d}^n). \tag{2.6}$$

In the most general terms, this approach approximates (2.4) and can be viewed as a weighted form of SAA (wSAA). The weight function $w_n(\cdot)$ can be considered a similarity function; as such, it has strong similarities with the kernel functions $K(\cdot, \cdot)$ used in kERM approaches, as we will see in Section 2.4.2. Obviously, the performance of a wSAA approach is determined by the choice of similarity function. BK constructed a number of weight functions based on k-nearest-neighbor regression, kernel regression, local linear regression, regression trees, and random forests.

We do not know a priori whether an ERM approach or a wSAA approach is more appropriate for our particular capacity planning problem, and we cannot make any claims regarding differences in their performance. BK made a number of arguments in favor of local learning approaches and described the limitations of ERM approaches applied to OR/MS problems. We show that an ERM approach is suitable for solving the prescriptive analytics problem stated in (2.3) and that it also has some properties—including the capability to extrapolate, the universal approximation property, and performance guarantees—that could make it a similarly attractive choice.

2.4.1 Weighted Sample Average Approximation Approach

Based on Problem 2.6, we state the wSAA approach for our capacity planning problem as:

$$\begin{aligned}\vec{q}^{\,\mathrm{wSAA}}(\vec{x}) = \arg\min_{\vec{q}\in\mathcal{Q}} \min_{\{y_{ij}^{tn}\}} \sum_{n=1}^{N} w_n(\vec{x}) \left(\sum_j f_j q_j - \sum_{t=1}^{T} \left(\sum_{i,j} a_{ij} y_{ij}^{tn} - \sum_i c_i d_i^{tn} \right) \right) \\ \text{s.t.} \quad \sum_j y_{ij}^{tn} \leq d_i^{tn} \ \forall i, n, t \\ \sum_i y_{ij}^{tn} \leq q_j \ \forall j, n, t \\ y_{ij}^{tn} \geq 0 \ \forall i, j, n, t \\ y_{ij}^{tn} = 0 \text{ if } i < j \ \forall n, t. \end{aligned} \tag{2.7}$$

We neglected $\Pi^*(\boldsymbol{d})$ because it is independent of $\vec{q}$.

Proposition 2.3. *The objective function of the wSAA approach is jointly convex in $\vec{q}$.*

For any given weight function $w_n(\vec{x}) \geq 0$, the wSAA approach (2.7) is a linear program. Using the results presented in BK, we show the asymptotic optimality[7] of our particular wSAA approach for the same classes of weight

[7] A wSAA approach is considered asymptotically optimal when, in the limit of $N \to \infty$, the expected cost of a decision $\vec{q}(\vec{x})$ equals the minimum expected cost under full knowledge of the distribution of $\boldsymbol{D}$, given $\vec{X} = \vec{x}$. See Definition 1 in BK for details.

functions as BK—that is, for weight functions that are based on k-nearest-neighbors (kNN), (recursive) kernel methods, and local linear methods (see Proposition A.1 in Appendix A.7.1).

Clearly, asymptotic optimality of a function $\vec{q}(\vec{x})$ is a desirable property of a prescriptive analytics approach. However, the convergence rate of these wSAA approaches may be prone to the curse of dimensionality; that is, it may decrease exponentially with the dimensionality p of the feature vector $\vec{x}$. We cannot prove this case for arbitrary distributions of $\vec{X} \times \boldsymbol{D}$ because a general expression for the rate of convergence of $\vec{q}^{\text{wSAA}}(\vec{x})$ cannot be derived (Györfi et al. 2002). However, we can show that the individual lower rate of convergence[8] decreases exponentially in p for a specific class of distributions and a general loss function.

Proposition 2.4. *Assume $\mathcal{Q} = \mathcal{D} = \mathbb{R}$, and a data set S_N drawn iid from a distribution $(\vec{X}, D) \in \mathcal{P}^{(l,C)}$. For a loss function $L(q,d) = |q-d|^2$, an individual lower rate of convergence of q^{wSAA} is given as*

$$a_N = N^{-\frac{2l+1}{2l+p}}. \tag{2.8}$$

Proposition 2.4 states that we cannot generally expect a convergence of $q^{\text{wSAA}}(\vec{x})$ that is faster than $N^{-\frac{2l+1}{2l+p}}$, which declines exponentially in p (see Györfi et al. 2002 for details). Therefore, in big data regimes with a large number of features p, convergence may be slow, and the number of observations N required to achieve a certain performance level may be high.

Therefore, in the context of our problem, where the number of historical observations of demand data is relatively small—we cannot expect to have more than $N \approx 260$ relevant observations (assuming one observation describes one week and that data older than five years is no longer relevant), and the number of features p may be large—convergence may be slow and the property of asymptotic optimality appears to have limited practical relevance. Results of an experiment that BK reported for a stylized setting with only three fea-

[8] The individual lower rate of convergence is the fastest rate with which an approach can converge to the optimal solution over all possible distributions of a class $\mathcal{P}$. See Definition A.3 in Appendix A.1 for further details.

tures suggested that their wSAA approaches converge to the full-information optimum after more than 10^4 observations. For regimes with substantially fewer observations, as is the case in our context, the property of asymptotic optimality does not allow inferences to be made regarding the approaches' performance.

In BK's numerical experiments, a weight function based on random forests, for which asymptotic optimality of wSAA could not be shown, led to the best performance. In contrast to the weight functions for which asymptotic optimality can be shown (see Proposition A.1), the random forest weight function is learned from the data, implicitly identifies a subset of the features that are most relevant, and can therefore provide a better measure of similarity, especially in regimes with a small number of historical observations. Based on these considerations and the numerical results BK presented, we propose using the random forest weight function that BK introduced:

$$w_n^{\mathrm{RF}}(\vec{x}) = \frac{1}{L}\sum_{l=1}^{L}\frac{\mathbb{1}[\mathcal{R}^l(\vec{x}) = \mathcal{R}^l(\vec{x}^n)]}{\sum_{j=1}^{N}\mathbb{1}[\mathcal{R}^l(\vec{x}) = \mathcal{R}^l(\vec{x}^j)]} \tag{2.9}$$

for a random forest with L trees and $\mathcal{R}^l(\vec{x})$ the terminal node of tree l containing $\vec{x}$. The numerator in (2.9) captures the instances in which the feature vectors $\vec{x}$ and $\vec{x}^n$ are assigned to the same terminal node in tree l, while the denominator captures the number of training samples in the terminal node of $\vec{x}$. The weight $w_n^{\mathrm{RF}}(\vec{x})$ is computed as an average of this fraction for all L trees of the random forest. This definition ensures normalization of the weights, such that $\sum_n w_n^{\mathrm{RF}}(\vec{x}) = 1$.

By design, wSAA approaches prescribe feasible solutions because they rely on re-optimization over the feasible set $\mathcal{Q}$ for each new instance of $\vec{x}$. While this is clearly an attractive property, it may have a downside in our context, where demand can have either a strong negative or a strong positive trend. Assuming scalar-valued demand realizations and capacity decisions, we can show that $q^{\mathrm{wSAA}}(\vec{x})$ is restricted to convex combinations of optimal solutions for individual demand realizations d_n $(n = 1, ..., N)$—see Proposition A.2 in Appendix A.7.1—so it interpolates between feasible solutions that would have

been optimal in the past. However, in the presence of a positive or negative trend in demand, interpolation may lead to prescriptions that are inferior to those that are based on an approach that allows for extrapolation (see Appendix A.7.2 for additional illustrations and discussion). As we discuss below, in contrast to wSAA, an ERM approach allows for extrapolation but does not guarantee feasible solutions.

2.4.2 Kernelized Empirical Risk Minimization Approach

Based on the general description of the ERM approach (Problem 2.5), we can formulate the following ERM model for our particular capacity planning problem:

$$\begin{aligned}
\min_{\vec{q}(\cdot)\in\mathcal{F},\{y_{ij}^{tn}\}} \; & \lambda||\vec{q}(\cdot)||_{\mathcal{F}}^2 + \frac{1}{N}\sum_{n=1}^{N}\left(\sum_j f_j q_j(\vec{x}^n) - \sum_{t=1}^{T}\left(\sum_{i,j} a_{ij} y_{ij}^{tn} - \sum_i c_i d_i^{tn}\right)\right) \\
\text{s.t.} \; & \sum_j y_{ij}^{tn} \leq d_i^{tn} \; \forall i,n,t \\
& \sum_i y_{ij}^{tn} \leq q_j(\vec{x}^n) \; \forall j,n,t \\
& y_{ij}^{tn} \geq 0 \; \forall i,j,n,t \\
& y_{ij}^{tn} = 0 \text{ if } i < j \; \forall n,t \\
& q_j(\vec{x}^n) \geq 0 \; \forall j,n.
\end{aligned} \tag{2.10}$$

Because Problem 2.5 is ill-posed for many function spaces and is prone to overfitting for function spaces of sufficient complexity, we follow the standard procedure of Tikhonov regularization (Vapnik 1998) and include a regularization term $\lambda||\vec{q}(\cdot)||_{\mathcal{F}}^2$ in the objective function of (2.10).

Proposition 2.5. *The objective function of the ERM approach is jointly convex in* $\vec{q}(\cdot)$.

Solving the ERM model (2.10) requires choosing a function space $\mathcal{F}$, which we assume to be a Banach space (similar as for Equation 2.3). In the following we provide a solution for linear and non-linear function spaces through kernelization, which corresponds to an (implicit) projection of feature vectors into

a reproducing kernel Hilbert space, employing the well-known machine learning *kernel trick* (e.g., as used in kernel ridge regression and support vector machines, see Smola and Schölkopf 2004 or Hastie et al. 2009 for details).

Definition 2.1. *A kernel $K(\vec{x}_1, \vec{x}_2)$ is a symmetric, positive semi-definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.*

The Moore-Aronszajn theorem states that each kernel K following this definition is a reproducing kernel and that it defines a unique reproducing kernel Hilbert space $\mathcal{H}_K$—see Part I, Section 2 (4) in Aronszajn (1950)—which is a function space. In Appendix A.7.4 we provide a formulation of the ERM model for a linear function space. The optimal solution (provided in Theorem A.2) depends only on the inner product of feature vectors $\vec{x}$, so an implicit projection of these feature vectors into a reproducing kernel Hilbert space can be accomplished by replacing the inner product with the kernel function, leading to the solution stated in Theorem 2.1.

Theorem 2.1. *Assume a reproducing kernel function K with reproducing kernel Hilbert space $\mathcal{H}_K$. Then the optimal kernelized solution to the ERM approach is:*

$$\vec{q}^{kERM}(\vec{x}) = \sum_{n=1}^{N} \vec{u}^n K\left(\vec{x}^n, \vec{x}\right) - \vec{b}, \tag{2.11}$$

for some $\vec{b}$, where the components of $\vec{u}^n$ are defined as $u_j^n = \frac{1}{2\lambda_j}\Big(\sum_{t=1}^{T}(\beta_j^{tn}) + \epsilon_j^n - f_j\Big)$, and β_j^{tn}, ϵ_j^n is the optimal solution to the kernelized dual problem

$$\begin{aligned}
\max_{\{\alpha_i^{tn}\},\{\beta_j^{tn}\},\{\epsilon_j^n\}} L_{dual} := & -\sum_{j=1}^{I}\lambda_j \sum_{p,q=1}^{N}\Big(u_j^p u_j^q K(\vec{x}^p, \vec{x}^q)\Big) + \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{t=1}^{T}(c_i - \alpha_i^{tn})d_i^{tn} \\
s.t.\ & \alpha_i^{tn},\ \beta_j^{tn},\ \epsilon_j^n \geq 0\ \forall i, n, t \\
& \alpha_i^{tn} + \beta_j^{tn} \geq a_{ij}\ \forall i \geq j,\ \forall n, t \\
& \sum_{n=1}^{N} u_j^n = 0\ \forall j.
\end{aligned} \tag{2.12}$$

Corollary 2.1. *The objective function L_{dual} of the kernelized dual problem is concave in the Lagrange multipliers α_i^{tn}, β_j^{tn}, ϵ_j^n.*

Problem 2.12 is a quadratic optimization problem that can be solved efficiently using standard non-linear programming techniques. Given the optimal $\vec{w}^n$, the offset $\vec{b}$ can be determined by applying (2.11) to (2.10) and solving the resulting problem, which is linear in b_j.

The performance of kERM and the applicable performance guarantees depend on the choice of the kernel function that defines $\mathcal{H}_K$. Similar to the weight functions used in wSAA, the kernel function $K(\vec{x}_1, \vec{x}_2)$ can be interpreted as a similarity function that measures the similarity between feature vectors $\vec{x}_1$ and $\vec{x}_2$ (see Section 1.2.4 in Vert et al. 2004). In general, we can employ either *data-independent kernels*, for which the kernel function $K(\vec{x}_1, \vec{x}_2)$, following Definition 2.1, is not learned from the data, or *data-dependent kernels*, for which the kernel function $K_{S_N}(\vec{x}_1, \vec{x}_2)$ depends on the training data S_N (see, e.g., Bengio et al. 2004). Note the similarity to the weight functions presented in Section 2.4.1 that included data-independent weight functions—see Proposition A.1 in Appendix A.7.1—and the random forest weight function (2.9), which is learned from the data.

The class of data-independent kernels includes the well-known general-purpose linear, polynomial, or radial basis function (RBF) kernels (see, e.g., Schölkopf and Smola 2002). When using these kernels, we can derive general bounds on the Rademacher complexity of the respective function space for the kERM approach, which allows us to establish out-of-sample performance guarantees. In Section 2.5.1 we derive such guarantees for kERM using data-independent kernels.

However, using these data-independent kernels for kERM may have two limitations. First, kERM is restricted to the function space $\mathcal{F}$, corresponding to the chosen kernel $K(\vec{x}_1, \vec{x}_2)$, so it will determine the optimal $\vec{q}^*(\vec{x}) \in \mathcal{F}$, which is not necessarily the optimal prescription function of all possible function spaces. Second, kERM with data-independent kernels may not perform well in regimes with a small number of historical observations and a large number of features because the data-independent kernels assign equal importance to all individual features, although some features are likely to have more prescriptive content than others.

We can overcome the first limitation by employing a (data-independent)

universal kernel (Micchelli et al. 2006), for which the associated reproducing kernel Hilbert space is dense in the space of continuous functions. In Section 2.5.2 we show that such a kernel allows us to obtain the optimal prescription function of all continuous functions for $N \to \infty$, a characteristic that is commonly termed *universal approximation property* (Micchelli et al. 2006, see Section 2.5.2 for further details). While this is clearly an attractive theoretical property, we show that the rate of convergence may decline exponentially in p, similar to our results in Proposition 2.4 for wSAA. Therefore, in regimes where p is large, convergence may be slow, and for small numbers of historical observations the universal approximation property appears to have limited practical relevance—in such high-dimensional regimes $\vec{q}^{\,\mathrm{kERM}}(\vec{x})$ based on a universal kernel may be a poor approximation of the optimal prescription function $\vec{q}^{\,*}(\vec{x})$, so it may lead to inferior prescriptions.

We therefore propose the random forest kernel, a kernel that is learned from the training data and is similar to the random forest-based weight function defined in Section 2.4.1. This kernel reduces the dimensionality of the problem through feature selection and, in contrast to data-independent kernels, accounts for varying predictive contents of the individual features. Therefore, it can lead to a better measure of the similarity between feature vectors, which may translate into performance that is superior to that of kERM with data-independent kernels. While for the random forest-based weight function we could propose only that feature selection enhances the performance of wSAA, research in the domain of machine learning has provided more substantial evidence of the importance of feature selection for kernelized approaches (e.g., Weston et al. 2000 and Chen and Lin 2006). Breiman (2000) was first to mention the random forest kernel, and its applicability has been demonstrated in various studies (e.g., Vens and Costa 2011, Gray et al. 2013, Davies and Ghahramani 2014, Scornet 2016). Analogous to the random forest-based weight function (2.9), we define the random forest kernel as:

$$K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2) = \frac{1}{L} \sum_{l=1}^{L} \frac{\mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_2)]}{\sum_{j=1}^{N} \mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_j)]}, \tag{2.13}$$

where $\mathcal{R}^l(\vec{x})$ describes the terminal node of tree l for feature vector $\vec{x}$.

Proposition 2.6. *$K^{RF}(\vec{x}_1, \vec{x}_2)$ is symmetric and positive semi-definite, so $K^{RF}(\vec{x}_1, \vec{x}_2)$ defines a reproducing kernel Hilbert space $\mathcal{H}_{K^{RF}}$.*

Based on Proposition 2.6, we can select K^{RF} as a kernel function and use the expression stated in Theorem 2.1 to obtain the kERM solution. For the real-world application with multivariate demand presented in Section 2.6, we found that using a multivariate random forest model leads to the best performance. In Section 2.6.6 we present a performance comparison of kERM with different kernel functions, including data-independent linear and polynomial kernels and the universal RBF Gauss kernel. Unfortunately, the random forest kernel does not exhibit the universal approximation property, and we can only derive out-of-sample performance guarantees using a sample-splitting approach (see Section 2.5.3).

In Section 2.4.1 we showed (for the case of scalar-valued demand realizations and capacity decisions) that wSAA interpolates between feasible solutions that would have been optimal in the past and that this may be a drawback in situations where demand has a strong positive or negative trend. In contrast, kERM allows for extrapolation because it provides an explicit function $\vec{q}^{\,\mathrm{kERM}}(\vec{x})$ and does not rely on re-optimization for each new instance of $\vec{x}$. However, the ability to extrapolate also entails that the feasibility of the prescribed solutions (i.e., $\vec{q}^{\,\mathrm{kERM}}(\vec{x}) \in \mathcal{Q}$) for a new instance $\vec{x}$ is not guaranteed when the decision space is constrained, so the prescriptions may turn out to be infeasible (see BK). In our case, the solution space is only restricted to $\mathbb{R}^I_+$, and negative capacity prescriptions can be corrected through postprocessing, as BK described. In our numerical experiments we did not face any negative capacity prescriptions.

Based on this brief discussion of the theoretical and practical properties of kERM (that we detail in Sections 2.5 and 2.6), we perceive kERM to be a promising approach to solving our complex capacity planning problem because it does not require re-optimization for every new instance of the feature vector $\vec{x}$ but derives a prescription function $\vec{q}(\cdot)$ that allows for extrapolation.

2.5 Performance Guarantees for the Kernelized ERM Approach

This section provides out-of-sample performance guarantees for kERM using data-independent kernels, the universal RBF Gauss kernel, and the data-dependent random forest kernel, as defined in Section 2.4.2. We base our analyses on the concept of multivariate Rademacher complexities and the general performance guarantees introduced in BK, which we adapt to our capacity management problem.

While the capacity planning problem stated in Section 2.3 does not include upper bounds on the prescribed capacity, it is reasonable to assume some upper bounds on capacity and demand, although they may be large.[9]

Definition 2.2. *Let $\mathcal{Q}$ and $\mathcal{D}$ be defined as*

$$\begin{aligned} \mathcal{Q} &= \{\vec{q} = \{q_i\} \in \mathbb{R}^I : 0 \leq q_i \leq \bar{q}\ \forall i\} \\ \mathcal{D} &= \{\boldsymbol{d} = \{d_i^t\} \in \mathbb{R}^{I \times T} : 0 \leq d_i^t \leq \bar{d}\ \forall i, t\}. \end{aligned} \tag{2.14}$$

Based on this definition, we can show that the loss function $L(\vec{q}, \boldsymbol{d})$ is bounded and equi-Lipschitz, which allows us to apply the theoretical results established by BK and derive out-of-sample performance guarantees for the function spaces defined by the various data-independent kernels and the random forest kernel (using a sample-splitting approach).

Lemma 2.1. *The loss function $L(\vec{q}, \boldsymbol{d})$ is*

a) bounded over $\mathcal{Q}$ and $\mathcal{D}$, and there is some $\bar{l} < \infty$ such that

$$\sup_{\vec{q} \in \mathcal{Q}, \boldsymbol{d} \in \mathcal{D}} L(\vec{q}, \boldsymbol{d}) \leq \bar{l}, \tag{2.15}$$

b) equi-Lipschitz in $\vec{q}$ over $\mathcal{Q}$ and $\mathcal{D}$, and there is some $M_{Lip} < \infty$ such

[9] The assumption is reasonable, as the amount of reservable capacity is typically limited because of factors like factory space, and the demand is typically limited by the current market or the maximum number of customers and their demand.

that

$$\sup_{\vec{q},\vec{q}'\in\mathcal{Q},\vec{q}\neq\vec{q}',\boldsymbol{d}\in\mathcal{D}} \frac{|L(\vec{q},\boldsymbol{d}) - L(\vec{q}',\boldsymbol{d})|}{||\vec{q}-\vec{q}'||_\infty} \leq M_{Lip}. \tag{2.16}$$

2.5.1 Data-independent Kernels

When solving Problem 2.12 using a data-independent kernel $K(\vec{x}_1, \vec{x}_2)$ with reproducing kernel Hilbert space $\mathcal{H}_K$, the function space, over which kERM optimizes, is given as $\mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$, with

$$\begin{aligned} \mathcal{F}_U &= \left\{\vec{q}_U(\cdot) : \mathcal{X} \to \mathbb{R}^I : q_{U,j}(\cdot) \in \mathcal{H}_K\right\}, \text{ and} \\ \mathcal{F}_C &= \left\{\vec{q}_C(\cdot) : \mathcal{X} \to \mathbb{R}^I : \vec{q}_C(\vec{x}) = -\vec{b}\right\}. \end{aligned} \tag{2.17}$$

Theorem 2.2 provides an out-of-sample performance guarantee for the kERM prescription function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ by bounding the true risk.

Theorem 2.2. *(Following BK) Assume S_N, generated by iid sampling from a joint distribution of $\vec{X} \times \boldsymbol{D}$, $L(\vec{q},\boldsymbol{d})$, as defined in (2.2); a function space $\mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$ with $||\vec{b}||_\infty \leq B_C$, $||q_{U,j}||_K \leq B_U\ \forall j$; and let $\delta > 0$. Then, with probability of at least $1-\delta$ for any function $\vec{q}(\cdot) \in \mathcal{F}$, the true risk is bounded as*

$$\begin{aligned} R(\vec{q}(\cdot)) \leq & R_N(\vec{q}(\cdot)) + 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} \\ & + M_{Lip}\left(\frac{2\sqrt{2}IB_Ce}{\sqrt{\pi}\sqrt{N}} + \frac{2IB_U}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{n=1}^{N} K(\vec{x}^n,\vec{x}^n)}\right), \end{aligned} \tag{2.18}$$

where $\bar{l}$ is the bound and M_{Lip} is the Lipschitz constant of $L(\vec{q},\boldsymbol{d})$.

Theorem 2.2 provides a bound on the true risk $R(\vec{q}(\cdot))$ when using a prescription function $\vec{q}(\cdot)$. The first term on the RHS of (2.18) is the in-sample empirical risk, and the third term captures the complexity of the function space $\mathcal{F}$, measured as Rademacher complexity and scaled by the Lipschitz-constant M_{Lip} of the loss function $L(\vec{q},\boldsymbol{d})$. These two terms reflect the well-known bias-variance trade-off the decision-maker faces when choosing the model complexity (see Hastie et al. 2009 for details). The second

term in (2.18) expresses the finite sample bias; because the finite sample bias depends only on the number of historical observations N and the aspired confidence $1-\delta$ (scaled by the bound on the loss function $\bar{l}$), it is independent of the function space $\mathcal{F}$. It is straightforward to see that the performance bound tends to be loose when the number of historical observations N is comparatively small.[10] Therefore, similar to the guarantees of asymptotic optimality (Section 2.4.1) or the universal approximation property (Section 2.5.2), these performance guarantees do not allow inferences to be made regarding the performance of the approach in small data regimes. This limitation appears to be a general issue for many problems in OM, where one can hardly expect to obtain a sufficiently large number of relevant observations of historical demand and corresponding features.[11]

However, the out-of-sample guarantee ensures consistency of the ERM principle; the risk of the estimated function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ will converge in probability[12] to the risk of the function $\vec{q}^{\,*}_{\mathcal{F}}(\cdot) \in \mathcal{F}$, which solves (2.3) for $N \to \infty$ —see Proposition A.6 in Appendix A.7.6—and allows us to bound the rate of convergence. If the data-independent kernel function is bounded, the kERM solution converges with $1/\sqrt{N}$ (see Appendix A.7.6 for details). While the performance guarantees ensure convergence of kERM when a fixed function space $\mathcal{F}$ is used, this convergence occurs only to the best function of $\mathcal{F}$. The next section presents a universal kernel that allows us to overcome this limitation.

2.5.2 The Universal RBF Gauss Kernel

The reproducing kernel Hilbert space associated with a universal kernel is dense in the space of continuous functions, so any continuous function can be

[10] The out-of-sample guarantees provided decrease with a rate of $1/\sqrt{N}$. However, they also contain a term $\propto \bar{l}/\sqrt{N}$, with $\bar{l}$ being a bound on the loss function that requires a large N to yield practically relevant bounds.

[11] We do not expect data to be relevant for more than five years, corresponding to a maximum of 1500 daily observations.

[12] Convergence in probability of the true risk of $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ to the optimal true risk $R(\vec{q}^{\,*}(\cdot))$ means that the probability that the absolute difference is larger than some $\epsilon > 0$ converges to zero (see Section 1.11.1 in Vapnik 1998).

approximated with arbitrarily high accuracy. This characteristic is also called *universal approximation property* (Micchelli et al. 2006).

Proposition 2.7. *The function space*

$$\mathcal{F}_{K_{RBFG}} = \left\{ \vec{q}(\cdot) : \mathcal{X} \to \mathbb{R}^I : q_j(\vec{x}) = q_{\mathcal{H},j}(\vec{x}) - b_j,\ q_{\mathcal{H},j}(\cdot) \in \mathcal{H}_{K_{RBFG}} \right\} \quad (2.19)$$

is dense in $C(\mathcal{X}, \mathbb{R}^I)$ for a reproducing kernel Hilbert space $\mathcal{H}_{K_{RBFG}}$ corresponding to the RBF Gauss kernel $K_{RBFG}(\vec{x}_1, \vec{x}_2) := \exp\left(-\gamma^2 |\vec{x}_1 - \vec{x}_2|^2\right)$ and compact $\mathcal{X} \subset \mathbb{R}^p$.

Proposition 2.8. *Assume S_N, generated by iid sampling from a joint distribution of $\vec{X} \times \boldsymbol{D}$, $L(\vec{q}, \boldsymbol{d})$, as defined in (2.2); a function space $\mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$ with $||\vec{b}||_\infty \leq B_{C,N}$, $||q_{U,j}||_K \leq B_{U,N}\ \forall j$ with sequences $B_{U,N}, B_{C,N}$ that satisfy $\lim_{N\to\infty} B_{U,N}, B_{C,N} = \infty$ and $\lim_{N\to\infty} B_{U,N}/\sqrt{N}, B_{C,N}/\sqrt{N} = 0$; and $K_{RBFG}(\vec{x}_1, \vec{x}_2)$ the RBF Gauss kernel. Then kERM fulfills the universal approximation property and the risk convergences for $N \to \infty$ in probability towards the risk of*

$$\vec{q}^*(\cdot) = \underset{\vec{q}(\cdot) \in C(\mathcal{X}, \mathbb{R}^I)}{\arg\min}\ \mathbb{E}_{\vec{X} \times \boldsymbol{D}} \left[L\left(\vec{q}(\vec{X}), \boldsymbol{D} \right) \right]. \quad (2.20)$$

Clearly, the universal approximation property of kERM with the RBF Gauss kernel is a desirable property. However, we find that the convergence rate may decrease exponentially with dimensionality p; in fact, a similar individual lower rate of convergence, as presented in Proposition 2.4, applies.

Proposition 2.9. *Assume $\mathcal{Q} = \mathcal{D} = \mathbb{R}$ and a data set S_N drawn iid from a distribution $(\vec{X}, D) \in \mathcal{P}^{(l,C)}$. For a loss function $L(q, d) = |q - d|^2$, and the RBF Gauss kernel $K_{RBFG}(\vec{x}_1, \vec{x}_2) := \exp\left(-\gamma^2 |\vec{x}_1 - \vec{x}_2|^2\right)$ with $\gamma > 0$, an individual lower rate of convergence of $R(q^{kERM}(\cdot)) \to \inf_{q(\cdot):\mathcal{X}\to\mathbb{R}} R(q(\cdot))$ is given as*

$$a_N = N^{-\frac{2l+1}{2l+p}}. \quad (2.21)$$

These results are similar to those presented in Section 2.4.1 for wSAA with data-independent weight functions: We obtain a strong theoretical result

regarding the performance of kERM using an RBF Gauss kernel in the limit of $N \to \infty$, but the convergence may be slow, and the RBF Gauss kernel may not be an appropriate choice in regimes with large p and small N. In such a regime, a data-dependent kernel, such as the random forest kernel, may lead to superior performance.

2.5.3 The Data-dependent Random Forest Kernel

The random forest kernel defined in (2.9) is data-dependent and, as such, is not universal (see Appendix A.7.5), so the universal approximation property does not apply to kERM using the random forest kernel; however, we can derive out-of-sample performance guarantees. The out-of-sample performance guarantee presented in Theorem 2.2 holds for all kernel functions that are independent of the training data, so it does not apply to the random forest kernel. One approach to establish performance bounds could be to consider the training of kERM with the data-dependent random forest kernel as an instance of *kernel learning* (Lanckriet et al. 2004), which considers the function space of all possible kernel functions that can be learned from the data. However, because our kernel function (2.13) depends on both the random forest and the data set, due to the normalization, we would have to consider the function space of all kernel functions defined by all possible random forests and all possible data sets defining the normalization. The Rademacher complexity of such a function space does not converge to zero for $N \to \infty$.

One way to circumvent this issue is to follow a sample-splitting approach, which entails using $N_{RF} < N$ data samples of S_N to train the random forest and to derive the kernel function (2.13), including the normalization, and using the remaining $N - N_{RF}$ data samples of S_N to learn the kERM prescription function. Then, from the perspective of kERM, the random forest kernel is fixed and independent of the data set $S_{N-N_{RF}}$, and we can apply the results of Theorem 2.2 to derive an out-of-sample performance guarantee (Theorem 2.3).

Theorem 2.3. *(Following BK) Assume $S_N = S_{N_{RF}} \uplus S_{N-N_{RF}}$, generated by iid sampling from a joint distribution of $\vec{X} \times \boldsymbol{D}$, $L(\vec{q}, \boldsymbol{d})$, as defined in (2.2); a function space $\mathcal{F}^{RF} = \mathcal{F}_U^{K^{RF}} + \mathcal{F}_C$ with $||\vec{b}||_\infty \leq B_C$, $||q_{U,j}||_K \leq B_U \; \forall j$,*

$K^{RF}(\vec{x}_1, \vec{x}_2)$, as defined in (2.13) and computed using $S_{N_{RF}}$; and let $\delta > 0$. Then, with probability of at least $1 - \delta$ for any function $\vec{q}(\cdot) \in \mathcal{F}^{RF}$, the true risk is bounded as

$$\begin{aligned} R(\vec{q}(\cdot)) \leq & R_{N-N_{RF}}(\vec{q}(\cdot)) + 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2\,(N - N_{RF})}} \\ & + M_{Lip}\left(\frac{2\sqrt{2}IB_C e}{\sqrt{\pi}\sqrt{N - N_{RF}}} + \frac{2IB_U}{\sqrt{N - N_{RF}}}\right), \end{aligned} \tag{2.22}$$

where $\bar{l}$ is the bound and M_{Lip} is the Lipschitz constant of $L(\vec{q}, \boldsymbol{d})$.

In contrast to the performance guarantee presented in Theorem 2.2, the terms that capture the finite sample bias and the complexity of the function space (second and third term on the RHS of Equation 2.22) decrease with the number of data samples $(N - N_{RF})$ used to compute the kERM prescription function. A decision-maker that has N historical demand observations therefore faces a trade-off in choosing the number of samples N_{RF} to train the random forest kernel: a large N_{RF} increases the second and third term on the RHS of (2.22), while a small N_{RF} may lead to a larger in-sample risk $R_{N-N_{RF}}(\vec{q}(\cdot))$, because the random forest kernel is only trained on $S_{N_{RF}}$.

Based on the result of Theorem 2.3, we conclude that kERM with a random forest kernel trained on a fixed $S_{N_{RF}}$ is consistent; that is, the kERM prescription function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ will converge with $1/\sqrt{N - N_{RF}}$ in probability to a function that minimizes the true risk (see Appendix A.7.6 for further details). For a fixed function space this rate of convergence is independent of the number of features p, but convergence occurs only to the best function of the function space $\mathcal{F}^{\mathrm{RF}}$, not to the best prescription function of all continuous functions, because a random forest kernel is not universal.

2.6 A Real-World Application and Numerical Insights

This section applies wSAA and kERM to the capacity planning problem of a logistics service provider. We use historical demand data and realistic values of our case company's cost parameters to demonstrate these approaches' applicability and to compare their performance to two traditional two-step approaches and a conventional SAA approach that does not incorporate feature data. To clarify the underlying drivers of the approaches' performance and to generate insights that go beyond the specific parameter settings of our case company, we present the results of additional numerical analyses in which we vary the (exogenous) cost parameters of our model while maintaining the case company's historical demand and feature data. The results of these analyses provide insights into the approaches' performance drivers and allow us to evaluate their robustness.

2.6.1 Problem Statement and Motivation

Our research was inspired by our work with a logistics service provider in Germany that collects, sorts, and delivers mail (letters, parcels), newspapers, advertising material, and so on. We focus on the sorting operations that approximately fifteen workers carry out in the company's main facility. The company receives an average of approximately 175,000 items per day that are sorted manually (service line 3), semi-automatically (service line 2), or on a fully automated line (service line 1). While sufficient sorting-machine capacity for each service line is available, the operation of these machines and the manual sorting require a certain level of staffing (i.e., capacity). Labor costs and the required skill levels differ among the three service lines, as operation of the fully automated sorting machine requires highly skilled staff, while semi-automatic and manual sorting have lower skill requirements. Since the staff of service line 1 can also operate service lines 2 and 3, and the staff of service line 2 can also operate service line 3, the company has an upgrade option, as described in Section 2.3. Every week, the company has to determine the

Table 2.1: Logistics provider parameter setting.

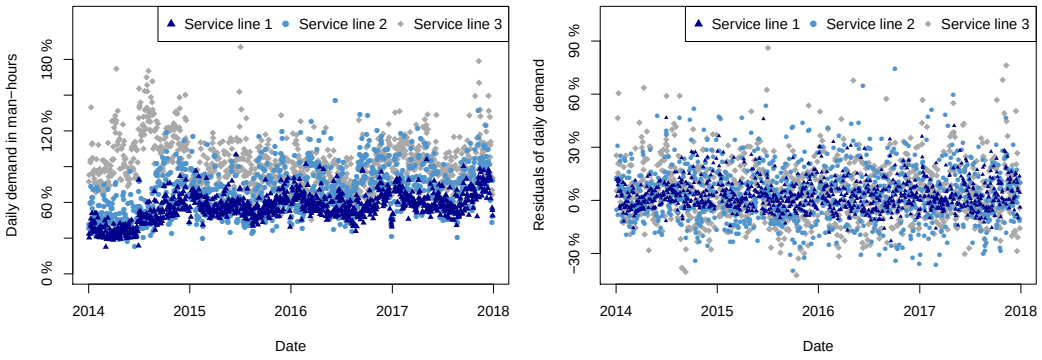
f	v	p	c	$a_{i,i}$
$\begin{pmatrix} 200 \\ 150 \\ 75 \end{pmatrix}$	0	$\begin{pmatrix} 1000 \\ 100 \\ 20 \end{pmatrix}$	$\begin{pmatrix} 500 \\ 50 \\ 10 \end{pmatrix}$	$\begin{pmatrix} 1500 \\ 150 \\ 30 \end{pmatrix}$

staffing levels (capacity) of each production line for the subsequent week that will lead to a fixed and constant capacity for each service line on each day ($t = 1, .., T = 5$). Demand that arrives on day t is processed by the staff designated to a service line, while demand that exceeds the staff capacity of the designated service line can be "upgraded" to more expensive service line staff. All items that arrive during day t must be processed by the end of the day—if necessary, by employing overtime that is not only costly in terms of wages but is also highly undesirable because of its negative impact on employee retention. The company faces a severe shortage in labor supply, so they want to limit overtime operations to maintain employee satisfaction. In the past the company used a relatively simple approach for taking their capacity decisions. Based on historical demand they obtained an estimate of the number of different mail items and converted these into capacity requirements for individual weekdays, which served as a basis for the weekly capacity plans. This process was carried in a mostly manual way and relied strongly on the planners' previous experience and expertise; the company did not use sophisticated tools for forecasting or capacity planning. Ad hoc capacity adjustments (by switching or rescheduling shifts, or recruitment of additional temporal workers) occurred frequently.

Using information provided by the company, we obtained estimates of the company's revenues and costs (Table 2.1). The revenue per item sorted amounts to approximately 0.1 EURO. The (full) cost per worker ranges from 15 EURO to 40 EURO per hour, depending on the skill set and other factors. Jointly with the company, we defined a penalty cost of 0.05 EURO per unsorted

item to reflect the negative impact of overtime on employee satisfaction.[13] We also assumed a capacity usage cost of $v_j = 0$ because planned capacity is fixed and must be paid for during the entire week.

Because it is similar to the capacity planning problem described in Section 2.3, we can employ the traditional two-step approach of estimating a multivariate demand distribution and solving a two-stage stochastic optimization model, as stated in (2.1), to determine the "optimal" capacities of the three production lines for the subsequent week. The practical difficulties of this approach are rooted primarily in the estimation of the multivariate demand distribution as described in Section 2.1. We support our arguments in Section 2.1 with some descriptive analyses of the case company's demand data. Figure 2.1a plots the daily demand of all three production lines (converted into required man-hours) from 2014 to 2017.



(a) Daily demand in required man-hours. (b) Residuals of daily demand.

Figure 2.1: Daily demand and residuals of de-trended and de-seasonalized time series (2014-2017).

We observe trends and seasonal patterns in all three time series, suggesting that demand is non-stationary. A more detailed analysis reveals that the time series contain a superposition of seasonalities at differing frequencies (Appendix A.2) and that variations in daily demand differ considerably for the three production lines. To highlight these variations in daily demand,

[13] This choice of penalty cost results in a total penalty $p_j + c_j$ of 150 percent of revenue, which includes the cost of overtime and the intangible costs of employee dissatisfaction. These penalty costs lead to imposition of high service-level targets.

Figure 2.1b plots the de-trended and de-seasonalized time series that were obtained using a TBATS model (see Appendix A.2 for details).

Across the four years, the coefficients of variation (CV) in the three service lines' daily demands are moderate to low, ranging from $CV_2 = 0.25$ for service line 2 to $CV_1 = 0.17$ for service line 1. The coefficient of correlation (CC) during the entire time period is comparatively low, amounting to $CC \approx 0.18$ for all three combinations of service lines. Although CV and CC are low to moderate across all observations, for shorter time periods we observe substantial variations (see Appendix A.2), but we do not know whether they occur at random or can be explained by certain features. Interviews with experts in the company indicated that some of these variations may be predictable. For instance, in weeks 50 and 51 of each year, demands for service lines 1 and 3 are typically highly correlated, as large amounts of year-end business mail (service line 1) and private holiday mailings (service line 3) arrive. On the other hand, proximity to public holidays often leads to a negative correlation because private mailings increase, while business mailings decrease because businesses' employees tend to take vacation during that time. We expect many such relationships and that they may be predictable, given appropriate features. However, it is difficult to incorporate these relationships into an estimate of the multivariate demand distribution, which illustrates the need for prescriptive approaches that can implicitly incorporate feature-dependent distributions.

2.6.2 Demand Data and Feature Engineering

The case company gave us a historical data set that contained demand d_i^t in number of mail items for each service line i for each day t between 2014 and 2017 to solve the planning problem we described. From this historical data, we constructed a data set $S_N = \{(\boldsymbol{d}^1, \vec{x}^1), ..., (\boldsymbol{d}^N, \vec{x}^N)\}$, with demand matrices $\boldsymbol{d}^n$ in units of man-hours for $N = 209$ weeks and feature vectors $\vec{x}^n \in \mathbb{R}^{162}$. As elements of the feature vectors $\vec{x}^n$, we first constructed date-based features that describe the temporal dimension of the observed demand. In particular, we used the year number and the half, quarter, and month of the

year that contained the particular week as features. The week number may also be a relevant feature, as we learned from interviews with experts that demand is high in some weeks near the end of the year (see Section 2.6.1). We included lagged demands (e.g., demand for each service line in the same week one year ago) in the second group of features to account for the sequential character of the time series. The third group of features encoded information on public holidays, which are also known to affect demand (see Section 2.6.1). We constructed indicators for public holidays and relative indicators (e.g., if a public holiday is a few days before or after the week of interest). A detailed description of the 162 features included in our analysis and an analysis of their importance can be found in Appendix A.3.

2.6.3 Evaluation Procedure

We split the resulting data set S_N into $N = 157$ weeks of training data (2014-2016) and 52 weeks of test data (2017) to facilitate out-of-sample performance evaluation. Because the number of features $p > N$, we face a high-dimensional problem. We evaluated and compared the prescription performance of the following approaches:

1. kERM: Estimate prescription function (2.11) by solving (2.12) with random forest kernel (2.13) and prescribe capacity decisions for each week of the test period.

2. wSAA: Solve (2.7) with random forest weights (2.9) for each week of the test period.

3. SAA: Estimate the SAA prescription of the training data set by solving (2.7) for $w_n(\vec{x}) = 1/N$.

4. SVR-SEO[14]: Train support vector regression models using the random forest kernel (2.13), estimate CV and CC on in-sample residuals, and

[14]Traditional two-step approaches are also referred to as *sequential estimation and optimization* (SEO).

predict multivariate demand distributions for each week of the test period. Solve (2.1) using Monte Carlo sampling and SAA on $N_{MC} = 300$ samples.

5. ARIMA-SEO: Train ARIMA time series models, estimate CV and CC on in-sample residuals, and predict multivariate demand distributions for each week of the test period. Solve (2.1) following the same procedure as for SVR-SEO.

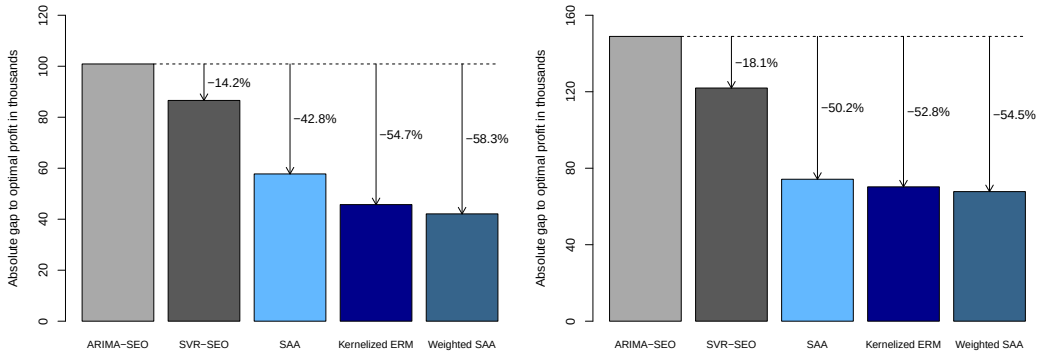
We determined the maximum achievable profit for all prescribed capacity decisions by solving the second stage of Problem 2.1 for each day of the test period, with the total profit as the sum of weekly profits over all weeks. We also calculated the ex-post optimal profit $\Pi^*(\boldsymbol{d})$ for the test period, so we can report the absolute gap to optimal profit $\Delta_{\Pi,\text{abs}} = \Pi^*(\boldsymbol{d}) - \Pi(\vec{q}, \boldsymbol{d})$ for all approaches.[15]

2.6.4 Results and Discussion

Figure 2.2a shows the absolute gap to optimal profit in the 2017 test period that is associated with the approaches' capacity prescriptions.[16] The results suggest that, in our setting, the gap to optimality can be reduced by more than half by using prescriptive analytics approaches (kERM, wSAA) instead of traditional two-step approaches (ARIMA-SEO, SVR-SEO). wSAA leads to the lowest performance gap, which is 3.6 percent lower than that of kERM, the second-best approach, and more than 58 percent lower than that of ARIMA-SEO, the approach with the worst performance. The potential improvements, documented in Figure 2.2a, could lead to a substantial increase in the company's financial performance. As a provider of basic logistics services, the

[15] While one could also use the relative gap $\Delta_{\Pi,\text{rel}} = 1 - \frac{\Pi(\vec{q},\boldsymbol{d})}{\Pi^*(\boldsymbol{d})}$ between actual and optimal profit to compare performances, this quantity is misleading in our real-world application, as fixed costs (e.g., machine costs, factory rental costs) are not considered but are expected to lower the profits Π and Π^* substantially.

[16] The ex-post optimal profit Π^* in 2017 amounts to 3.9 mio. Our calculations do not include overhead costs, so the company's true maximum achievable profit is substantially lower than that reported here.

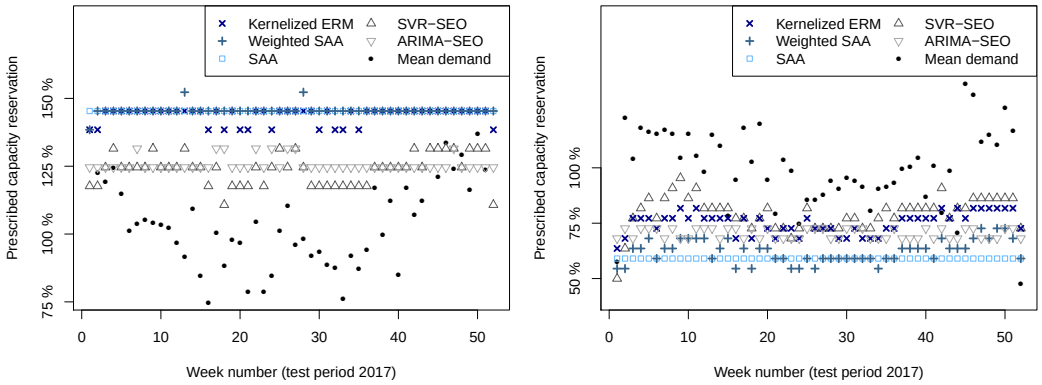


(a) With upgrading. (b) Without upgrading.

Figure 2.2: Absolute gap to optimal profit for all approaches for the real-world application.

company operates on low margins of 2-3 percent and had revenues of approximately EURO 4.3 mio. in 2017. Therefore, the increase in net profits achieved through the use of prescriptive analytics instead of traditional methods could amount to 30-50 percent.

The largest performance improvement occurs between the traditional parametric approaches (SVR-SEO, ARIMA-SEO) and the most basic non-parametric approach (SAA), which does not benefit from features. Comparing the results of SAA and the prescriptive approaches allows us to assess the additional value of incorporating features into the prescription. BK introduced the coefficient of prescriptiveness to quantify "the prescriptive content of data and the efficacy of a policy" (BK, p. 1025) in leveraging this prescriptive content. The coefficient of prescriptiveness P measures the reduction of the gap to optimal profit relative to SAA, which does not use features. Using the results displayed in Figure 2.2a, we can directly compute the coefficients of prescriptiveness $P_{wSAA} = .271$ and $P_{kERM} = .208$ for wSAA and kERM. These values suggest that the features have substantial value and that wSAA is more efficient in exploiting their prescriptive content than kERM is. However, one must take care in drawing conclusions about the two approaches' efficacy. Section 2.6.5 shows that the performances of SAA, wSAA and kERM depend heavily on the exogenous cost parameters and that we obtain $P_{kERM} > P_{wSAA}$



(a) Service line 1. (b) Service line 3.

Figure 2.3: Prescribed capacity of all approaches for the test period (with upgrading). The label "100%" denotes the mean demand of the respective service line for the test period.

for other parameter settings.

To elucidate the approaches' performance and the underlying drivers, Figure 2.3 plots the approaches' prescribed capacities for service lines 1 and 3 in the individual weeks of the test period.[17] Figure 2.3 highlights structural differences in capacity prescriptions, which can explain the large difference in performance between traditional parametric approaches and non-parametric approaches shown in Figure 2.2a. SAA, wSAA, and kERM prescribe capacities for service line 1 that are substantially higher than the mean demand and the prescriptions of the parametric approaches (ARIMA-SEO, SVR-SEO). However, they prescribe much lower capacities for service line 3. A traditional perspective would assume that the capacity prescriptions are composed of two elements: (i) an estimate of the mean demands $\hat{\vec{\mu}}$ (i.e., a forecast) and (ii) some safety capacity $\vec{\Lambda}(\vec{C}_O, \vec{C}_U, \{a_{i,j}\}_{j<i}, \hat{\mathbf{\Sigma}})$ that depends on the overage and underage costs $\vec{C}_O$ and $\vec{C}_U$, the marginal profits for upgrading $\{a_{i,j}\}_{j<i}$, and the estimate of the covariance matrix $\hat{\mathbf{\Sigma}}$.[18] Following this basic newsvendor logic,

[17] Our analysis focuses on service lines 1 and 3 because service line 2 exhibits a superposition of increased and decreased optimal capacity decisions as a result of "outgoing" (to service line 3) and "incoming" (from service line 1) upgraded capacity.

[18] Technically, this decomposition holds true only for ARIMA-SEO and SVR-SEO because they rely on an explicit estimate of a multivariate normal distribution's mean demand and covariance matrix, while the other approaches integrate estimation and optimization.

we can approximate the overage and underage costs of the individual service lines and infer approximate "optimal" service levels of 97 percent for service line 1 and 50 percent for service line 3.[19] Moreover, as Netessine et al. (2002) showed for the case of single-level upgrading, the optimal capacity prescription for the most flexible capacity (service line 1 in our case) should be higher than its optimal newsvendor quantity, and the prescription for the capacity with least flexibility (service line 3 in our case) should be lower than its optimal newsvendor quantity (Proposition 3 in Netessine et al. 2002). While this logic explains the higher (lower) prescriptions for service line 1 (3), it does not explain why the non-parametric approaches prescribe substantially more capacity for service line 1 and less for service line 3. We should be able to explain the varying performances in Figure 2.2a by means of differences in the accuracy of the estimated means $\hat{\vec{\mu}}$ or the safety capacities $\vec{\Lambda}(\cdot)$. An analysis of SVR-SEO's and ARIMA-SEO's forecasting performance reveals that SVR-SEO produces more accurate estimates of $\hat{\vec{\mu}}$ than ARIMA-SEO does, as the former's out-of-sample RMSE is 18 percent lower than that of the latter. However, this higher forecast accuracy only partially translates into better overall performance, as evidenced by the results in Figure 2.2a.

Following our previous line of reasoning, we attribute the differences in capacity prescriptions and performance to varying safety capacities (i.e., $\vec{\Lambda}(\cdot)$) through which the various methods account for both uncertainty and upgrading effects. To disentangle the joint effect of uncertainty and upgrading, we determined the approaches' performance without allowing for upgrading. Figure 2.2b plots the corresponding performance gaps. Comparing the results in Figures 2.2a and 2.2b, we observe a larger optimality gap when upgrading is not permitted, but the increase is roughly the same across all approaches under consideration; the change in performance from the traditional parametric approaches (SVR-SEO, ARIMA-SEO) to the non-parametric approaches (SAA, wSAA, kERM) remains, so it cannot be explained by up-

[19] We calculate the approximate service level as $\mathrm{SL}_i = \frac{C_{U,i}}{C_{U,i}+C_{O,i}}$, with overage and underage cost factors $C_{O,i} \approx f_i/5$ and $C_{U,i} \approx p_i + c_i - v_i - f_i/5$. This approximation underestimates the service level for service line 1, as the option to upgrade lowers the overage cost, thus increasing the "optimal" service level.

grading effects. In fact, the results suggest that the differences in capacity prescriptions and performance are driven by how the approaches explicitly (SVR-SEO, ARIMA-SEO) or implicitly (SAA, wSAA, kERM) account for demand uncertainty. To substantiate this conjecture, we compare SVR-SEO's and ARIMA-SEO's estimated CVs for service line 1 with an implicit measure of uncertainty for SAA[20] and observe a significant difference between ARIMA-SEO ($CV_1^{\text{ARIMA-SEO}} = 0.147$) and SVR-SEO ($CV_1^{\text{SVR-SEO}} = 0.151$) and SAA ($CV_1^{\text{SAA}} = 0.221$), which explains why SAA's capacity prescriptions are larger than those of the traditional parametric approaches for service line 1 (Figure 2.3a).[21] Although we cannot obtain implicit estimates for the CVs of wSAA and kERM, we assume that, like SAA, their higher capacity prescriptions can be explained by their implicit consideration of demand uncertainty based on the empirical demand distribution. The initial results from our case study suggest that the performance differences are rooted primarily in how the approaches account for uncertainty in the multivariate demand, which translates into differing safety capacities (via $\vec{\Lambda}(\cdot)$) and overall capacity prescriptions.

2.6.5 The Impact of Service Levels and Upgrade Profitability

The preceding section discussed the performance of our prescriptive analytics approaches for the (cost) parameters of our case company and conjectured that the performance differences are driven primarily by how the various approaches account for demand uncertainty. In this section we determine whether this conjecture holds under different cost parameters, leading to different implicit service level targets (SL) and upgrade profitabilities $\alpha_i = \frac{a_{i+1,i}}{a_{i,i}}$.

[20] We approximate the empirical distribution of historical demand for service line 1 using a stationary normal distribution, calculate the mean $\hat{\mu}^{\text{SAA}}$ and standard deviation $\hat{\sigma}^{\text{SAA}}$, and derive the CV as $CV^{\text{SAA}} = \hat{\sigma}^{\text{SAA}}/\hat{\mu}^{\text{SAA}}$.

[21] All CVs are estimated relative to the historical mean demand $\hat{\mu}^{\text{SAA}}$ as $CV = \hat{\sigma}/\hat{\mu}^{\text{SAA}}$. Because the safety capacity for SVR-SEO and ARIMA-SEO $\vec{\Lambda}(\cdot)$ is based on the uncertainty of daily demand $\sigma_d^2 = \hat{\mathbf{\Sigma}}_{i,i}$ and the average variation of predicted demand values σ_w^2 in each week, we use $\tilde{\sigma} = \sqrt{\sigma_d^2 + \sigma_w^2}$ (see Appendix A.7.3 for details).

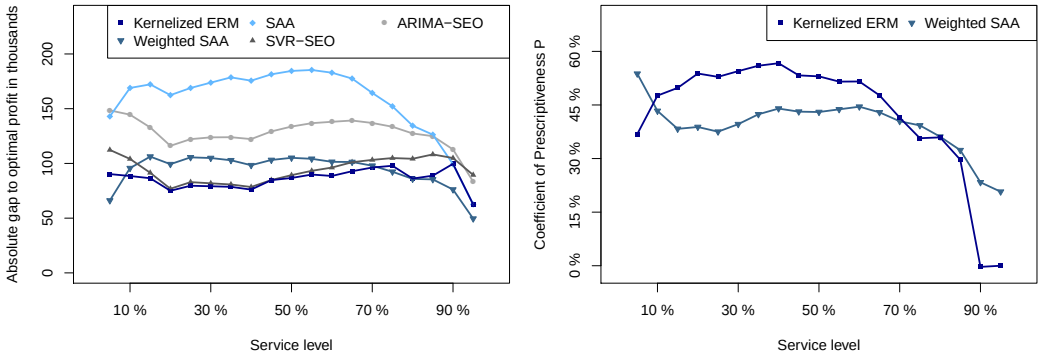
Table 2.2: Parameters for the service level and upgrade profitability variation.

Figure	f	v	p	c	$a_{i,i}$	**SL**	α
2.4a)	$\begin{pmatrix} 2700..142 \\ 1130..60 \\ 500..26 \end{pmatrix}$	$\begin{pmatrix} 40 \\ 30 \\ 20 \end{pmatrix}$	$\begin{pmatrix} 600 \\ 260 \\ 120 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 7.5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 570 \\ 238 \\ 105 \end{pmatrix}$	5%..95%	40%
2.5a)	$\begin{pmatrix} 94500..237 \\ 4750..250 \\ 263 \end{pmatrix}$	$\begin{pmatrix} 40 \\ 30 \\ 20 \end{pmatrix}$	$\begin{pmatrix} 37830..125 \\ 1920..123 \\ 120 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 7.5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 37800..95 \\ 1900..100 \\ 105 \end{pmatrix}$	50%	5%..95%
2.5b)	$\begin{pmatrix} 18900..47 \\ 950..50 \\ 53 \end{pmatrix}$	$\begin{pmatrix} 40 \\ 30 \\ 20 \end{pmatrix}$	$\begin{pmatrix} 37830..125 \\ 1920..123 \\ 120 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 7.5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 37800..95 \\ 1900..100 \\ 105 \end{pmatrix}$	90%	5%..95%

To carve out individual effects and to enhance comprehensibility, we first study the impact of a service level variation for a fixed upgrade profitability. Thereafter, we examine the effect of a variation of the upgrade profitability for a fixed service level across all service lines. Table 2.2 shows the cost parameters used in these analyses and their corresponding approximate service levels and upgrade profitabilities. We carried out extensive additional analyses for heterogeneous service levels and for other combinations of α and SL. The results are reported in Appendix A.8. All of our analyses use the same demand and feature data as in Section 2.6.4.

The results in Figure 2.4a show the approaches' optimality gaps for each service level and provide a more differentiated picture of the approaches' performance than those presented in the preceding section. At a service level of ≈ 50 percent the two featureless approaches (SAA, ARIMA-SEO) lead to substantially lower performance than the approaches that account for feature data (kERM, wSAA, SVR-SEO). Among the latter, kERM and SVR-SEO lead to similar performance levels, and they both outperform wSAA. Because the impact of the safety capacity $\vec{\Lambda}(\cdot)$ is negligible in this service level regime[22],

[22]Without upgrading, the safety capacity $\vec{\Lambda}(\cdot)$ would be equivalent to the newsvendor safety capacity in accounting for uncertainty in demand. While upgrading impacts the prescriptions, all approaches' performance levels are impacted similarly by upgrading, as shown in Figure 2.5a for a service level of 50 percent. Therefore, we focus on the service level effect to explain the performance differences.



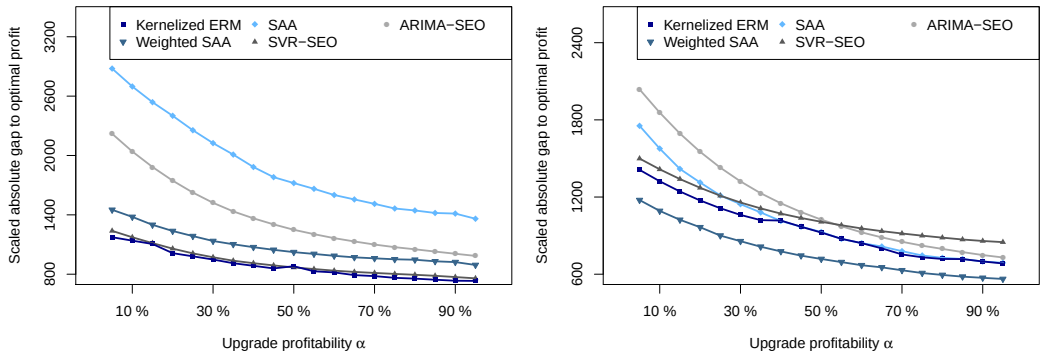
(a) Absolute gap to optimal profit. (b) Coefficient of prescriptiveness.

Figure 2.4: Variation of the service level for $\alpha = 40\%$.

the approaches' performance is dominated by the accuracy of the estimation of $\hat{\vec{\mu}}$, which appears to be higher for kERM and SVR-SEO than for wSAA. At a service level of 50 percent, we obtain $P_{kERM} = .53$ and $P_{wSAA} = .43$, which suggests that the value of features is greater at this particular service level than it is for the higher service levels considered in Section 2.6.4, and that kERM is more efficient in exploiting this value (Figure 2.4b plots the coefficient of prescriptiveness of wSAA and kERM for the varying service levels).[23] However, as we move to higher service levels, SAA's performance improves significantly; for high service levels (e.g., SL=95%), SAA achieves a performance level similar to that of the prescriptive approaches kERM and wSAA, despite being a featureless approach. Thus, the value of incorporating features declines because it becomes more important to account for demand uncertainty to improve safety capacity estimations $\vec{\Lambda}(\cdot)$ (which increase non-linearly in the service level) than it is to obtain accurate estimates of $\hat{\vec{\mu}}$. This effect is also reflected in decreasing values of the coefficient of prescriptiveness for service levels higher than 60 percent (Figure 2.4b). This finding matches our observation in Section 2.6.4, where we considered a regime with comparatively high service levels, and supports the conjecture that the non-parametric approaches are better at accounting for demand uncertainty, which leads to substantial performance improvements over those of traditional non-parametric

[23] Appendix A.8.3 explores the statistical confidence in these results.

approaches. The results in Figure 2.4a and Figure 2.4b also show that, for a service level range of 10-70 percent, kERM outperforms wSAA, while the latter results in better performance for high or very low service levels. For service levels of 90 percent and above, Figure 2.4b suggests that kERM no longer exploits the prescriptive content of the features, resulting in $P_{kERM} \approx 0$. In this regime, and at least for this particular data set, it appears for kERM to be optimal to choose a higher regularization parameter, and therefore to adapt less to the variations of the historical observations. Doing so has two consequences: It leads to higher average capacity prescriptions, driven by a higher (implicit) safety buffer, and a stronger regularization suppresses the impact of features on (the implicit) estimates of $\hat{\tilde{\mu}}$.[24] This leads to the interesting effect that, at least in this particular instance, kERM results in the same prescriptions and the same performance as SAA. At this service level, wSAA can still exploit some of the prescriptive content of the features, which leads to a slightly better performance compared to kERM and SAA. We will see in Section 2.6.6 that this effect is mainly associated with the random forest kernel.



(a) Service level SL=50%. (b) Service level SL=90%.

Figure 2.5: Variation of the upgrade profitability α.

Figures 2.5a and 2.5b plot the results for the variation of the upgrade prof-

[24] To explain this effect, consider the breakdown of kERM's prescriptions into forecasting and safety buffer estimations. Higher regularization leads to higher bias and, therefore, to higher in-sample errors for the forecasting part, which leads, in turn, to higher safety buffers.

itability α at fixed service levels of 50 percent and 90 percent, respectively. Because the optimal profits vary strongly across α, we scale the absolute gap to optimal profit to obtain meaningful results.[25] Overall, we find that all approaches benefit similarly from an increase in the upgrade profitability. A closer inspection of the capacity prescriptions revealed that all approaches prescribe an increasing amount of capacity for service line 1 as α increases; at higher levels of α it becomes more attractive to use the capacity of service line 1 to meet the demand of service lines 2 and 3. A similar argument as made before can be made to explain the performance differences observable in Figures 2.5a and 2.5b. At a service level of 50 percent (Figure 2.5a), an accurate estimate of $\hat{\vec{\mu}}$ is central to achieving good capacity prescriptions, which explains why the feature-based approaches (SVR-SEO, wSAA and kERM) outperform SAA and ARIMA-SEO. In this setting, kERM dominates all other approaches. We observe a different picture for a service level of 90 percent (Figure 2.5b), where it is more important to obtain better estimations of the safety capacity $\vec{\Lambda}(\cdot)$ than it is to exploit the predictive content of the features to generate accurate demand forecasts. In this setting, wSAA clearly outperforms all other approaches, and kERM leads to almost identical performances as SAA for $\alpha \geq 40$ percent, an outcome that is consistent with the results displayed in Figure 2.4b, where we see $P_{kERM} \approx 0$ at a service level of 90 percent, which again can be explained by the "regularization effect".

In summary, all of our results indicate that the approaches' performance depends on how well they are able to (implicitly) estimate $\hat{\vec{\mu}}$ and $\vec{\Lambda}(\cdot)$, that the importance of estimating $\hat{\vec{\mu}}$ and $\vec{\Lambda}(\cdot)$ depends predominantly on the service levels, and that the upgrade profitability has a much less pronounced effect than the service levels. However, we must be careful about generalizing this finding beyond our specific case, as we found a relatively low demand correlation (see Section 2.6.1) in our data set and cannot rule out that the performance effect of upgrading may be more pronounced when the demand correlation is high

[25] Because a meaningful measure of performance is the gap to optimality in units of average overage and underage costs, we divide the gap to optimal profit by the scaling factor $\gamma_{Scale} = \overline{\frac{1}{2}(C_U + C_O)} = \frac{\sum_{i \leq j} a_{ij}\mu_i}{2\sum_{i \leq j} \mu_i}$.

and feature-dependent.

Our analyses and discussion in Sections 2.4 and 2.5 shed light on the theoretical properties of wSAA and kERM, especially regarding their theoretical performance for $N \to \infty$, and our numerical study provides clear evidence that prescriptive approaches are superior to traditional parametric and non-parametric approaches for a finite number of historical observations N. However, our numerical study is inconclusive as to whether one should prefer kERM or wSAA, because the former achieves better performance under medium-range service levels, which we attribute to kERM's superior ability to exploit features' prescriptive content, while the latter achieves superior performance for very high or very low service levels. Additional numerical studies showed that both approaches are fairly robust to an increase in the number of features (see Appendix A.3), which we attribute to the feature selection implicitly done by the random forest weight function or kernel. We did not find evidence that wSAA suffers from interpolation in the presence of a slight positive demand trend or that kERM benefits from its ability to extrapolate. Of course, we cannot rule out that this issue becomes important in other practical instances with stronger positive or negative trend.

2.6.6 Performance Analysis of kERM with Alternative Kernel Functions

In Sections 2.4.2 and 2.5 we discussed data-independent as well as data-dependent kernel functions for the kERM approach and provided performance guarantees for the different classes of kernels. While we were able to show that a universal data-independent kernel (the RBF Gauss kernel) enjoys a universal approximation property, we argued that data-dependent kernels (such as a random forest kernel) may lead to better performance in regimes with a low number of historical observations and a large number of features, because they allow for feature selection. This section explores, how the choice of a particular (data-independent or data-dependent) kernel function impacts the performance of kERM in our specific case example. In line with our discussion in Sections 2.4.2 and 2.5, we compare the gap to optimal profit of kERM when

using (i) data-independent linear and polynomial (homogeneous 3rd degree) kernels, (ii) the data-independent, universal RBF Gauss kernel, and (iii) the data-dependent random forest kernel.[26] Figure 2.6 plots the absolute gap to optimal profit for kERM with the alternative kernel functions across different service levels.

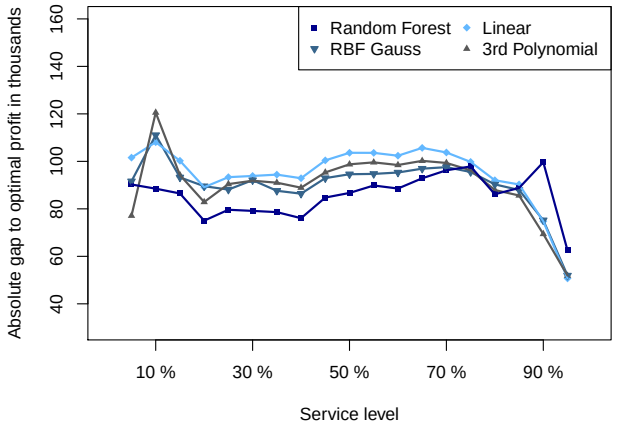


Figure 2.6: Absolute gap to optimal profit of kERM for various kernels across service levels.

We observe that the random forest kernel leads to better performance than all other kernels for service levels between 10 and 70 percent and that the linear and polynomial kernels perform worse than the RBF Gauss kernel. At medium range service levels the value of features is particularly high (see Section 2.6.5) and the random forest kernel appears to benefit from its ability to account for the varying predictive content of individual features and implicit feature selection, as conjectured in Section 2.4.2. To support this conjecture and to shed more light on the performance differences between data-dependent and data-independent kernels, we analyzed the kernel values of the random forest kernel and the various data-independent kernels and found that the random forest kernel leads to very heterogeneous similarity measures—typically, few data samples have a high kernel value, reflecting high similarity, while most others have a low kernel value—and that the data-independent kernels lead

[26] Definitions and further details for these alternative kernel functions are provided in Appendix A.5.

to more homogeneous similarities (see Appendix A.5 for details). This is clearly a direct consequence of how the kernels account for the predictive content of individual features. While the random forest kernel benefits from this property under medium range service levels where forecast accuracy is more important, the more heterogeneous similarities come at a disadvantage for very high service levels because of the regularization effect identified in Section 2.6.5. Due to the heterogeneous similarity measures, kERM needs to regularize more to prescribe the required safety buffers when using a random forest kernel. In contrast, the more homogeneous similarity measures of the data-independent kernels induce a higher implicit uncertainty and therefore also higher implicit safety buffers without the need for strong regularization. In other words, the lower forecast accuracy of these kernels does not imply a disadvantage at very high service levels where the value of features is lower, but rather allows them to prescribe higher implicit safety buffers and overall capacities, which translates into better performance compared to the data-dependent random forest kernel. This finding is in line with our results and discussions in Sections 2.6.4 and 2.6.5.

2.7 Conclusion

This work proposes and studies two data-driven, distribution-free prescriptive analytics approaches for solving a complex two-stage capacity planning problem with multivariate demand and vector-valued capacity decisions. Our main theoretical contribution pertains to the kERM approach, for which we provide solutions for linear and non-linear function spaces, demonstrate the universal approximation property of the approach when using a universal kernel, and derive out-of-sample guarantees for various kernels. The results of our numerical study, using data from a logistics service provider, suggest that substantial performance improvements can be achieved by our prescriptive analytics approaches and that they are more robust to variations of exogenous cost parameters than their traditional counterparts are—which is an attractive property for decision-makers in practice. Our interpretation of the results

sheds light on the two approaches' underlying dynamics and their performance drivers.

Our work has a number of limitations that should be addressed by future research in this relatively new field of prescriptive analytics (in Operations Management). First and foremost we face the problem of generalization. Although we can provide theoretical performance guarantees, especially for our kernelized ERM approach, we cannot make inferences as to which of the prescriptive analytics methods under consideration should be employed in "small data" regimes that are common in OM. In addition to the common machine learning issue of generalization beyond a specific data set that we also face, the application of prescriptive analytics approaches to OM problems raises another issue related to generalization: Because these approaches typically rely on non-standard (asymmetric) loss functions, the results may be sensitive to the choice of the loss function and its (exogenous) parameters, making generalization even more difficult. This issue should be addressed in future work that applies prescriptive analytics in OR/MS.

3 Prescriptive Analytics for a Multi-Shift Staffing Problem

Motivated by the work with a leading maintenance service provider in the aviation industry, this paper examines novel data-driven approaches to solving a certain type of capacity-sizing problem—the *multi-shift staffing problem*—with uncertain, time varying arrival rates and patient "customers" that do not abandon the queue while waiting for a service, but who must be served by a pre-defined time. Drawing on established methods in both capacity management and prescriptive analytics, we propose to use fluid and stationary approximations to apply tailored prescriptive analytics approaches to determine staffing levels for multiple interrelated shifts. The prescriptive analytics approaches rely on machine learning techniques that incorporate a detailed representation of the non-stationary structure of arrivals and leverage extensive auxiliary data that may be predictive of demand. In particular, we adapt established prescriptive analytics approaches—weighted sample average approximation and kernelized empirical risk minimization—and propose a new *optimization prediction* approach to solving the multi-shift staffing problem. Using a case study that is based on extensive data from our project partner, the maintenance service provider, we demonstrate the applicability of these approaches, highlight their benefits over traditional "estimate then optimize" approaches, and shed light on their structural properties and performance drivers. In the context of our real-world application, we derive a clear recommendation for the choice of method with which to solve the multi-shift staffing problem.[27]

[27]This paper is co-authored by Peter K. Wolf and Richard Pibernik.

3.1 Introduction

This paper proposes and examines novel data-driven approaches to solve a certain class of capacity-sizing problems, where a company has to staff multiple shifts for each workday in the presence of uncertain arrival rates that vary throughout the day and patient "customers" that do not abandon the queue while waiting for a service, but who must be served by some pre-defined time (for example, all those who arrive before 5 p.m. must be processed by the end of the day). This problem is common among logistics companies and other service providers, such as those in online fulfillment or bring-in maintenance. We term this problem the multi-shift staffing problem (MSSP).

The research presented in this paper is motivated by our work with Lufthansa Technik Logistik Services, a subsidiary of Lufthansa Technik, a leading provider of maintenance, repair, and overhaul (MRO) services in the aerospace industry. The company faces an MSSP in its inbound logistics operations, where new, refurbished, and defective parts arrive throughout the day and must be processed by the end of the day. How long the individual parts "wait" before being processed is largely irrelevant, as long as they are processed the same day. The company currently operates a morning shift and an afternoon shift of employees who process the incoming parts. If the company cannot process all of the parts that arrive on one day during the day's regular working hours, it has to use costly overtime to process any remaining parts or incur a penalty cost.

The company has access to a time series of historical demand observations in the form of individual daily demand arrivals and associated time stamps, and to a data set that contains numerous co-variates ("features") that correspond to the time series of historical demand observations. These features may be predictive of the arrival rates on a certain day and can include the day of the week, indicator variables for public holidays, and process-related variables like the number of advanced shipping notifications. In other cases, such as in online fulfillment, features may also be derived from weather data, click streams, and other available data sources. The company wants to determine staffing levels for the two shifts every day in such a way that personnel costs

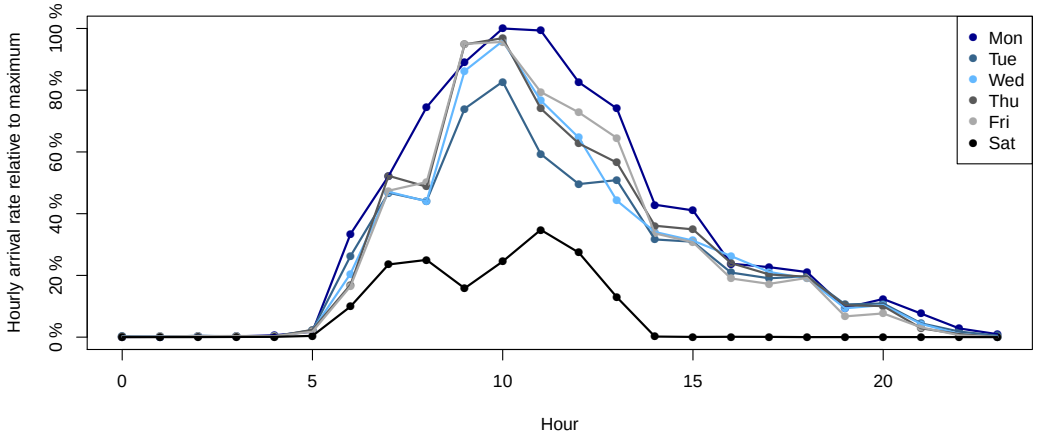


Figure 3.1: Average hourly arrival rate by weekday.

and overtime/penalty costs are minimized.

The arrival rates are time-varying and uncertain and may be correlated throughout the day. For example, Figure 3.1 displays the mean arrival rates for our case company for time periods of one hour on all weekdays based on 89 weeks of data. Although the patterns are similar across weekdays, arrival rates clearly depend on the day of the week. Moreover, the arrival rates during individual time periods vary substantially. Table 3.1 contains estimates of the mean number of arrivals ($\bar{\lambda}$) in the period from 9 a.m. to 10 a.m. on each weekday, along with their empirical coefficient of variation ($CV_{\bar{\lambda}}$) and an estimate of the latter (CV_{Poisson}) that assumes that the underlying process is Poisson. We observe that the empirical CV is much larger than the CV under the Poisson assumption, suggesting considerable parameter uncertainty.

Although the MSSP may seem simpler than many other queuing problems, especially because it does not have to deal with abandonments, it is difficult to solve for two main reasons: i) when formulated as a queuing model with a time-varying, doubly stochastic arrival process, the MSSP is analytically intractable and difficult to solve numerically, even with the assumption that the distribution of the arrival rates is known; ii) in most practically relevant cases, where the availability of historical demand is limited to a certain period of time (e.g., three years), the decision-maker cannot derive an accurate estimate of the arrival rates' feature-dependent and time-varying multivari-

Table 3.1: Mean arrivals between 9 a.m. and 10 a.m. by weekday with estimated processing rate for 10 servers processing all demand in one day.

Day	$\bar{\lambda}$ **in % of maximum**	$CV_{\bar{\lambda}}$ **in %**	$CV_{\textbf{Poisson}}$ **in %**	$1/\sqrt{\mathcal{E}_\lambda}$ **in %**	$\bar{D}$
Monday	98.9	44.2	7.4	19.8	3158.54
Tuesday	78.0	48.0	8.3	22.3	3097.63
Wednesday	88.7	54.6	7.8	20.9	4666.42
Thursday	99.5	48.7	7.3	19.7	4083.87
Friday	100.0	48.5	7.3	19.7	4059.97
Saturday	16.1	114.1	18.2	48.7	3745.33

ate distribution from the available demand and feature data. We overcome both of these difficulties by following a two-step approach: first deriving an approximated MSSP (AMSSP) that can be solved and then using appropriate prescriptive analytics approaches that derive capacity decisions directly from the available demand and feature data without estimating an underlying demand distribution. We outline these two steps in what follows.

Our representation of the AMSSP is based on a fluid approximation, which is a valid approximation when the parameter uncertainty of the doubly stochastic arrival process is the dominant source of uncertainty ("uncertainty-dominated regime"; Bassamboo et al. 2010). In fact, Lufthansa Technik Logistik Services operates under an "uncertainty-dominated regime" because the empirical CV is much larger than $1/\sqrt{\mathcal{E}_\lambda}$, where $\mathcal{E}_\lambda = \lambda/\mu$ is the load of the system and μ its empirical service rate.[28] Bassamboo et al. (2010) show that, under these conditions a (single-shift) capacity optimization problem can be effectively reduced to a newsvendor problem, where "the logic underlying this

[28] One can also employ the dispersion test to estimate the goodness-of-fit of a Poisson distribution for the arrivals within a single time period for a single weekday, as Kim and Whitt (2014) propose. Under the Poisson assumption, the index of dispersion $\bar{D} = (N-1)\bar{\sigma}^2/\bar{\lambda}$ with mean arrival rate $\bar{\lambda}$ and standard deviation of arrival rates $\bar{\sigma}$ for N data samples "is distributed as χ^2_{N-1}, a chi-squared random variable with $N-1$ degrees of freedom" (Kim and Whitt 2014, p. 475). Because the observed dispersion values $\bar{D}$ are well above $\chi^2_{N-1;1-\alpha} \approx 136$ for a 1% significance level and 100 degrees of freedom (Table 3.1), we reject the null hypothesis of observed arrivals' being independent Poisson-distributed variables. In Appendix B.4, we demonstrate that these findings also extend to almost all other time periods.

reduction germinates in viewing stochastic variability as a 'lower order' effect in comparison with parameter (demand) uncertainty" (Bassamboo et al. 2010, p. 1670). However, in our case, matters are more complicated because we have to make interrelated staffing decisions for multiple shifts under time-varying arrival rates (Figure 3.1). Therefore, we introduce a stochastic fluid model that splits the planning horizon of length τ_{max} into multiple periods with length $\Delta\tau = \tau_{max}/T$, for which we can assume stationary arrival rates. While this approach is similar to the stationary independent period-by-period (SIPP) approach, we cannot optimize the capacity levels independently for each period; instead, we want to determine staffing levels for longer shifts that typically span multiple periods, and we have to account for the carry-over between subsequent periods and the fact that arrival rates may not be independent. As a result, we intend to solve a multi-period stochastic optimization problem with a T-dimensional distribution of the arrival rates. Solving such a problem with conventional means is impractical or even infeasible: Even under the highly restrictive assumption of a T-dimensional normal distribution, estimating the parameters of the arrival rate distribution (i.e., the means and the covariance matrix) in particular is difficult because the arrival rates depend on other factors, such as the day of the week (as evidenced by Figure 3.1), whether a particular week includes one or more public holidays, and so on. While these challenges render the approach impractical, the problem also becomes intractable from an analytical and computational viewpoint—even for a small number of periods T and a given T-dimensional distribution (see Section 3.3).

Instead of following the traditional approach of first estimating a (state-dependent) multivariate distribution and then trying to solve a stochastic optimization problem with some approximate (dynamic programming) approach, we propose three data-driven prescriptive analytics approaches that solve the AMSSP by "learning" a prescription rule that derives prescriptions directly from the available demand and feature data without estimating an underlying demand distribution. Two of these approaches adapt methods proposed in Bertsimas and Kallus (2020) (*weighted sample average approximation*, wSAA) and Notz and Pibernik (2021) (*kernelized empirical risk minimization*, kERM). We also propose and study an additional approach—the *Optimization Predic-*

tion (OP) approach—that first derives decisions that would have been optimal in the past and then learns a prescription function from these ex-post optimal decisions. These approaches have two attractive properties: First, they facilitate an efficient computation of a vector-valued capacity prescription while allowing for a detailed representation of the underlying stochastic system with time-varying, uncertain, and potentially correlated arrival rates and a non-continuous cost function. In fact, we can show that our approaches allow for an arbitrarily large T that is restricted only by the time resolution of the data set that contains the historical arrivals. Thus, we can solve the problem over historical observations of the arrival process and overcome what we term the (negative) time-structure effect that is typically associated with fluid and stationary approximations. The second and perhaps more attractive property is that prescriptive analytics approaches make full use of the prescriptive value of a potentially large set of features, so they can overcome a (negative) feature effect that is caused by omitting some or all of the relevant features when estimating an arrival rate distribution, as traditional approaches to solving problems with a similar structure have done.

To determine whether these purported advantages translate into improved prescriptions, we conduct a comprehensive case study based on historical data from Lufthansa Technik Logistik Services and compare the performance of our prescriptive analytics approaches to that of two (traditional) benchmark approaches—*Partitioned Distribution Estimation* (PDE), which relies on estimating demand distributions for the T periods of our planning horizon and solving the AMSSP numerically, and sample average approximation (SAA). We provide structural insights into the time-structure effect and the feature effect and show that prescriptive analytics approaches have clear advantages over traditional approaches in overcoming these effects. Based on the results of additional analyses in which we vary the MSSP's relevant cost parameters, we can also explain differences in performance across the three prescriptive analytics approaches. We find that, among these approaches, wSAA outperforms the others across a wide range of parameter settings, and conclude that, at least in our example, wSAA should be preferred over all other approaches in terms of its performance, intuitiveness, and ease of use.

3.2 Literature Review

The MSSP can be represented by an $M(\tau)/M/b(\tau)$ queuing model with an uncertain arrival rate that is varying in continuous time τ. Such a model is an extension of the $M/M/b+M$ model Bassamboo et al. (2010) study because it assumes a non-stationary τ-dependent distribution of the arrival rate Λ_τ and a capacity $b(\tau)$ that can vary in time. Bassamboo et al. (2010) show that their stationary queuing model cannot be solved analytically because of difficulties in characterizing the distribution of the number of customers in the system under steady state. While the MSSP is simpler than Bassamboo et al.'s (2010) model in one regard—only the queue length at the end of the workday is relevant because customers do not abandon the queue—it is also not analytically tractable because the queue length at the end of the workday depends on the uncertain, non-stationary arrival rates and the multiple shifts' staffing levels (we provide details in Section 3.3). This is one reason to depart from traditional queuing approaches for solving the MSSP and to apply prescriptive analytics approaches instead. However, the more important reason is that, in our setting, we cannot assume we have an accurate estimate of the arrival process—an assumption that is typically made by traditional queuing approaches. We have a data set that contains only historical observations of demand arrivals and potentially predictive features, so obtaining an accurate estimate of a non-stationary arrival rate process with significant parameter uncertainty and feature-dependence is impossible or at least impracticable, especially when there are many features (as in our case) and a comparatively small number of historical observations (e.g., two or three years of daily arrival processes).

As we described in Section 3.1, we follow a two-step approach in solving the MSSP. The first step draws on the literature that addresses capacity planning problems under parameter uncertainty and the literature on capacity planning under non-stationary demand in order to obtain an approximation of the MSSP that can then be solved by means of prescriptive analytics approaches without having to make assumptions about the arrival rate distribution. We review the literature that is relevant to our work and clarify our contribution

in what follows.

Defraeye and Van Nieuwenhuyse (2016) provide a comprehensive overview of queuing approaches for staffing under non-stationary demand—which is frequently the case in real-life settings, as exemplified by our case company (Figure 3.1)—and identify several metrics that are commonly used to evaluate the performance of a capacity configuration that includes fluid approximations, stationary approximations, and simulation. Our derivation of the MSSP first uses a fluid approximation because we experience temporal overload when the queue is building up, and fluid models are "particularly useful to assess performance in systems that are temporarily overloaded" (Defraeye and Van Nieuwenhuyse 2016, p. 15) and because we face significant parameter uncertainty. In such fluid models, in which "the discrete processes of customer arrivals and service completions are replaced by continuous processes" (Stolletz 2008, p. 482), the focus shifts from the stochasticity of the arrival and server processes to the uncertainty in the estimated parameters, that is, the arrival rate. Harrison and Zeevi (2005) study a call center staffing problem with various pools of customers and agents, and use a stochastic fluid model to obtain a newsvendor formulation with linear overage and underage costs. Bassamboo et al. (2010) build on this fluid model approximation and develop an analytical justification: In the "uncertainty-dominated" regime, defined by $CV_\lambda \gg \sqrt{\frac{\mu}{\lambda}}$, where μ is the processing rate and λ is the arrival rate, the variability of the arrival process can be neglected and the fluid approximation is justified, leading to newsvendor-like solutions with an uncertainty hedge that is "extremely accurate" (see Bassamboo et al. 2010, p. 1669). This observation closes the gap between two common ways of formalizing capacity planning problems: queuing models that typically "focus on flow times and responsiveness" (Van Mieghem 2003, p. 280) and newsvendor models that "focus on the impact of multivariate demand uncertainty, while assuming deterministic processing" (Van Mieghem 2003, pp. 281-282). Therefore, our work draws on Bassamboo et al. (2010), demonstrates that the conditions for applying a fluid approximation are fulfilled in the case of Lufthansa Technik Logistik Services—the company operates under an uncertainty-dominated regime (Section 3.1 and Table 3.1)—and derives the AMSSP using a fluid model approach.

However, because the uncertain arrival rate is non-stationary during each day and each shift, we cannot employ a simple newsvendor model. Therefore, we use a stationary approximation, which is "by far the most widely adopted approach for performance evaluation in non-stationary systems" (Defraeye and Van Nieuwenhuyse 2016, p. 12). Similar to the common SIPP approach, we assume the arrival rate is stationary for small periods of time and derive a multi-period stochastic optimization problem termed the AMSSP. As we show in Section 3.3.2, the AMSSP remains analytically intractable even for small numbers of periods T and a known T-dimensional demand distribution, but it can be solved efficiently using various prescriptive analytics approaches or a Monte Carlo sampling approach (as proposed by, for example, Shapiro and Homem-de-Mello 1998) that we will use as a benchmark for the prescriptive analytics approaches.

The stream of research on "prescriptive analytics", which evolved only recently in the operations research and management science (OR/MS) domain, proposes a variety of new solution approaches to stochastic optimization problems, where the objective function is difficult to evaluate because it has a complex functional form, and the decision-maker has access to a potentially large set of feature data that may be predictive of the random variable of interest (e.g., demand). Bertsimas and Kallus (2020) provide a motivation for and a comprehensive overview of prescriptive analytics methods, while Ban and Rudin (2019) propose two prescriptive analytics methods for solving what they term the "Big Data Newsvendor Problem". The idea behind these prescriptive analytics methods is that, instead of solving the decision-making problem by first estimating a distribution of the variable of interest (e.g., demand), and then solving a potentially complex stochastic optimization problem (with a non-linear objective function and/or a very large state space), one considers the empirical counterpart of the problem and uses machine learning techniques to "learn" a prescription rule that derives prescriptions directly from the available demand and feature data without estimating a probability distribution. A variety of such prescriptive analytics approaches have performed well on real-world data sets (Ban and Rudin 2019, Bertsimas and Kallus 2020, Notz and Pibernik 2021).

In contrast to previous work in the domain of prescriptive analytics, our problem setting has a number of complexities that have not been addressed yet and that make it difficult to apply prescriptive analytics approaches that have previously been proposed directly to the MSSP: We face a queuing-type problem in which demand during a day (and, more important, during a shift) may be highly non-stationary (as Figure 3.1 demonstrates), which differs from previous applications of prescriptive analytics, where the time-structure of demand arrivals during a planning period was of no concern, or at least of little concern (e.g., Bertsimas and Kallus 2020, Notz and Pibernik 2021). One approach to such a queuing situation is proposed in Taigel et al. (2019), who study the staffing problem of a public service office and use a decision tree to prescribe optimal staffing levels. This approach is also applied to a call center staffing problem, where Taigel and Meller (2020) use an approximate cost function that solves the queuing problem numerically and train a decision tree to prescribe a staffing decision for a single shift. However, this approach is not applicable to multi-shift problems, so it cannot be applied to the MSSP, where we make multiple interrelated capacity decisions with potentially complex constraints during a single planning period (day). Taigel and Meller's (2020) approach is also restricted to training decision trees and cannot be extended easily to random forests. Bertsimas and Kallus (2020) observed that—at least when one is using wSAA—random forest weight functions can provide significantly better performance than decision tree weight functions can.

We propose three prescriptive analytics approaches to solving the AMSSP. The first approach, wSAA, was initially proposed by Bertsimas and Kallus (2020), who demonstrated, based on a real-world application, that it can lead to superior performance over traditional approaches. Notz and Pibernik (2021) were the first to apply wSAA to a capacity management problem and, inspired by the work of Bertsimas and Kallus (2020), Notz and Pibernik (2021) developed and studied a second prescriptive analytics approach, termed kERM. They found that both prescriptive analytics approaches can have good results when they are used to solve a two-stage capacity planning problem with multivariate demand and upgrading. In contrast to the MSSP, the problem

Notz and Pibernik (2021) studied did not require detailed modelling of the demand-arrival process, and they did not face the problem of staffing multiple shifts for each workday. We adapt both wSAA and kERM to the requirements of the AMSSP and provide a structural comparison and numerical evidence regarding their performance. We also introduce and study a new prescriptive analytics approach, the OP approach, that appears attractive because of its simplicity relative to the other two approaches; it requires only solving a deterministic optimization problem and employment of standard machine learning methods.

To the best of our knowledge, this paper is the first to propose solving a practical queuing type of problem by deriving a formulation (the AMSSP) so we can employ wSAA and kERM, both of which can account for time-varying demand and vector-valued decisions and exploit the prescriptive value of extensive data available to the decision-maker. Our approach may also be useful in solving more complex queuing problems, such as those that include abandonments. While we contribute to the field of prescriptive analytics, especially by providing a new approach and providing insights into the behavior of prescriptive analytics approaches under uncertain and time-varying demand, the main contribution of our paper lies in the combination of both "worlds": that of queuing theory and that of prescriptive analytics.

3.3 The (Approximated) Multi-Shift Staffing Problem

Section 3.1 outlined the two steps of our approach to solving the MSSP. This section describes the first step, the derivation of the AMSSP. We provide a queuing formulation of the MSSP and demonstrate how fluid and stationary approximations can be used to reformulate this problem so it can be solved with traditional numerical techniques (assuming that the demand distribution is known) or by means of prescriptive analytics approaches (if the decision-maker has access only to historical demand and feature data and does not know the demand distribution).

3.3.1 Queuing Formulation of the Multi-Shift Staffing Problem

Let us assume, in the most general terms, that demand follows a non-stationary doubly stochastic Poisson process (similar to Bassamboo et al. 2010), with variability and an uncertain and time-dependent arrival rate Λ_τ with distribution F_{Λ_τ}, and $\tau = 0...\tau_{max}$ representing the operating time of one day. Going forward, assume that the decision-maker knows F_{Λ_τ} (e.g., estimated from the historical data).

Assuming that demand processing occurs at rate μ by $b(\tau)$ identical staff members, we have an $M(\tau)/M/b(\tau)$ queuing model with an uncertain arrival rate. At $\tau = 0$, the queue contains M_0 items of demand, which arrived before the beginning of the work day (e.g., during the night). The decision-maker wants to staff S non-overlapping shifts, where shift s operates from times τ_{s-1} to τ_s with b_s identical staff members:

$$b(\tau) = b_s \text{ for } \tau \in [\tau_{s-1}, \tau_s), \tag{3.1}$$

such that the total expected costs are minimized.

Therefore, the decision-maker intends to solve the optimization problem:

$$\begin{aligned} \min_{\vec{b}=\{b_s\}} C(\vec{b}) &:= \frac{1}{\tau_{max}} \int_0^{\tau_{max}} c_1 b_\tau d\tau + c_2 \mathbb{E}[M_{\tau_{max}}] \\ \text{s.t. } b_\tau &:= b(\tau) = b_s \text{ for } \tau \in [\tau_{s-1}, \tau_s) \; \forall s = 1...S, \end{aligned} \tag{MSSP}$$

where $M_{\tau_{max}}$ is the random variable that describes the number of customers in the queue at $\tau = \tau_{max}$, c_1 is the staffing cost factor, and c_2 is the overtime cost factor.

The objective function of the MSSP is similar to Equation 1 in Bassamboo et al. (2010), but it accounts for time-dependent uncertain arrival rates and does not consider the costs that are associated with abandonments. The constraint ensures a constant staffing level in each shift. In practice, additional constraints and/or parameters may be required, for example, to account for upper bounds on the capacity of an individual shift or to ensure that work-

loads between shifts are balanced, but for the purpose of our exposition, this generic formulation will suffice. Despite its simplicity, the MSSP cannot be solved analytically, even if we assume to know Λ_τ, because of the difficulty in characterizing the distribution of $M_{\tau_{max}}$, which depends on the capacities of all S shifts and the non-stationary uncertain arrival rate Λ_τ.

3.3.2 Approximated Multi-Shift Staffing Problem

Because we cannot solve the MSSP, we propose an approximation based on fluid and stationary approximations that can be solved using prescriptive analytics and traditional numerical approaches. The MSSP can be simplified using the fluid approximation (Bassamboo et al. 2010) if the number of arrivals has an empirical CV that is much larger than $1/\sqrt{\mathcal{E}_\lambda}$, indicating that the "uncertainty-dominated regime" applies so we can neglect the variability (Appendix B.4 shows that this requirement is fulfilled in our real-world application). Based on this fluid model approximation, which assumes a continuous flow with uncertain rate Λ_τ and continuous processing with rate μ, the distribution of the number of demand items in the queue M_τ at time τ is determined by the non-continuous differential equation

$$dM_\tau = \left[(\Lambda_\tau - \mu b_\tau)^+ - \mathbb{1}_{M_\tau > 0} \cdot (\mu b_\tau - \Lambda_\tau)^+\right] d\tau, \tag{3.2}$$

where $(\Lambda_\tau - \mu b_\tau)^+$ is the uncertain queue build-up rate, and $(\mu b_\tau - \Lambda_\tau)^+$ is the uncertain rate at which demand is removed from the queue, given that $M_\tau > 0$. Because the RHS of this differential equation is non-continuous at $M_\tau = 0$ when $\mu b_\tau - \Lambda_\tau > 0$, we cannot obtain an analytical expression for $\mathbb{E}[M_{\tau_{max}}]$ at time τ_{max} for non-stationary Λ_τ in a system in which the queue length may be zero.

Therefore, we further simplify the model using a stationary approximation that partitions the planning horizon into T periods of duration $\Delta\tau = \tau_{max}/T$ and assumes that $\Lambda_\tau = \Lambda_t$ and $b_\tau = b_t$ are constant within each period t. This approximation allows us to derive an expression for $\mathbb{E}[M_{\tau_{max}}] = \mathbb{E}[M_T]$ and transition from uncertain arrival rates Λ_t to uncertain demand $D_t = \Lambda_t \Delta\tau$

within a time period t of length $\Delta\tau$.

The choice of the number of periods T determines the accuracy of this approximation: For $T \to \infty$, the duration $\Delta\tau$, for which Λ_τ and b_τ are assumed to be constant, becomes arbitrarily small, corresponding to the continuous non-stationary setting. We provide a more detailed discussion on the impact of T on accuracy and problem complexity in Section 3.4. Based on (3.2), we can express the uncertain demand in the queue M_t at the end of each discrete period $t = 1...T$ as

$$\begin{aligned} M_t &= M_{t-1} + dM_t/dt \cdot \Delta\tau = (M_{t-1} + \Lambda_t \Delta\tau - \mu b_t \Delta\tau)^+ \\ &= (M_{t-1} + D_t - q_t)^+, \end{aligned} \tag{3.3}$$

where $q_t := \mu b_t \Delta\tau$ is the capacity in period t, expressed as the number of units that can be processed in period t. Using this recursive expression, we can formulate the approximated multi-shift staffing problem (AMSSP) as:

$$\begin{aligned} &\min_{\vec{q} \in \mathcal{Q}} C(\vec{q}) := \frac{1}{T} \sum_{s=1}^{S} c_q (t_{s+1} - t_s) q_s + c_2 \mathbb{E}[M_T] \\ &\text{s.t. } M_t = (M_{t-1} + D_t - q_s)^+ \ \forall t \in [t_s, t_{s+1}) \ \forall s = 1...S, \end{aligned} \tag{AMSSP}$$

where $c_q := c_1/(\mu\Delta\tau)$ is the cost of processing capacity per time period, $\vec{q} = (q_1, ..., q_S) \in \mathcal{Q} \subset \mathbb{R}^S$ is the capacity of all shifts, and t_s is the time period in which shift s starts. By definition, $t_{s+1} > t_s \ \forall s$, $t_1 = 1$, and $t_{S+1} = T + 1$, such that the shifts are non-overlapping and span across all time periods. The AMSSP has the structure of a stochastic inventory-like problem with demand backlogging.

Proposition 3.1. *The cost function $C(\vec{q})$ of the AMSSP is jointly convex in $\vec{q}$.*[29]

The classical approach to solving the AMSSP analytically requires deriving the distribution of M_T and then solving the optimization problem. While the fluid and stationary approximations allow us to derive the expected queue length of the last period $\mathbb{E}[M_T]$, the resulting optimization problem can still

[29] All proofs can be found in Appendix B.1.

not be solved analytically. However, this approximation is still useful because it provides the foundation for prescriptive analytics approaches to solving the MSSP.

Proposition 3.2 provides an expression for $\mathbb{E}[M_T]$ for $T = 2$ with $S = 2$ shifts and independent normally distributed demands D_t.

Proposition 3.2. *Assume $S = T = 2$, an empty queue at the beginning of the horizon ($M_0 = 0$) and independent, normally distributed demands D_1, D_2 with mean μ_1, μ_2 and standard deviation σ_1, σ_2. Then the expected queue length at the end of the horizon is given as:*

$$\mathbb{E}[M_T] = \int_0^\infty m \left[F_{D_1}(q_1) f_{D_2}(m + q_2) + \frac{\exp\left[-\frac{(m-\mu_1-\mu_2+q_1+q_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]}{2\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} \cdot \left(1 + \text{Erf}\left[\frac{\sigma_1^2(m-\mu_2+q_2)+\sigma_2^2(\mu_1-q_1)}{\sqrt{2}\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}\right]\right) \right] dm, \tag{3.4}$$

where $F_{D_1}(d) = \int_{-\infty}^{d} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(\tilde{d}-\mu_1)^2}{2\sigma_1^2}\right] d\tilde{d}$ is the cumulative distribution function of D_1; $f_{D_2}(d) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{(d-\mu_2)^2}{2\sigma_2^2}\right]$ is the density distribution function of D_2; and $\text{Erf}[\cdot]$ is the error function.

The results presented in Proposition 3.2 show that, even under the restrictive assumption of only $T = 2$ periods with $S = 2$ shifts and independent normal demand, the estimation of $\mathbb{E}[M_T]$ becomes complex and results in a non-linear problem that cannot be solved analytically. To solve the AMSSP for $T \geq 2$, one must resort to numerical approaches, as presented in the next section.

3.3.3 Monte Carlo Sampling Solution to the AMSSP

This section uses a numerical solution approach to solve the AMSSP based on Monte Carlo sampling, as Shapiro and Homem-de-Mello (1998) and Shapiro (2003) propose. This numerical approach is used to establish a benchmark for evaluating the prescriptive analytics approaches presented in the next section.

In Section 3.3.1, we assumed that the decision-maker knows the distribution of Λ_τ for all τ. Analogous to the transition from the MSSP to the AMSSP, we can define the T-dimensional uncertain demand $\vec{D}$ as:

$$D_t = \int_{(t-1)\Delta\tau}^{t\Delta\tau} \Lambda_\tau d\tau \ \forall t = 1...T. \tag{3.5}$$

Assuming the decision-maker knows the T-dimensional distribution $F_{\vec{D}}$, the Monte Carlo sampling approach draws N_{MC} samples $\vec{d}^n$ from $F_{\vec{D}}$ and then solves the sample average approximation problem:

$$\begin{aligned} \vec{q}^* := \arg\min_{\vec{q}\in\mathcal{Q}} \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} \left[\frac{1}{T} \sum_{s=1}^{S} c_q(t_{s+1} - t_s) q_s + c_2 m_T^n \right] \\ \text{s.t. } m_t^n = (m_{t-1}^n + d_t^n - q_s)^+ \ \forall n \ \forall t \in [t_s, t_{s+1}) \ \forall s = 1...S \\ m_0^n = M_0 \ \forall n. \end{aligned} \tag{3.6}$$

Assuming that the N_{MC} samples are drawn iid from $F_{\vec{D}}$ and that $\mathcal{Q}$ is compact, then, because $|M_T(\vec{q}, \vec{D})| \leq \sum_t D_t \ \forall \vec{q}$ (dominating function), the uniform strong Law of Large Numbers holds (Lemma 2.4 in Newey and McFadden 1994). Consequently, $\frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} m_T^n$ converges in probability to $\mathbb{E}[M_T]$ for $N_{MC} \to \infty$ and so is a consistent estimator (see, e.g., Shapiro and Homem-de-Mello 1998). Therefore, if the decision-maker knows $F_{\vec{D}}$, the accuracy of this numerical approach is determined only by the number of samples N_{MC} drawn from $F_{\vec{D}}$.

3.4 Prescriptive Analytics Approaches to the AMSSP

The derivation of the AMSSP and the numerical solution approach described in Section 3.3.3 are based on the assumption that the distribution $F_{\vec{D}}$ of the demand arrivals in each period t is known. Although theoretically appealing, this assumption does not hold in most practical settings. When demand is non-stationary during a single planning period (e.g., one day) and depends on numerous "features", the decision-maker is not likely to obtain an accurate

estimate of the distribution $F_{\vec{D}}$.

In this section, we relax the assumption that the distribution $F_{\vec{D}}$ is known and consider the more practical case described in Section 3.1, where the decision-maker has access only to a set of historical demand observations and a data set of corresponding features (explanatory variables) but does not know the relationship between the features and the uncertain arrival rate. More formally, we denote the available data set by $\tilde{S}_N = \{(\vec{\delta}^n, \vec{x}^n)\}$, where $\vec{\delta}^n$ is the vector of arrivals that occurred on an individual day n (e.g., as time stamps),[30] and $\vec{x}^n$ denotes the associated p-dimensional feature vector. The p elements of this feature vector $\vec{x}$ are variables that describe the day of the week, whether the day is close to a public holiday, lagged demands from the previous day or from the same day in the previous week, etc.[31]

For the purpose of solving the AMSSP, which is based on a discrete T-period time structure, we transform the data set $\tilde{S}_N$ as follows: We assign all arrivals to their respective time period t and denote the number of arrivals in period t by d_t. The resulting data set $S_N^T = \{(\vec{d}^n, \vec{x}^n)\}$ contains the number of demand arrivals $\vec{d}^n = \{d_t^n\}$ in each period t on day n, and the same corresponding feature vectors $\vec{x}^n$ as in $\tilde{S}_N$. The transformation from $\tilde{S}_N$ to S_N^T ensures that the level of aggregation of the demand data in S_N^T is aligned with the AMSSP's T-period time structure. This alignment of the data structure with the model structure is a precondition for employing traditional data-driven and prescriptive analytics approaches to solve the AMSSP.

The data set S_N^T is assumed to contain N iid samples from the underlying joint distribution $F_{\vec{X}\times\vec{D}}$ of the T-dimensional uncertain demand $\vec{D}$ and the p-dimensional feature vectors $\vec{x} \in \mathcal{X} \subseteq \mathbb{R}^p$. However, because the joint distribution $F_{\vec{X}\times\vec{D}}$ is $(T+p)$-dimensional, and demand and features may exhibit a complex functional relationship that is driven by interaction effects of multiple features, the limited number of data samples N contained in S_N^T is typically not sufficient to characterize the joint distribution $F_{\vec{X}\times\vec{D}}$. If the decision-maker

[30] The dimensionality of the vector $\vec{\delta}^n$ of the arrivals on day n equals the number of arrivals on this day, and the individual entries of this vector denote the time stamps of the arrivals.

[31] See Appendix B.2 for a detailed description of features used in our real-world numerical analyses.

cannot infer $F_{\vec{X}\times\vec{D}}$ or $F_{\vec{D}|\vec{X}=\vec{x}}$ from S_N^T, the AMSSP cannot be readily solved using the Monte Carlo sampling solution outlined in Section 3.3.3. A simple way to deal with the problem that is caused by $F_{\vec{X}\times\vec{D}}$ being high-dimensional is to reduce the dimensionality of the feature vector $\vec{x}$: Simply speaking, the decision-maker can, for example, identify the most "important" feature $\hat{x}$ contained in $\vec{x}$ (e.g., the day of the week) and neglect the prescriptive value of the remaining features. This simplification allows the decision-maker to estimate conditional distributions $F_{\vec{D}|\hat{X}=\hat{x}}$ by partitioning the data set S_N^T along $\hat{x}$ and fitting a T-dimensional distribution to each partition, and to solve the AMSSP for each $\hat{x}$ using $F_{\vec{D}|\hat{X}=\hat{x}}$ instead of the distribution of $\vec{D}$, as outlined in Section 3.3.3—that is, the PDE approach. We provide additional details on PDE in Appendix B.3.5 and use this approach as a benchmark in our numerical analyses. Clearly, the number of features that can be taken into account when PDE is employed is limited by the number of observations N contained in S_N^T.[32] Therefore, the PDE approach will inevitably forgo some of the prescriptive value of the features contained in the feature vectors $\vec{x}$, which can result in lower performance than that of an approach that considers all relevant features and interaction effects. We term the performance dependence on the number of features the solution approach accounts for as the *feature effect*, which we illustrate in Section 3.5.3.

Moreover, PDE requires estimating a T-dimensional conditional demand distribution, and a limited number of observations N may require the decision-maker to choose a small T to avoid issues that occur when fitting a high-dimensional distribution based on a small number of demand observations.[33] Choosing a small number of periods T can imply that the demand arrivals' non-stationary structure cannot be modeled appropriately, which may lead to lower performance than that of an approach that can incorporate the non-stationary

[32] In practical situations we cannot expect more than 1,000 relevant observations of demand and corresponding feature values. What's more, partitioning, such as by the weekday and some continuous feature like lagged demand using a 10-class binning rule will reduce the number of observations in each partition to an average of less than 17, limiting the accuracy of the estimated conditional distribution.

[33] In addition, the natural assumption of normally distributed total demand arrivals within a time period (as the sum of several random variables) becomes less accurate when T is large and the time periods are short.

structure in more detail. We term the dependence of the performance on T the *time-structure effect*.

As an alternative to this comparatively "hands-on" PDE approach, prescriptive analytics approaches work to exploit the (potential) prescriptive value of all of the p features in the feature vectors $\vec{x}$ without estimating a T-dimensional conditional distribution $F_{\vec{D}|\vec{X}=\vec{x}}$. Instead, they directly prescribe decisions based on the data set S_N^T by deriving a prescription function $\vec{q}(\vec{x})$ that maps from the feature space $\mathcal{X}$ to the decision space $\mathcal{Q}$. Knowing such a prescription function allows the decision-maker to derive a capacity prescription for each new feature vector. As shown in the next section, prescriptive analytics approaches overcome both limitations of the PDE approach by being able to incorporate a large number of features simultaneously, so they are more suitable to exploiting the prescriptive value of the features $\vec{x}$ contained in the data set S_N^T, and by supporting an arbitrarily high time resolution (reflected by very high values of T) that is constrained only by the granularity of the demand data in $\tilde{S}_N$.

Next, we introduce and discuss three prescriptive analytics approaches that prescribe staffing levels directly based on the data set S_N^T.

3.4.1 Weighted Sample Average Approximation

This section adopts the wSAA approach introduced by Bertsimas and Kallus (2020) to solve the AMSSP. In contrast to the PDE approach, wSAA derives prescriptions using an implicit estimate of the conditional demand distribution based on a weight function $w_n(\vec{x})$ that represents the similarity between feature vectors. As such, it fully exploits the prescriptive content of S_N^T. Moreover, because it relies on the empirical demand distribution, and the weight function $w_n(\vec{x})$ is independent of T, the choice of T is restricted only by the time resolution of the data set $\tilde{S}_N$, which allows for a detailed representation of the non-stationary arrival rates. Therefore, it can overcome the time-structure effect that is associated with a small T. The approach is presented in Algorithm 3.1.

In Steps 1 and 2, the decision-maker selects the number of time periods T

Algorithm 3.1 Weighted SAA

1: Choose the number of time periods T.
2: Transform the data set $\tilde{S}_N$ into S_N^T.
3: Select the weight function $w_n(\cdot)$.
4: **for** each new feature vector $\vec{x}$ **do**
5: Compute the weights $w_n(\vec{x})$ for all $n = 1...N$.
6: Solve:

$$\begin{aligned} \vec{q}^{\,\text{wSAA}}(\vec{x}) = \underset{\vec{q}\in\mathcal{Q}}{\arg\min} \sum_{n=1}^{N} w_n(\vec{x}) \left[\frac{1}{T} \sum_{s=1}^{S} c_q(t_{s+1} - t_s) q_s + c_2 m_T^n \right] \\ \text{s.t.} \quad m_t^n = (m_{t-1}^n + d_t^n - q_s)^+ \ \forall n \ \forall t \in [t_s, t_{s+1}) \ \forall s = 1...S \\ m_0^n = M_0 \ \forall n. \end{aligned} \tag{3.7}$$

7: **end for**

and transforms the data set $\tilde{S}_N$ according to our description in the previous section. The feature-based weight function $w_n(\vec{x})$ (selected in Step 3, evaluated in Step 5) computes a measure of similarity between the feature vectors $\vec{x}^n$ in the data set S_N^T and a new feature vector $\vec{x}$. These weights are then used in the subsequent optimization (Problem 3.7 in Step 6), where the total costs associated with demand observations $\vec{d}^n$ are weighted depending on how similar their corresponding feature vector $\vec{x}^n$ is to $\vec{x}$.

Bertsimas and Kallus (2020) point out that, whenever the weights are non-negative, they can be understood to correspond to an estimated conditional distribution of $\vec{D}$, given $\vec{X} = \vec{x}$. Thus, wSAA is based on an implicit estimate of the conditional distribution based on the similarity of the data samples and is in contrast to PDE, which is based on a partitioning of the data set S_N^T into similar observations along individual features and an explicit estimate of a conditional distribution.

The performance of a wSAA approach is determined by the choice of the weight function $w_n(\vec{x})$. Weight functions can be broadly classified into data-independent weight functions, which have a pre-defined functional form, and weight functions that are learned from a specific data set such that the function itself is data-dependent. Bertsimas and Kallus (2020) proved asymptotic optimality of a wSAA approach using data-independent weight functions based on,

for example, k-nearest-neighbors, kernel methods, and local linear regression. Asymptotic optimality can also be shown for the wSAA approach in (3.7) when it is based on data-independent weight functions (Appendix B.7). In regimes with a limited number of historical demand observations and a potentially large number of features p, the property of asymptotic optimality is of less practical relevance[34], and data-dependent weight functions that are based, for example, on regression trees or random forests may produce better performance, especially because they implicitly identify a subset of the features that are most relevant to predicting a variable of interest (e.g., demand). Data-independent weight functions, on the other hand, assign equal importance to all p features (see the discussion in Notz and Pibernik 2021 for details). However, asymptotic optimality could not be shown for data-dependent weight functions.

In numerical experiments by Bertsimas and Kallus (2020) and Notz and Pibernik (2021), the following data-dependent weight function based on random forests (introduced by Bertsimas and Kallus 2020), leads to superior performance:

$$w_n^{\mathrm{RF}}(\vec{x}) = \frac{1}{L}\sum_{l=1}^{L}\frac{\mathbb{1}[\mathcal{R}^l(\vec{x}) = \mathcal{R}^l(\vec{x}^n)]}{\sum_{j=1}^{N}\mathbb{1}[\mathcal{R}^l(\vec{x}) = \mathcal{R}^l(\vec{x}^j)]} \tag{3.8}$$

for a random forest trained on the data set S_N^T, containing L trees, and with $\mathcal{R}^l(\vec{x})$ the terminal node of tree l containing $\vec{x}$. The numerator in (3.8) counts the instances in which the feature vectors $\vec{x}$ and $\vec{x}^n$ are assigned to the same terminal node in tree l, while the denominator captures the number of training samples in the terminal node of $\vec{x}$. The weight $w_n^{\mathrm{RF}}(\vec{x})$ is computed as an average of this fraction for all L trees of the random forest. This definition ensures normalization of the weights, such that $\sum_n w_n^{\mathrm{RF}}(\vec{x}) = 1$. Simply speaking, similarity between $\vec{x}^n$ and $\vec{x}$ is measured by how often these two feature vectors are assigned to the same terminal nodes of a random forest. Because this random forest is trained on the data set S_N^T and uses only the most important features (by selecting a single feature for each split), the weight function is

[34] Convergence may be slow when the number of features p is large; consequently, when N is small, the performance achieved may be far from the optimum. See Notz and Pibernik (2021) for additional details.

data-dependent and based on (implicit) feature selection. The superior performance of wSAA with the weight function $w_n^{\mathrm{RF}}(\vec{x})$ in other applications, a limited number of historical observations of demand $N = 425$, and a large number of features $p = 142$, which warrants a weight function with implicit feature selection, lead us to use $w_n^{\mathrm{RF}}(\vec{x})$ as defined in (3.8) to solve the AMSSP in our real-world application.

3.4.2 Kernelized Empirical Risk Minimization

This section introduces an alternative to wSAA that we term *kERM*. While wSAA relies on re-optimization for each new feature vector $\vec{x}$, the kERM approach determines a prescription function $\vec{q}(\cdot) : \mathcal{X} \to \mathcal{Q}$ that directly maps from the feature space $\mathcal{X}$ to the decision space $\mathcal{Q}$ so that, for any new $\vec{x}$, the function $\vec{q}(\cdot)$ outputs a prescription $\vec{q}(\vec{x})$. The approach selects such a function $\vec{q}(\cdot)$ from a function space $\mathcal{F}$, which is often restricted to the space of linear functions for reasons of tractability and interpretability. However, "there is no reason to expect that optimal solutions will have a linear structure" (Bertsimas and Kallus 2020, p. ec2, e-companion), so we choose to solve the AMSSP for non-linear function spaces through kernelization. Kernelization corresponds to an (implicit) projection of feature vectors into a reproducing kernel Hilbert space, employing the well-known machine learning *kernel trick* (e.g., as used in kernel ridge regression and support vector machines, see Smola and Schölkopf 2004 or Hastie et al. 2009 for details).

The kernel Hilbert space $\mathcal{H}_K$ is defined by a kernel function $K(\vec{x}_1, \vec{x}_2)$.[35] Similar to the weight function $w_n(\vec{x})$ used for wSAA, the kernel function of kERM provides a measure of similarity between two feature vectors $\vec{x}_1$ and $\vec{x}_2$. Algorithm 3.2 summarizes the kERM approach.

Steps 1 and 2 are identical to Steps 1 and 2 of the weighted SAA algorithm, while Step 3 selects the kernel function that defines the reproducing kernel Hilbert space $\mathcal{H}_K$. The problem of selecting a kernel function is similar to the problem of selecting a weight function for wSAA. As is common in

[35] See, for example, Schölkopf and Smola (2002) for the theoretical foundations of using reproducing kernel Hilbert spaces.

Algorithm 3.2 Kernelized ERM

1: Choose the number of time periods T.
2: Transform the data set $\tilde{S}_N$ into S_N^T.
3: Select the kernel function $K(\cdot,\cdot)$.
4: Determine the regularization parameter $\vec{\lambda}$.
5: Compute the coefficients $u_s^n = \frac{1}{2\lambda_s}\left(\sum_{t\in[t_s,t_{s+1})}\alpha_t^n + \delta_s^n - \frac{c_q(t_{s+1}-t_s)}{NT}\right)$ by solving:

$$\begin{aligned}\max_{\{\alpha_t^n\},\{\delta_s^n\}} & -\sum_{s=1}^{S}\lambda_s\sum_{p=1}^{N}\sum_{q=1}^{N}(u_s^p u_s^q K(\vec{x}^p,\vec{x}^q)) + \sum_{n,t}\alpha_t^n d_t^n + \sum_n \alpha_1^n M_0 \\ \text{s.t. } & \alpha_t^n,\ \delta_s^n \geq 0\ \forall t,n,s \\ & \alpha_{t+1}^n \geq \alpha_t^n\ \forall t,n \\ & \alpha_T^n \leq \frac{c_2}{N}\ \forall n.\end{aligned} \tag{3.9}$$

6: Compute the parameter $\vec{b}$ by solving:

$$\begin{aligned}\vec{b} = \arg\min_{\vec{b}} \min_{\{m_t^n\}\in\mathbb{R}^{T\times N}} & \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{T}\sum_{s=1}^{S}c_q(t_{s+1}-t_s)\right. \\ & \left.\cdot\left(\sum_{p=1}^{N}u_s^p K(\vec{x}^p,\vec{x}^n) - b_s\right) + c_2 m_T^n\right] \\ \text{s.t. } & m_t^n \geq m_{t-1}^n + d_t^n - \left(\sum_{p=1}^{N}u_s^p K(\vec{x}^p,\vec{x}^n) - b_s\right) \\ & \forall n\ \forall t\in[t_s,t_{s+1})\ \forall s = 1...S \\ & m_t^n \geq 0\ \forall n,t \\ & m_0^n = M_0\ \forall n \\ & \sum_{p=1}^{N}u_s^p K(\vec{x}^p,\vec{x}^n) - b_s \geq 0\ \forall n,s.\end{aligned} \tag{3.10}$$

7: Compute the prescription $\vec{q}^{\,\mathrm{kERM}}(\vec{x}) = \sum_{n=1}^{N}\vec{u}^{\,n}K(\vec{x}^n,\vec{x}) - \vec{b}$ for each new feature vector $\vec{x}$.

machine learning, we select Tikhonov regularization (Vapnik 1998) to restrict the complexity of the function space $\mathcal{H}_K$. In Step 4, the regularization parameter $\vec{\lambda}$ is obtained through validation and is used in the subsequent steps. In

Step 5 the coefficients u_s^n and in Step 6 the parameters b_s of the prescription function are computed. For this computation, a linearized equivalent of the cost function is required, which we present in Proposition 3.3.[36]

Proposition 3.3. *The cost function of the AMSSP can be linearized as:*

$$\begin{aligned} C(\vec{q},\vec{d}) =& \frac{1}{T}\sum_{s=1}^{S} c_q(t_{s+1}-t_s)q_s \\ &+ \min_{\{m_t\}\in\mathbb{R}^T} c_2 m_T \\ &\quad s.t. \quad m_t \geq m_{t-1} + d_t - q_s \;\forall t \in [t_s, t_{s+1}) \;\forall s = 1...S \\ &\qquad\quad m_t \geq 0 \;\forall t \\ &\qquad\quad m_0 = M_0. \end{aligned} \tag{3.11}$$

Proposition 3.4. *The linearized cost function $C(\vec{q},\vec{d})$, as stated in (3.11), is jointly convex in $\vec{q}$.*

Based on this linearized cost function, we can solve the kERM approach to the AMSSP for a non-linear reproducing kernel Hilbert space $\mathcal{H}_K$. We provide the solution in Proposition 3.5, which forms the basis for Algorithm 3.2's Steps 5 and 6.

Proposition 3.5. *Assume a kernel function $K(\vec{x}_1,\vec{x}_2)$. Then the kERM solution to the AMSSP is given as*

$$\vec{q}^{kERM}(\vec{x}) = \sum_{n=1}^{N} \vec{u}^n K(\vec{x}^n,\vec{x}) - \vec{b}, \tag{3.12}$$

where $u_s^n = \frac{1}{2\lambda_s}\left(\sum_{t\in[t_s,t_{s+1})} \alpha_t^n + \delta_s^n - \frac{c_q(t_{s+1}-t_s)}{NT}\right)$ and α_t^n, δ_s^n being the solution

[36]The linearized cost function ensures that the primal Lagrangian is differentiable.

to

$$\begin{aligned} \max_{\{\alpha_t^n\},\{\delta_s^n\}} & -\sum_{s=1}^{S} \lambda_s \sum_{p=1}^{N}\sum_{q=1}^{N} (u_s^p u_s^q K(\vec{x}^p, \vec{x}^q)) + \sum_{n,t} \alpha_t^n d_t^n + \sum_n \alpha_1^n M_0 \\ s.t.\ & \alpha_t^n,\ \delta_s^n \geq 0\ \forall t, n, s \\ & \alpha_{t+1}^n \geq \alpha_t^n\ \forall t, n \\ & \alpha_T^n \leq \frac{c_2}{N}\ \forall n. \end{aligned} \tag{3.13}$$

Step 7 of Algorithm 3.2 uses (3.12) and the coefficients u_s^n and parameters b_s (computed in Steps 5 and 6) to derive prescriptions for new feature vectors $\vec{x}$. Similar to the possible choices of weight functions for wSAA, kERM allows the decision-maker to select from a variety of kernel functions. We can distinguish between data-independent kernels—among which linear, polynomial, or Gaussian radial basis function kernels are the most popular—and kernels that are learned from the data. Analogous to the property of asymptotic optimality, some of the data-independent kernels may allow for the universal approximation property of the kERM approach. The universal approximation property implies that, in the limit of $N \to \infty$, the kERM approach determines the optimal prescription function of all continuous functions. However, this property is again of limited practicality in settings with a small number of historical observations of demand (e.g., $N = 425$ in our real-world application) and a large number of features ($p = 142$). In such "small data" regimes, a data-dependent kernel may lead to superior performance (see Appendix B.7 for theoretical properties for a number of kernel functions, and Notz and Pibernik 2021 for a more detailed discussion of the choice of the kernel function).

For the same reasons as stated in Section 3.4.1 in favor of a data-dependent weight function in small data regimes, a data-dependent kernel function is likely to lead to better performance in our real-world application. Therefore, we use a random forest kernel, which is similar to the random forest weight function in Algorithm 3.1, for our numerical experiments:

$$K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2) = \frac{1}{L}\sum_{l=1}^{L} \frac{\mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_2)]}{\sum_{j=1}^{N} \mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_j)]}, \tag{3.14}$$

where $\mathcal{R}^l(\vec{x})$ describes the terminal node of tree l for feature vector $\vec{x}$.

We cannot establish the universal approximation property of the kERM approach when using the random forest kernel $K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2)$, but we can derive out-of-sample performance guarantees that bound the expected costs that are associated with staffing levels determined by the prescription function $\vec{q}^{kERM}(\cdot)$ based on a sample-splitting approach (see Appendix B.7).

3.4.3 Optimization Prediction Approach

As an alternative to the two prescriptive approaches presented in Sections 3.4.1 and 3.4.2, which entail re-optimization for each new prescription (wSAA) or complexity in deriving the prescription function (kERM), this section presents an easier approach—the OP approach, which relies only on solving a deterministic optimization problem once and applying a standard machine learning method to predict optimal decisions. More specifically, the OP approach computes ex-post optimal decisions $\vec{q}^*_{S_N^T}(\vec{x}^n)$ for each $\vec{x}^n$ and then, based on these ex-post optimal decisions, learns a prescription function $\vec{q}^{OP}(\cdot)$ that prescribes decisions for new feature vectors $\vec{x}$. The approach is summarized in Algorithm 3.3.

Algorithm 3.3 Optimization Prediction Approach

1: Choose the number of time periods T.
2: Transform the data set $\tilde{S}_N$ into S_N^T.
3: Compute the ex-post optimal decisions for all training data samples $\vec{x}^n$ by solving:

$$\begin{aligned} \vec{q}^*_{S_N^T}(\vec{x}^n) := \arg\min_{\vec{q}\in\mathcal{Q}} & \frac{1}{T}\sum_{s=1}^{S} c_q(t_{s+1}-t_s)q_s + c_2 m_T \\ \text{s.t.} \quad & m_t = (m_{t-1} + d_t^n - q_s)^+ \ \forall t \in [t_s, t_{s+1}) \ \forall s = 1...S \\ & m_0 = M_0. \end{aligned} \tag{3.15}$$

4: Select a machine learning method (e.g., random forest) and learn the prescription function $\vec{q}^{OP}(\cdot)$ that minimizes $\frac{1}{N}\sum_{n=1}^{N}\left|\vec{q}^{OP}(\vec{x}^n) - \vec{q}^*_{S_N^T}(\vec{x}^n)\right|^2$.
5: Compute the prescription $\vec{q}^{OP}(\vec{x})$ for each new feature vector $\vec{x}$.

Steps 1 and 2 are the same as those in Algorithms 3.1 and 3.2, but in Step 3 the OP approach computes the ex-post optimal decisions $\vec{q}^*_{S_N^T}(\vec{x}^n)$ for each $\vec{x}^n$ of the training data set S_N^T. The $\vec{q}^*_{S_N^T}(\vec{x}^n)$ can be regarded as supporting points of the prescription function $\vec{q}^{OP}(\cdot)$, which is learned using a standard machine learning method (Step 4). For learning this prescription function, we use the l_2 norm as loss function so we can employ any standard machine learning technique, including random forests, neural networks, and support vector machines. The prescription function $\vec{q}^{OP}(\cdot)$ is then used to prescribe staffing levels $\vec{q}^{OP}(\vec{x})$ for new observations $\vec{x}$ (Step 5).

The attractiveness of this approach lies in its simplicity, comprehensibility, and low implementation effort. Because the learning problem described in Step 4 is based on a standard loss function, the OP approach does not require the problem-specific development and implementation of machine learning algorithms. In contrast to wSAA and kERM, which require non-standard implementations, the OP approach can be solved directly with standard solvers and common machine learning techniques for regression problems (e.g., neural networks, gradient boosting, or random forests) that are widely available in standard packages. Therefore, the OP approach is most suitable for rapid deployment of prescriptive analytics.

In the next section we explore whether and when this intangible benefit comes at the cost of lower prescriptive performance.

3.5 A Real-World Application and Additional Numerical Insights

This section applies wSAA, kERM, and the OP approach to the MSSP faced by Lufthansa Technik Logistik Services. We use their historical demand data and realistic values of the company's cost parameters to demonstrate the approaches' applicability and to compare their performance to two traditional PDE approaches and a conventional SAA approach that does not incorporate feature data. To elucidate the underlying drivers of the approaches' performance along the time-structure and feature effects (Section 3.4), we vary the

number of time periods T and study the importance of including an extensive set of features. Using additional experiments with varying cost parameters, we are able to shed light on the robustness of the approaches and to explain differences in their performance. We conclude this section with a recommendation on which approach to choose for solving the MSSP, at least in the context of our particular real-world application.

3.5.1 Problem Statement

This section adapts the general MSSP presented in Section 3.3 to the real-life problem of Lufthansa Technik Logistik Services. The case company has two shifts ($S = 2$), the first of which ($s = 1$) runs from 6 a.m. to 2 p.m., and the second ($s = 2$) from 2 p.m. to 8 p.m. We specify the staffing levels of the shifts for the AMSSP as:

$$\begin{aligned} &q_1 \text{ for } t = 7, ..., 14 \\ &q_2 \text{ for } t = 15, ..., 20. \end{aligned} \tag{3.16}$$

We track demand from midnight to 8 p.m. and initially set $T = 20$ with $t = 1, ..., 20$ having a duration of $\Delta\tau = 1$ hour. Based on data provided by the company, including an average personnel cost of 25 EUR/hour and a processing rate of $\mu \approx 7.2$ units/hour, we determined the average personnel costs of processing one unit of demand and set the capacity cost to $c_q = 3.5$ EUR/unit, which is the same for both shifts.[37]

The second shift can be extended to process any remaining workload at 8 p.m. but at a surcharge of 50 percent, so that $c_2 = 5.25$ EUR/unit. Because demand arrivals after 8 p.m. will be backlogged for the next day, we add any of these late arrivals to the demand of the first period of the next day. As indicated in Section 3.3.2, a real-world application of an approach to solve the MSSP typically requires additional capacity constraints. Without such constraints, the solution to the AMSSP is trivial because it would be optimal to set $q_1 = 0$ and to simply solve a newsvendor problem to determine the optimal

[37] Capacity q_t is expressed as the number of units that can be processed in period t: $q_t := \mu b_t \Delta\tau$.

$q_2 > 0$ (see Appendix B.5 for additional details). However, this trivial solution is often not an option in practice because, for example, the storage capacity of unprocessed demand is limited, and congestion needs to be prevented. In our particular case, the company wants to ensure employee satisfaction, so it wants to staff both shifts in a more or less balanced way to accommodate the preferences of its workers. We incorporate this soft workload-balancing constraint by including an intangible cost, denoted by c_3, that penalizes carry-over from the first to the second shift. We acknowledge that there are other ways to account for such load-balancing objectives, such as by enforcing a minimum capacity for the first shift, but we found the carry-over penalty to be the most flexible way to model such a constraint, especially because it allows us to quantify the consequences of more or less load-balancing between the shifts on the overall capacity required, as well as the total costs. For our initial analyses, we set c_3 to a relatively low value of 0.45 EUR/unit and study how higher values of c_3 impact the optimal staffing levels in additional analyses. Based on these specifications, we formulate the AMSSP for our real-world application as:

$$\begin{aligned} &\min_{\vec{q}=(q_1,q_1)} C(\vec{q}) := c_q\left((t_2-t_1)q_1 + (t_3-t_2)q_2\right) + \mathbb{E}[c_2 M_T + c_3 M_{14}] \\ &\text{with } M_t = \begin{cases} 0 \text{ for } t = 0 \\ (M_{t-1} + D_t)^+ \text{ for } 1 \le t \le 6 \\ (M_{t-1} + D_t - q_1)^+ \text{ for } 7 \le t \le 14 \\ (M_{t-1} + D_t - q_2)^+ \text{ for } 15 \le t \le 20, \end{cases} \end{aligned} \tag{3.17}$$

with $t_2 - t_1 = 8$ hours and $t_3 - t_2 = 6$ hours the duration of the two shifts, and M_{14} the queue length at the end of the first shift.

3.5.2 Demand Data, Feature Engineering and Evaluation Procedure

We received from the case company a data set $\tilde{S}_N$ with demand arrivals $\vec{\delta}^n$ on $N = 532$ days between February 2016 and November 2017 with a one-minute

time resolution. From this raw data we constructed a data set S_N^T with demand values $\vec{d}^n$ for $T = 20$ time periods of one hour for each day. For each day, we constructed a feature vector $\vec{x}^n \in \mathbb{R}^{142}$ with features based on the date, public holidays, lagged demands, and features related to the company's processes. A detailed description of the 142 features we constructed for our analysis can be found in Appendix B.2.

We split the data set S_N^T of 532 days into $N = 425$ days of training data, and $N_{test} = 107$ days of evaluation data. We evaluated the performance of the following prescriptive approaches:

1. Weighted SAA: Apply Algorithm 3.1 with random forest weight function (3.8) to the specific capacity planning problem (3.17).

2. Kernelized ERM: Apply Algorithm 3.2 with random forest kernel (3.14) and use (3.12) to solve the specific capacity planning problem (3.17).

3. OP Approach: Apply Algorithm 3.3 using random forest models to the specific capacity planning problem (3.17).

These three prescriptive analytics approaches are compared to three benchmark approaches:

4. SAA: Apply Algorithm 3.1 with $w_n(\vec{x}) = 1/N$ to the specific capacity planning problem (3.17) to obtain the SAA prescription (constant for the test period).

5. PDE-T2 Approach: Estimate a single bivariate normal demand distribution for each weekday and solve (3.17) using Monte Carlo simulation with $N_{MC} = 300$ samples.

6. PDE-T20 Approach: Estimate $T = 20$ independent normal demand distributions for each weekday (resulting in a total of 120 demand distributions) and solve (3.17) using Monte Carlo simulation with $N_{MC} = 300$ samples.

Our first benchmark approach, SAA, is an attractive, data-driven, but featureless approach; the second and third benchmark approaches are traditional

PDE approaches that were presented in Section 3.4. We used PDE first, with $T = 2$ (PDE-T2), which corresponds to the number of shifts to be planned. Then we used PDE with $T = 20$ to incorporate a higher time resolution but assumed independence between the demands of the individual time periods and so fit one normal distribution for each weekday and hour. Appendix B.3 provides a more detailed description of all of the approaches we considered.

We evaluated each approach's prescriptive performance in terms of the gap to optimal cost, that is, in comparison to the ex-post optimal cost

$$C^*(\vec{d}) = \min_{\vec{q} \in \mathcal{Q}} C(\vec{q}, \vec{d}). \tag{3.18}$$

3.5.3 Base Line Results

Figure 3.2 plots the gap to optimal cost of all approaches for the test period using the realistic cost parameters specified in Section 3.5.1. The largest performance gap is between SAA and all other approaches: Compared to SAA, the prescriptive approaches wSAA and kERM reduce the gap to optimal cost by 52.1 percent and 48.3 percent, respectively. kERM performs similar to the PDE-T20 approach, which outperforms its two-period counterpart, PDE-T2, by almost 10 percentage points.

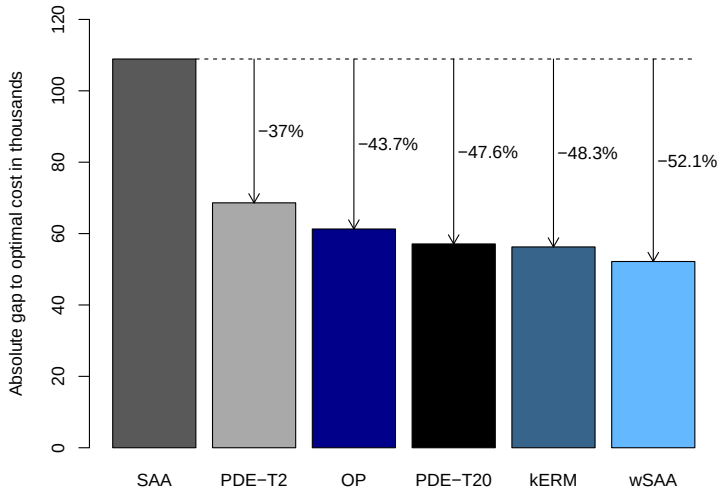


Figure 3.2: Absolute gap to optimal cost for realistic cost parameters.

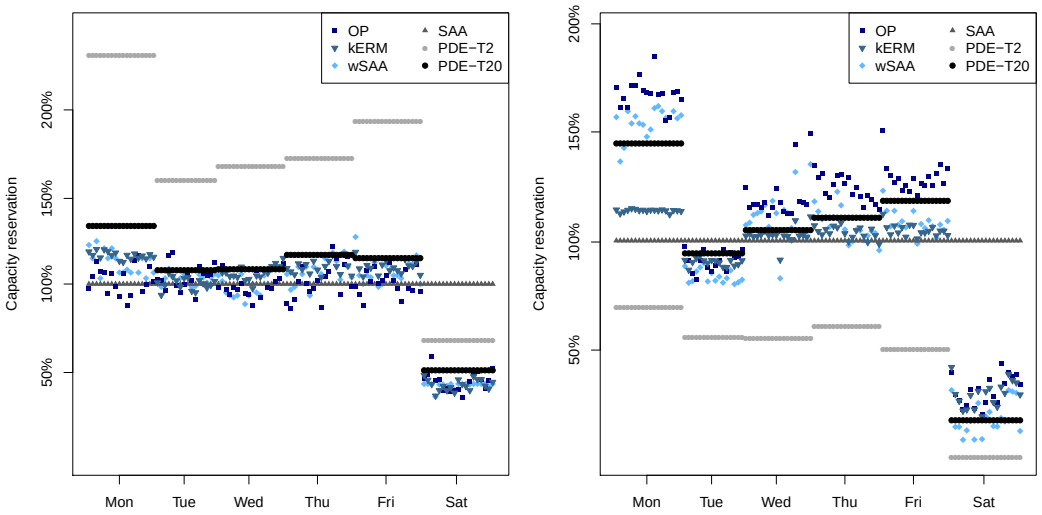
The results for this particular parameter setting, presented in Figure 3.2, point to the time-structure effect and the feature effect introduced in Section 3.4: The difference between the traditional PDE-T2 approach and the PDE-T20 approach in terms of performance reflects the time-structure effect because the two methods are identical, apart from their time resolution of demand. The extent of the feature effect becomes most obvious when SAA and wSAA are compared; both approaches are based on the same time resolution of demand ($T = 20$), but while wSAA accounts for the prescriptive contents of the feature data contained in S_N^T (by re-optimizing against the weights $w_n(\vec{x})$), SAA is featureless and based only on historical demand observations. These clear-cut comparisons provide quantitative evidence for the time-structure and the feature effect, but only for these particular instances with a low vs. high time resolution (of PDE-T2 vs. PDE-T20) and featureless vs. feature-based approaches (SAA vs. wSAA).

The next two sections provide a more nuanced picture of the two effects and how they impact the various approaches' overall performance. These analyses also substantiate our claim that prescriptive analytics approaches have an advantage over traditional approaches in terms of both the time-structure effect and the feature effect.

3.5.4 The Time-structure Effect

The time-structure effect becomes most obvious when PDE-T2 and PDE-T20 are compared. To facilitate an in-depth discussion of the various approaches with regard to the time-structure effect, Figure 3 plots all approaches' staffing levels for shift 1 (Figure 3.3a) and shift 2 (Figure 3.3b), relative to the staffing levels of SAA.

The staffing levels prescribed by PDE-T2 and PDE-T20 provide a straightforward explanation for the more than ten percentage-point differences in performance shown in Figure 3.2. As Figure 3.3 shows, the PDE-T2 approach prescribes higher staffing levels for shift 1 and substantially lower levels for shift 2 than the PDE-T20 approach does consistently across all days of the week. Both approaches face the trade-off between the cost incurred by idle ca-



(a) Shift 1.

(b) Shift 2.

Figure 3.3: Staffing levels relative to SAA for real-world case.

pacity during shift 1, where we have a non-stationary demand structure with increasing rates (Figure 3.1), and the cost incurred from backlogging demand at the end of the first shift. Because the PDE-T2 approach assumes stationary demand during the first shift, it ignores the idling cost during the ramp-up phase at the beginning of shift 1 and so underestimates the overage cost. As a consequence, it prescribes a higher staffing level for the first shift to avoid the backlogging cost at the end of that shift. Because PDE-T20 and all other approaches explicitly account for the increasing demand rates, they prescribe lower staffing levels for shift 1 and higher staffing levels for shift 2 to balance the cost of (idle) capacity with the cost of backlogging at the end of the first shift.

While these results indicate that a low time resolution of demand (e.g., $T = 2$) leads to a (negative) time-structure effect on performance and that it is more beneficial to choose a higher resolution (e.g., $T = 20$), the decision-maker does not know, a priori, the "right" value of T. As discussed in Section 3.4, a T set at either a value that is too low and one that is too high can have adverse effects, depending on the solution approach. To shed light on how the choice of the time resolution impacts the approaches' performance, we carried

out additional analyses in which we varied T from $T = 2$ to $T = 1200$.[38] For $T = 1200$, the length of each period t is equal to one minute.[39] Figure 3.4 reports the gap to optimal cost for OP, kERM, wSAA, and PDE, dependent on T.

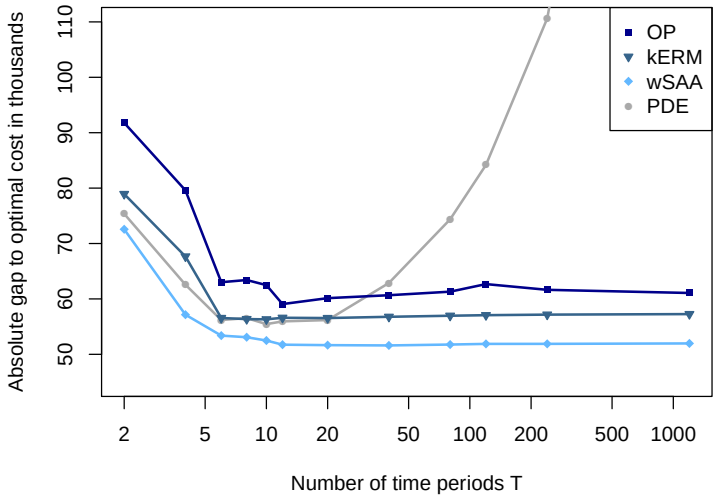


Figure 3.4: Absolute gap to optimal cost for a variation of T.

The results in Figure 3.4 indicate a time-structure effect for $T < 12$. Up to $T = 12$, all approaches benefit from a higher time resolution of demand, while for $T > 12$, there are no additional benefits associated with a more detailed representation of the time structure. However, this finding is specific to our problem setting, and the "right" number of time periods T that is required for optimal staffing of the shifts depends heavily on the demand arrivals' time structure and the number of shifts. While, in our setting, $T = 12$ time periods

[38] In our case, the two shifts have a length of eight and six hours, respectively. To account for the different shift lengths, we define T as the number of time periods between 0 a.m. and 8 p.m., as in Problem 3.17, for $T \geq 20$. When $T < 20$, T refers to the number of time periods in which the two shifts are separated, that is, when each shift is separated into $T/2$ time periods, which guarantees that both shifts extend over an integer number of time periods. We evaluate the prescriptions using the empirical arrivals of the test period with one-minute accuracy.

[39] Of course, the computational costs increase with the number of time periods, although only polynomial in T. In case of, e.g., wSAA, when using the linearized cost function (Proposition 3.3), we need to solve a deterministic linear program with $S + T$ variables, which can be done in polynomial time (Vaidya 1989).

seem sufficient, in other instances (e.g., in online fulfillment) with possibly multiple, short-term demand peaks throughout the day, an even higher time resolution may be warranted, especially when shifts are shorter than they are in our case.

The most relevant and interesting insight from the data shown in Figure 3.4 is that the performances of the three prescriptive analytics approaches (OP, kERM, and wSAA) are robust to an increase in the number of periods T (for $T > 12$), even for extremely large values of T. For example, for $T = 1200$, the prescriptive approaches effectively solve the problem over historical observations of the arrival process. As such, they can overcome the typical problems that are associated with stationary and fluid approximations while still being computationally tractable. The results in Figure 3.4 suggest that, when decision-makers use prescriptive analytics approaches, they may sacrifice performance by choosing a time resolution that is too low but do not suffer from choosing one that is unnecessarily high. We consider this a significant advantage over traditional approaches that rely on distribution fitting, such as PDE. Figure 3.4 shows that the PDE approach's performance diminishes for $T > 20$, which is intuitive because the number of demand arrivals in each time period is small, and a normal distribution becomes a poor approximation of the true demand distribution in each period t. Of course, one could try to use other theoretical distributions, but the well-known problems associated with fitting some theoretical distribution to a small number of historical observations will persist.

3.5.5 The Feature Effect

In conjunction with the performance results in Figure 3.2, the feature effect becomes most obvious when SAA is compared with wSAA. In our analysis, SAA is the only approach that does not account for any features, while wSAA can leverage all of the features contained in our data set S_N^T. Therefore, wSAA's reducing the gap to optimal cost by 52.1 percent can be attributed to the features' prescriptive value and wSAA's ability to leverage this prescriptive

value.[40]

We carried out additional analyses to quantify the importance of various classes of features and their prescriptive value and to clarify how the feature effect is driven by the features' prescriptive value. Figure 3.5 plots the importance of several classes of features (e.g., weekday-related features, public holiday-related features) based on the random forest model that was used for kERM and wSAA.[41] The results show that the features related to weekdays are significantly more important compared to the other classes of features. Figure 3.3 shows the consequences of SAA's not accounting for weekday features as, in contrast to all other approaches, it prescribes constant staffing levels that are independent of the weekday. PDE-T2 and PDE-T20 rely on weekday-specific demand distributions and prescribe staffing levels that vary substantially across weekdays but are constant for each individual weekday. wSAA leads to prescriptions with weekday-dependent levels and variations in individual weekdays. Loosely speaking, we can attribute these within-weekday variations to features that are not contained in the class of weekday-related features. However, we must be careful with this interpretation because other features, such as lagged demands, may be correlated with the weekday features, preventing us from fully separating the effect that weekday-related features and the remaining features have on wSAA's prescriptions.

Beyond these structural effects regarding the prescribed staffing levels, we are interested in how the importance of the class of weekday features and the importance of the remaining features translate into prescription performance. Therefore, we carried out an additional analysis in which we partitioned the data set S_N^T by the weekday and applied SAA to each partition. Comparing the performances of SAA, "weekday-SAA", and wSAA facilitates a consistent evaluation of the weekday-related features' and all other features' prescriptive

[40] Bertsimas and Kallus (2020) introduce the "coefficient of prescriptiveness" as a measure that quantifies "the prescriptive content of data and the efficacy of a policy". In our case, the coefficient of prescriptiveness of wSAA is equal its relative improvement over SAA—that is, .521.

[41] We measure the importance of individual features as the decrease in node impurity based on the residual sum of squares in the random forest that determines the weight or kernel function of wSAA or kERM, respectively. A detailed analysis of the individual features' importance is provided in Appendix B.2.

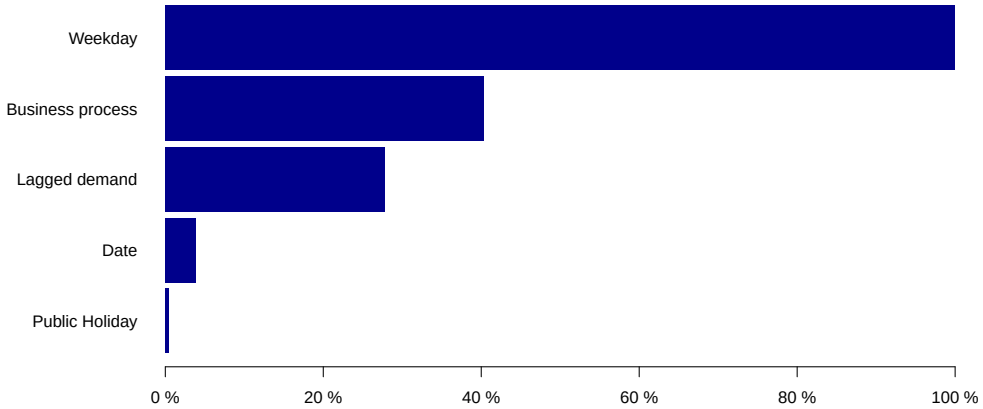


Figure 3.5: Importance of feature classes, measured as a decrease in node impurity in the underlying random forest model.

value. We find that the weekday-SAA approach's performance improvement over SAA is 50.7 percent, which suggests that the remaining features account for performance improvement of only 1.4 percent. Therefore, the feature effect in our particular case example is driven primarily by a single class of features, which is also reflected in the observation that wSAA's prescriptions (Figure 3.3) do not vary substantially on individual weekdays. An additional insight comes from comparing PDE-T20 and the weekday-SAA approach: The relatively good performance of PDE-T20 can be explained by its fitting a distribution for each weekday and each period t and, therefore, capturing most of the features' prescriptive value. However, using PDE-T20 leads to a lower performance than using the weekday-SAA approach does (by 3.1 percent), which can be explained only by the fact that PDE-T20 estimates theoretical demand distributions while the weekday-SAA approach solves the AMSSP over the empirical distribution.

Our results on the feature effect raise the question concerning whether the ability to incorporate a large set of features is a benefit of our prescriptive analytics approaches. However, we must be careful not to generalize the findings of our particular case study. Our situation, in which the weekday proves to be by far the most important feature, appears to be an exception when compared with other case studies presented in, for example, Bertsimas

and Kallus (2020) and Notz and Pibernik (2021). Therefore, we conclude that other features with a high prescriptive value may not be included in our data set S_N^T and that we cannot generally assume that the weekday features have such a predominant effect. In online fulfillment, for example, other features, such as variables for public or school holidays or variables that capture promotional efforts may have much larger prescriptive values than the weekday has in our particular maintenance case. The decision-maker does not know a priori which combination of features drives the feature effect. When using a traditional approach like PDE, the decision-maker has to identify the most important features beforehand and is severely restricted in the number of features that can effectively be incorporated (see our discussion in Section 3.4). In contrast, the prescriptive analytics approaches considered in this paper can deal with a larger set of features without having to make prior assumptions about their prescriptive value. The performances of the kERM and wSAA approaches in Figure 3.2 suggest that while this does not always lead to a large performance increase—as in our particular setting—it also does not come with negative consequences.

3.5.6 Performance Comparison of Prescriptive Analytics Approaches

The time-structure and feature effects described in the previous sections can explain the gaps in the performance levels of the prescriptive approaches (wSAA, kERM, and OP) and the (traditional) benchmark approaches (SAA, PDE-T2, PDE-T20). We showed that the prescriptive analytics approaches allow for a high time resolution of demand and the incorporation of a larger set of features than the traditional approaches allow, and that these capabilities can lead to superior performance. However, while all three prescriptive analytics approaches have these capabilities—they use the same high time resolution of demand ($T = 20$) and can leverage all of the features contained in S_N^T—we saw substantial performance differences in our particular case. Figure 3.2 suggests that using wSAA leads to the highest performance, followed by kERM and OP, the last of which has a gap to optimal profit that is 8.4 percentage

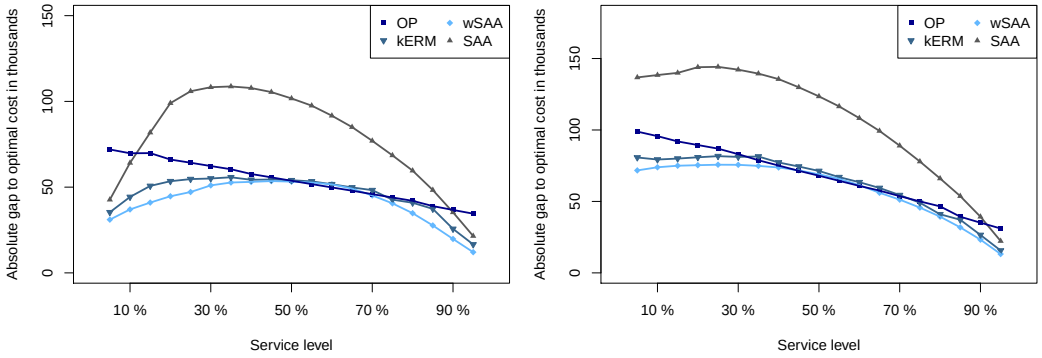
points larger than that of wSAA. In this section, we first explore whether our results obtained for a particular scenario of cost parameters are sensitive to variation in the cost parameters. Then we provide insights into and explanations for performance differences among OP, wSAA, and kERM. We conclude this section with a discussion of non-quantitative advantages and disadvantages of three prescriptive analytics approaches and a recommendation for practitioners who intend to use prescriptive analytics approaches to solve the MSSP.

To explore the prescriptive approaches' sensitivity to variations in the cost parameters, we vary the staffing costs c_q, the overtime costs c_2, and the backlogging costs c_3 to induce approximate newsvendor-type service levels ($SL = \frac{C_U}{C_U+C_O} = 1 - \frac{c_q}{c_2}$) that range from 5 percent to 95 percent and different load-balancing factors, defined as $LB = 1 - \frac{c_q}{c_3+c_q}$. The load-balancing factor LB describes the propensity to fulfill demand that arrives in the first shift within that shift, rather than backlogging the demand for the second shift. Table 3.2 depicts the cost parameters of our analyses and Figure 3.6 shows the performance of the three prescriptive analytics approaches that were trained and evaluated for each cost scenario based on the real-world data set S_N^T.

Table 3.2: Variations in Cost Parameters.

Figure	c_q	c_2	c_3	**Service Level SL**	**Load Balancing LB**
3.6a	5...0.3	5.25	0.6...0.06	5%...95%	0.1
3.6b	5...0.3	5.25	7.5...0.4	5%...95%	0.6

From the results presented in Figure 3.6, we observe that using wSAA leads to the best performance across almost all service level regimes and that using kERM leads to similar performance but that is slightly lower in most instances. For service levels ranging from 40 percent to 70 percent, using the OP approach achieves optimality gaps that are similar to those of wSAA and kERM, but also leads to substantially lower performance for very high or very low service levels. This finding is in line with the results reported in Figure 3.2 for our realistic cost scenario, where the approximated service level



(a) Load balancing $LB = 0.1$.

(b) Load balancing $LB = 0.6$.

Figure 3.6: Absolute gap to optimal cost across service levels (SL) for different levels of load-balancing (LB).

is low ($SL = 33.3\%$) and using OP leads to a large gap to optimality relative to that of using wSAA and kERM. Structurally, the results are similar under low and high load-balancing factors. Appendix B.6 provides more extensive explanations for the effect of the load-balancing factor and shows how that factor impacts the traditional approaches' performance.

The lower performance of the OP approach under very low or very high service level regimes can be explained based on traditional newsvendor logic. Simply speaking, a high service level ($SL > 50\%$) requires that the optimal staffing level includes a positive safety buffer that increases in both the service level and the demand uncertainty. Likewise, a low service level ($SL < 50\%$) requires a negative safety buffer. While OP accounts for a potential asymmetry in overage and underage costs (in Step 3 of Algorithm 3.3), demand uncertainty is not involved in this optimization because it determines the ex-post optimal staffing levels for (certain) historical observations of demand, so the ex-post optimal staffing levels obtained do not include any (positive or negative) safety buffers to hedge against demand uncertainty. In Step 4 of Algorithm 3.3, the OP approach generalizes these ex-post optimal decisions by learning a prescription function based on the common, symmetric l_2 loss function, which also does not provide for safety buffers because the symmetric l_2 loss implicitly induces an optimal service level of 50 percent. As a conse-

quence, the OP approach does not adjust the staffing levels appropriately for high or low service levels when demand is uncertain. This *safety buffer effect* can also be observed in Figure 3.3b, where OP prescribes staffing levels that are higher than those of either wSAA or kERM on each weekday. Intuitively, this effect is negligible—or at least less pronounced—under medium-range service level regimes, but it has a detrimental effect on performance when service levels are either very high or very low and demand uncertainty is considerable, as in our case example (Table 3.1).

We now turn to the differences in the performance that result from using wSAA and kERM. The results in Figure 3.6 suggest that using kERM or wSAA leads to lower performance when the service levels are high or low. However, the differences are much lower than they are for the OP approach. We attribute this lower performance to a *regularization effect* that Notz and Pibernik (2021) describe and that can also be explained using a traditional newsvendor logic. Under high or low service level regimes, it is optimal for kERM to choose a higher regularization parameter in Step 4 of Algorithm 3.2, which causes the prescription function to adapt less well to variations in the historical demand observations. This reduction in variation is generally desirable under high or low service level regimes, because it leads to a higher bias and therefore to a higher positive or negative (implicit) safety buffer (see Notz and Pibernik 2021 for additional details). However, in the case of our data set it also suppresses the variations in the prescriptions between the weekdays. Figure 3.3b shows both effects: a reduced variation in prescriptions for kERM (e.g., on Mondays) and a reduced variation between the weekdays; for example, kERM prescribes lower capacities than wSAA on Mondays and Wednesdays but higher capacities on Tuesdays and Saturdays.

Our numerical results lead us to conclude that, at least for our particular case example, wSAA is preferable for solving the (A)MSSP because it allows for a high time resolution, incorporates a large number of features, and is subject to neither the safety buffer effect nor the regularization effect. The superior performance of wSAA in our study supports the results of previous studies, where wSAA leads to very good, if not superior, results, when it is applied to other problems in Operations Management (see Bertsimas and Kallus 2020

and Notz and Pibernik 2021). wSAA also has some clear intangible benefits, especially over kERM, as it builds on the well-established SAA methodology in combination with a weight function that is highly intuitive. As such, not only does it have an intuitive logic and is fairly easy to explain to decision-makers, but it also requires a limited implementation effort. Although wSAA requires re-computation of the weights and re-optimization against these weights for every new feature vector $\vec{x}$, its computational complexity does not seem to pose restrictions for practical applications.

kERM is more complex and difficult to implement than wSAA is, as evidenced by Algorithm 3.2. In our setting, kERM's greater complexity is not compensated for by better performance because, in contrast to wSAA, kERM suffers from a regularization effect that appears to be particularly relevant because of the weekday structure in our data set.

As Section 3.4.3 explains, the OP approach's attractiveness lies in its simplicity and in its ability to be deployed rapidly and with little effort since it does not require the problem-specific development and implementation of machine learning algorithms but can be solved with standard solvers and standard machine learning packages. However, the approach suffers from a safety buffer effect, so it performs well only when target service levels are close to 50 percent, that is, service levels for which the safety buffer effect is not pronounced. In our case, using the OP approach lead to good performance for service levels between 40 percent and 70 percent, but because the safety buffer effect also depends on the demand uncertainty, we cannot generalize this finding to other problem instances. This issue limits the OP approach's general applicability to solving the (A)MSSP, despite its advantages in terms of comprehensibility and ease of implementation.

3.6 Conclusion

This paper proposes new data-driven approaches for solving a multi-shift staffing problem (MSSP) with uncertain, time-varying arrival rates and patient "customers" that do not abandon the queue while waiting for service. Our ap-

proaches have two steps: Derive an approximated MSSP (AMSSP) based on fluid and stationary approximations, and solve the AMSSP by means of tailored prescriptive analytics approaches, using machine learning techniques that allow for a detailed representation of the non-stationary structure of arrivals and that leverage extensive auxiliary data that may be predictive of demand. We adapted two established prescriptive analytics approaches—weighted sample average approximation and kernelized empirical risk minimization—and proposed a new *optimization prediction* approach to solve the (A)MSSP. We show that our approaches have clear structural benefits over traditional approaches that first estimate the arrival rates' distribution and then try to solve a complex stochastic optimization problem: they can help overcome both a negative time-structure effect and a negative feature effect that are likely to be associated with traditional approaches. Using a real-world application and additional numerical analyses, we exemplified and quantified these benefits. The numerical analyses also provided insights into differences between the performance levels of the three prescriptive analytics approaches we employ. Based on an extensive discussion of the numerical results, we conclude that prescriptive analytics approaches are superior to traditional approaches and that, among the prescriptive analytics approaches, wSAA is the most suitable method in terms of both performance and intangible criteria like ease of use, intuitiveness, and comprehensibility. Of course, we must be careful in generalizing these results beyond the boundaries of our particular application; additional comparative studies should be carried out to validate (or invalidate) our findings. Nonetheless, we see prescriptive analytics approaches as a promising avenue for solving complex capacity planning problems with a queuing type of structure and see the primary contribution of our paper as its combining both "worlds"—that of queuing theory and that of prescriptive analytics. Our approach may be the basis for solving capacity planning problems that are richer and more complex than the MSSP by means of prescriptive analytics approaches.

4 Explainable Subgradient Tree Boosting for Prescriptive Analytics in Operations Management

Motivated by the overwhelming success of gradient boosting approaches in machine learning and driven by the need for explainable prescriptive analytics approaches in operations management (OM), we propose an explainable prescriptive analytics approach to solving complex OM problems: subgradient tree boosting (STB). The STB approach combines the well-known method of subgradient descent in function space with sample average approximation, and directly prescribes decisions from a problem-specific loss function, historical demand observations, and prescriptive features. The approach is inherently explainable and allows a decision-maker to derive detailed explanations for the prescribed decisions, such as a breakdown of individual features' impact that clarifies the underlying drivers of the prescriptions, which is increasingly useful in practice. We show how subgradients can be derived for many common OM problems, demonstrate STB's applicability to two real-world, complex capacity planning problems in the service industry, benchmark its performance against those of two prescriptive approaches—weighted sample average approximation (wSAA) and kernelized empirical risk minimization (kERM)—and show how STB's prescriptions can be explained by estimating the impact of individual features. The results suggest that STB's performance is comparable to those of wSAA and kERM but also provides explainable prescriptions.

4.1 Introduction

Prescriptive analytics approaches to operations management (OM) problems use integrated machine learning algorithms to prescribe decisions based directly on historical observations of demand and numerous features (co-variates), without making assumptions about underlying demand distributions. These approaches have become popular (Bertsimas and Kallus 2020, Notz and Pibernik 2021) and aim to improve prescriptive performance by prescribing decisions that minimize expected costs or maximize expected profits. However, "understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications" (Lundberg and Lee 2017, p. 4768). According to Rudin (2019), especially in domains like healthcare and criminal justice, the explainability of prescription models can be of particular relevance to preventing incorrect decisions. Rudin (2019, p. 206) calls for the use of "inherently interpretable models" in making high-stakes decisions. However, this interpretability may come at the cost of performance, and a decision-maker may face a trade-off between prescription performance and explainability (Bertsimas et al. 2019).

Boosting approaches are particularly well-suited when both prescription performance *and* explainability are required: Boosting has often led to superior performance and is often termed "one of the most significant advances in machine learning" (as in Zhang and Yu 2005, p. 1538). Boosting is regularly among the winning approaches in machine learning competitions, such as those organized by *Kaggle* and others (see, e.g., Sandulescu and Chiru 2016 or Volkovs et al. 2017). At the same time, boosting models are generally explainable because they consist of an additive set of comparatively simple base learners (e.g., simple decision trees), which can be interpreted. This additive structure of a boosting model allows the decision-maker to derive explanations for the prescribed decisions that build trust and enable the traceability of decisions.

The foundations of boosting—that is, the improvement of a weak learning algorithm through iterative application and combination—date back to Schapire (1990), who shows that *weak learnability* (which means that a learn-

ing algorithm can perform at least slightly better than random guessing) is equivalent to the notion of *strong learnability*; in other words, given enough training data, "the learner with high probability is able to output an hypothesis that is correct on all but an arbitrarily small fraction of the instances" (Schapire 1990, p. 197). Freund and Schapire (1997) introduce the first boosting algorithm, termed *AdaBoost*, that focuses on learning an improved classifier based on weak classifiers. This algorithm opened up the field of research on boosting methods and led to the development of a multitude of algorithms for classification and regression problems (see, e.g., Schapire 2003, Bühlmann and Hothorn 2007, and Schapire and Freund 2012). One of the boosting algorithms that has been applied successfully is *Gradient Boosting*, introduced by Friedman (2001) and extended to stochastic gradient boosting in Friedman (2002). Stochastic gradient boosting led to one of the most popular gradient boosting approaches in use today, *XGBoost*, a scalable tree boosting approach proposed by Chen and Guestrin (2016). The concept of gradient boosting is similar to that of gradient descent methods of convex optimization: After initializing with a base hypothesis function, the algorithm performs "steps" in the opposite direction of the gradient of a loss function toward the minimum by adding base learners to the hypothesis function that approximate the negative gradient.

When using decision trees as base learners, as XGBoost does, one can derive *Shapley additive explanation* (SHAP) values that allow us to express any prediction as the sum of individual features' contributions. These SHAP values, proposed by Lundberg and Lee (2017), are shown to be the only possible feature attribution method that provides *local accuracy* (the sum of SHAP values equals the prediction), *consistency* (between models with differing feature attributions), and *missingness* (features that have no impact on the prediction must have zero value). Because the computation of these values can lead to a complexity that is exponential with the number of features, Lundberg et al. (2020) introduce an algorithm they call *TreeSHAP*, which allows us to compute the SHAP values efficiently in polynomial time.

Motivated by the success of boosting approaches in the domain of (traditional) machine learning and the inherent explainability of their predictions, we

apply the boosting idea of combining a set of iteratively trained, weak, explainable learners to the domain of prescriptive analytics. While classic gradient boosting approaches are based primarily on differentiable loss functions (e.g., the l_2 loss), we require a generalization of these approaches to more complex, non-differentiable loss functions that are common in OM, such as the well-known newsvendor problem, which leads to a non-differentiable loss function. Similar to the subgradient methods that generalize gradient descent algorithms to be applicable to non-differentiable cost functions, gradient boosting can be restated using subgradients, as in Ratliff et al. (2006) and Biau and Cadre (2017). However, the application of these subgradient boosting approaches for prescriptive analytics in OM raises two problems: First, such a generalized approach (e.g., Algorithm 2 in Biau and Cadre 2017) is applicable only to loss functions that are defined on a decision space $\mathcal{Q} = \mathbb{R}$. Such is not usually the case in OM settings, where the loss function (or cost function) is often defined only for a restricted decision space, such as $\mathcal{Q} = \{q \in \mathbb{R} : 0 \leq q \leq q_{max}\} \subset \mathbb{R}$. In such settings, when the loss function is defined only on $\mathcal{Q}$, as in the case of the real-world capacity problems presented in Section 4.5, the subgradients are undefined for $q \notin \mathcal{Q}$, and the traditional (sub)gradient boosting approach (as in Friedman 2001 and Biau and Cadre 2017) is not applicable. Second, deriving a subgradient of OM loss functions is not trivial. Many OM problems lead to complex loss functions, such as the seemingly simple problem of capacity planning with upgrading for a car rental provider, which leads to a (two-stage) stochastic problem with recourse (Netessine et al. 2002), for which a subgradient of the respective loss function cannot be obtained easily. Another example of this class of problems is a two-stage shipment planning problem, where products are allocated to individual warehouses under demand uncertainty (the first stage), and once demand is realized, the decision-maker chooses from which warehouse to ship products to satisfy it (second stage) (Bertsimas and Kallus 2020). Deriving a subgradient for the loss function of this complex, two-stage OM problem is, again, not trivial.

We address the first problem by proposing an approach we term *Subgradient Tree Boosting* (STB), which extends the (sub)gradient boosting approach by using sample average approximation (SAA) to estimate the descent di-

rection (in function space) and the step size, which guarantees that the prescriptions for the training data samples are always feasible and the respective subgradients are well-defined throughout the training procedure (Section 4.3). We address the second problem by providing methods with which to derive a subgradient of the loss functions for common OM problems; in particular, we use perturbation theory to derive a subgradient for two-stage stochastic problems with recourse (Section 4.4).

Our numerical analyses compare STB's performance to those of two prescriptive analytics approaches, weighted SAA (wSAA) and kernelized empirical risk minimization (kERM) (Bertsimas and Kallus 2020, Notz and Pibernik 2021), by means of two real-life case studies of capacity planning problems. We derive SHAP values using the TreeSHAP algorithm (Lundberg et al. 2020), provide exemplary breakdowns of prescriptions into individual features' impacts (based on their feature value), and use SHAP dependence plots to show how the individual features' value drives the prescription. These explanations based on SHAP values mark a significant advancement from analyses of features' importance conducted for wSAA or kERM using a random forest weight or kernel function (see, e.g., Notz and Pibernik 2021). Such analyses of importance shed light only on which feature is most important to characterize the similarity of training samples (e.g., the weekday could be more important than lagged demand). In contrast, the SHAP values derived for STB allow us to explain how the value of a feature impacts the decision; for example, because demand yesterday was 50 percent higher than average, we decrease our capacity prescription by 7 percent (Figure 4.9b, aviation-maintenance case).

The main contributions of this paper can be summarized as follows:

1. We propose STB, a novel prescriptive analytics approach to OM problems based on a combination of well-known (sub)gradient boosting methods and SAA for estimating step size and direction, and compare its prescription function structurally to that of kERM.

2. We show how a subgradient can be derived for many common OM problems, including the use of perturbation theory to derive a subgradient for the large class of two-stage stochastic problems with recourse.

3. We demonstrate the STB approach's applicability to real-life OM problems by applying it to two capacity planning problems (of a mail logistics provider and an aviation maintenance service provider) and benchmark its performance against those of SAA, wSAA and kERM.

4. We show that STB's prescriptions are explainable, compute the SHAP values, break selected prescriptions down into the contributions of individual features, and analyze how individual features' value drives the prescription.

4.2 Literature Review

The research presented in this paper is related to three streams of literature: the vast body of literature on boosting approaches for machine learning, the recently evolved stream of research on prescriptive analytics approaches for OM, and the stream of literature on explainable artificial intelligence (explainable AI).

The first stream, on boosting approaches for machine learning, begins with the idea of boosting introduced in Schapire (1990), in which weak and strong learnability are shown to be equivalent. Based on this insight, Freund and Schapire (1997) propose *AdaBoost*, the first boosting algorithm for classification. While research first considered boosting an iterative learning approach in which each additional weak learner improves the overall hypothesis function by successively correcting for prediction errors, Mason et al. (1999, p. 512) provide the perspective of "boosting algorithms as gradient descent on cost-functionals in an inner-product function space". This understanding facilitated Friedman's (2001) development of an explicit formulation of gradient boosting and its extension to stochastic gradient boosting in Friedman (2002). For additional details, we refer the reader to comprehensive reviews, including theoretical perspectives and statistical insights on several boosting approaches in Meir and Rätsch (2003), Bickel et al. (2006), Bühlmann and Hothorn (2007), and Schapire and Freund (2012).

In addition to the classic prediction tasks of regression and classification,

boosting has been applied to a variety of problems, including non-differentiable loss functions. Ratliff et al. (2006) use boosting approaches to solve the maximum margin planning algorithm for imitation learning; in particular, to solve their learning problem, they use "subgradient descent in the space of cost functions" (Ratliff et al. 2006, p. 1153), described as a "convex, but non-differentiable regularized risk function for the general margin" (Ratliff et al. 2006, p. 1154). The proposed methods are applied in Ratliff et al. (2007) and Ratliff et al. (2009).

An even more general view on boosting for convex loss functions, presented in Grubb and Bagnell (2011) and Biau and Cadre (2017), has since evolved. Grubb and Bagnell (2011) provide a general theory of functional gradient descent over the $L^2(\mathcal{X}, \mathcal{Y}, \mu)$ Hilbert space of functions $f : \mathcal{X} \to \mathcal{Y}$ with μ the probability measure. They propose two modified boosting algorithms for which they derive convergence results for non-differentiable convex risk functionals. However, Biau and Cadre (2017, p. 1) show that these modifications are not required and provide a "thorough analysis of two widespread versions of gradient boosting", including convergence results for these algorithms, where the key difference between the two versions is the method by which the functional (sub)gradient is projected onto the space of admissible functions. The first algorithm they analyze is based on the scalar-product-maximizing boosting steps originally proposed in Mason et al. (2000). Convergence is demonstrated under the assumption of a bounded loss function with a Lipschitz-continuous subdifferential in expectation (Theorem 3.1 in Biau and Cadre 2017). The second algorithm uses an l_2 norm minimization to project the functional (sub)gradient onto the space of admissible functions (originally proposed in Friedman 2001). In addition to the assumptions for the first algorithm, strong convexity of the loss function is required to prove convergence (Theorem 3.2 in Biau and Cadre 2017). Convergence is generally a valuable property, but when boosting approaches are applied to real-world problems and finite data sets, these properties are "numerical-analysis-type results, which do not provide information on the statistical properties of the boosting predictor" (Biau and Cadre 2017, p. 18). In fact, the boosting approaches may overfit when the iteration is continued until it converges, which

leads to predictor functions that do not generalize well beyond the training data set. Regularization methods that prevent overfitting include the popular technique of *early stopping* and limiting the complexity of the linear span of the base learners, which is the function space over which the boosting approaches optimize when the number of iterations is not limited. Using a specific function space with limited complexity and a strongly convex risk functional, Biau and Cadre (2017) show that, in such a setting, the boosting algorithm does not overfit when the parameters are carefully selected.[42]

Applying these (sub)gradient boosting algorithms becomes difficult, when the feasible region $\mathcal{Q}$ of the decision q is restricted (e.g., when $0 \leq q \leq q_{max}$), which is typically the case in OM problems. In such settings, the prescription function may lead to infeasible decisions q during training, which can render the subgradient undefined and, consequently, the algorithm inapplicable. Therefore, we propose a combination of the (stochastic) (sub)gradient boosting algorithm with l_2 norm minimization projection (Friedman 2001, Friedman 2002, Biau and Cadre 2017) with SAA to determine step size and direction, which guarantees feasibility of the prescribed decisions q for the training data and all boosting iterations during training (see Section 4.3.1 for details).

The second stream of related literature focuses on prescriptive analytics approaches to OM problems. Historically, planning problems in OM have been solved either under the assumption of a known demand distribution (e.g., Netessine et al. 2002) or using a data set of historical demand observations, with the implicit assumption of a stationary but unknown demand distribution (e.g., by employing the SAA approach, see Shapiro and Kleywegt 2002 and Shapiro 2003). When the demand distribution is neither known nor stationary, but the decision-maker has a set of historical demand observations and feature data (co-variates), prescriptive analytics approaches may lead to superior decisions. Key contributions to this stream of literature are those of Ban and Rudin (2019), who study a newsvendor model with feature data; Bertsimas and Kallus (2020), who propose a general prescriptive analytics

[42]The proposed function space is based on decision trees that are allowed to split only in the center of a node—which is in contrast to regular decision trees, where the splitting point is optimized—and partitions the feature space in equally sized hypercubes.

framework based on a weighted adaption of SAA (wSAA); Notz and Pibernik (2021), who compare a kERM approach to wSAA; and Notz et al. (2020), who apply wSAA and kERM to a queuing capacity planning problem and propose an additional approach they term Optimization Prediction (OP) approach.[43] A more detailed review of this stream of literature can be found in Notz and Pibernik (2021).

We extend this stream of literature by proposing STB as a novel prescriptive analytics approach (in addition to wSAA and kERM) and demonstrate its applicability using two complex capacity planning problems with vector-valued decisions, multivariate demand, and multiple constraints. The first problem is a capacity planning problem of a mail logistics provider that plans the (staff) capacity for three service lines under demand uncertainty with an upgrading option after demand has been realized (Bassok et al. 1999, Netessine et al. 2002, Notz and Pibernik 2021). The second problem is the multi-shift staffing problem (MSSP) of a maintenance service provider in the aviation industry that plans daily staff capacities for two shifts while facing uncertain hourly demand arrivals (Notz et al. 2020).

A third stream of literature, focused on explainable AI, has evolved recently. This research area follows the view that, in many applications, the "explainability" of a prediction can be as important as the prediction performance (Lundberg and Lee 2017). Typical motivations for wanting to understand a model include the need to trust the model's predictions, to understand the causality behind its decisions, and to ensure fair and ethical decision-making (Lipton 2018). Foundational concepts of explainability (Gilpin et al. 2018) and a perspective on what constitutes a good explanation based on research in the social sciences (Miller 2019) elucidate what makes a model understandable or explainable. Abdul et al. (2018) provide a general overview of research that is relevant to explainable AI across many domains. However, a model's interpretability may come at the cost of decreased prediction performance. Bertsimas et al. (2019, p. 1) "quantify the 'price' of interpretability, i.e., the tradeoff with predictive accuracy" because requiring a model to be in-

[43] The OP approach leads to significantly lower performance for exceedingly low or high service levels (Notz et al. 2020), so it is not included in our analysis.

terpretable in addition to predicting accurately may lead to a dual-objective problem. We may also face this trade-off in our research because, for example, when STB leads to lower prescription performance than wSAA, a decision-maker would have to choose between higher performance (wSAA) and more explainable prescriptions (STB). We explore this trade-off numerically, derive explanations of the STB prescriptions using SHAP values (Lundberg and Lee 2017, Lundberg et al. 2020), and present a detailed discussion in Section 4.5.5.

4.3 Subgradient Tree Boosting for Prescriptive Analytics

This section introduces the STB approach under the assumption that a subgradient of the loss function is known. The section also provides a structural comparison of its prescription function to kERM and shows how SHAP values can be derived (following Lundberg et al. 2020) to explain the STB prescriptions. Section 4.4 provides methods with which to derive a subgradient for the most common OM problems.

Assume a decision-maker has a historical data set S_N of N data samples $(\vec{x}^n, \boldsymbol{d}^n)$ with historical demand observations $\boldsymbol{d}^n \in \mathcal{D}$ and feature vectors (co-variates) $\vec{x}^n \in \mathcal{X}$. The decision-maker faces the planning problem of choosing a $\vec{q} \in \mathcal{Q}$ (e.g., a capacity decision) to minimize a loss function $L(\vec{q}, \boldsymbol{d}) : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$, which assigns an incurred loss to a decision $\vec{q}$ given a realized demand $\boldsymbol{d}$. For many OM problems in which the decision-maker wants to minimize some (expected) costs associated with a decision, the loss is equal to the cost reflected by the problem's objective function (e.g., as in the case of the aviation maintenance provider in Section 4.4.4). In other instances, where the decision-maker wants to maximize the (expected) profit that is associated with a decision, the loss can be expressed as the gap to the optimal profit under full information (e.g., as in the case of the mail logistics provider in Section 4.4.3).

One approach of prescriptive analytics is to determine a function $\vec{q}(\cdot)$ of a function space $\mathcal{F}$ that maps from $\mathcal{X}$ to $\mathcal{Q}$, such that the true risk (which is

defined as expected loss) is minimized:

$$\min_{\vec{q}(\cdot)\in\mathcal{F}} R(\vec{q}(\cdot)) := \min_{\vec{q}(\cdot)\in\mathcal{F}} \mathbb{E}_{\vec{X}\times\boldsymbol{D}}\left[L(\vec{q}(\vec{X}),\boldsymbol{D})\right]. \tag{4.1}$$

Because the joint distribution of $\vec{X}\times\boldsymbol{D}$ is unknown, but the decision-maker has a data set S_N that consists of N iid samples of this joint distribution, the principle of empirical risk minimization proposes to minimize the empirical risk instead of the true risk:

$$\min_{\vec{q}(\cdot)\in\mathcal{F}} R_N(\vec{q}(\cdot)) := \min_{\vec{q}(\cdot)\in\mathcal{F}} \frac{1}{N}\sum_{n=1}^{N} L(\vec{q}(\vec{x}^n),\boldsymbol{d}^n). \tag{4.2}$$

Gradient boosting approaches solve (4.2) by applying a gradient descent method in function space, which iteratively adapts a hypothesis function $\vec{q}_k(\cdot)$ along the direction of the gradient of the loss function $L(\vec{q},\boldsymbol{d})$ with respect to $\vec{q}$. Because the loss function of OM problems is often convex but not differentiable, we follow the generalized approach of subgradient boosting that employs subgradients to determine the direction of descent. Next, we introduce the concept of subgradients and subdifferentials for convex loss functions (following Rockafellar 1970).

Definition 4.1. *(Following Rockafellar 1970) A vector $\vec{s}\in\mathcal{Q}$ is called a subgradient of a convex function $L:\mathcal{Q}\to\mathbb{R}$ at $\vec{q}_0\in\mathcal{Q}$ if*

$$L(\vec{q}) \geq L(\vec{q}_0) + \langle \vec{s}, \vec{q}-\vec{q}_0\rangle \ \forall \vec{q}\in\mathcal{Q}. \tag{4.3}$$

Definition 4.2. *(Following Rockafellar 1970) The subdifferential $\partial L(\vec{q}_0)$ of a convex function $L:\mathcal{Q}\to\mathbb{R}$ at $\vec{q}_0\in\mathcal{Q}$ is the set of all subgradients:*

$$\partial L(\vec{q}_0) = \{\vec{s} : L(\vec{q}) \geq L(\vec{q}_0) + \langle \vec{s}, \vec{q}-\vec{q}_0\rangle \ \forall \vec{q}\in\mathcal{Q}\}. \tag{4.4}$$

Based on these definitions, we can state requirements that guarantee at least one subgradient.

Proposition 4.1. *(Following Rockafellar 1970) Let $L:\mathcal{Q}\to\mathbb{R}$ be a convex function with $\mathcal{Q}$ convex and non-empty, $L(\vec{q})<\infty$ for at least one $\vec{q}\in\mathcal{Q}$, and*

$L(\vec{q}) > -\infty$ for all $\vec{q} \in \mathcal{Q}$, and let $\vec{q}_0 \in \operatorname{int} \mathcal{Q}$. Then the subdifferential $\partial L(\vec{q}_0)$ of L at $\vec{q}_0$ is a non-empty, bounded set.[44]

Proposition 4.1 implies that, when $\mathcal{Q} = \mathbb{R}^I$, the subdifferential is non-empty and at least one subgradient exists for each $\vec{q} \in \mathcal{Q}$. The subgradient provides the foundation for the STB approach, which we describe in the next section.

4.3.1 The Subgradient Tree Boosting Approach

In this section we propose the STB approach, which is a combination of (stochastic) (sub)gradient boosting (Friedman 2001, Friedman 2002, Biau and Cadre 2017) and SAA, for solving Problem 4.2. Similar to other boosting approaches (e.g., XGBoost in Chen and Guestrin 2016), STB uses decision (regression) trees as base learners.

As discussed in Section 4.1, in OM settings the space of feasible decisions $\mathcal{Q}$ is almost always restricted by, for example, only allowing for positive capacities $q \geq 0$ and/or imposing an upper bound as $q \leq q_{max}$. In such settings, traditional (sub)gradient boosting approaches may lead to hypothesis functions that prescribe decisions $q \notin \mathcal{Q}$ during training for which the loss function is not defined, so a subgradient cannot be derived, making the approach non-applicable. We propose to overcome this problem by using SAA to estimate the step size and direction in each boosting iteration. This additional SAA step, which builds on the idea of line search in classic gradient descent algorithms, is an attractive choice because it guarantees that the prescriptions remain feasible ($q \in \mathcal{Q}$) throughout all training iterations while also optimizing the descent step size (similar to line search in gradient descent).

In what follows, we first define the space of all possible prescription functions, and then describe the STB approach in detail (Algorithm 4.1). Let K

[44] All proofs can be found in Appendix C.1.

be the number of iterations[45] and let each regression tree be defined as

$$\vec{q}_{Tree}(\vec{x}) = \sum_{l=1}^{L} \vec{\lambda}_l \mathbb{1}_{\vec{x} \in R_l}, \tag{4.5}$$

where L is the number of terminal (leaf) nodes R_l of the tree, and $\vec{\lambda}_l$ is the prescription corresponding to terminal node l. Then we can define the function space that contains all possible prescription functions of STB as

$$\mathcal{F}_{STB} = \left\{ \vec{q}(\cdot) : \mathcal{X} \to \mathcal{Q} \;\middle|\; \vec{q}(\vec{x}) = \vec{q}_0 + \sum_{k=1}^{K} \sum_{l=1}^{L} \vec{\lambda}_{lk} \mathbb{1}_{\vec{x} \in R_{lk}} \right\}, \tag{4.6}$$

where K equals the number of trees. Because solving Problem 4.2 over the function space $\mathcal{F}_{STB}$ directly is not feasible in many cases, we follow the stepwise *greedy* approach of iteratively learning tree after tree and updating the hypothesis function $\vec{q}_k(\cdot)$, as in Friedman (2001). This procedure is similar to the *steepest descent method* and the *subgradient method*, which are methods for solving convex optimization problems by iteratively adapting the solution hypothesis along the negative (sub)gradient direction (see Chapter 1.2 in Bertsekas 1995 and Chapter 8.9 in Bazaraa et al. 1993 for details).

In Step 1 the STB approach initializes the sequence of hypothesis functions $\vec{q}_k(\cdot)$ by using an SAA prescription (the optimal constant prescription for the training data set) as $\vec{q}_0(\cdot)$. Then, for each iteration $k = 0...K-1$, a random subset $\pi_k(n)$ of the training data is drawn (similar to stochastic gradient boosting, see Friedman 2002) and an arbitrary subgradient $\vec{s}_k(\vec{x}^{(\pi_k(n))})$ of the loss function $L(\vec{q}, \boldsymbol{d})$ with respect to $\vec{q}$ is calculated for each data point $(\vec{x}^{(\pi_k(n))}, \boldsymbol{d}^{(\pi_k(n))})$ of the random subset (Steps 3 and 4). The size of the random subset is determined by the fraction $0 < \xi \leq 1$. In Step 5 we train a regression tree that minimizes the l_2 loss so it best approximates the set of subgradients $\vec{s}_k(\vec{x}^{(\pi_k(n))})$. This step of the algorithm can be considered a projection of the point-wise defined functional subgradient onto the function space of regression trees, which are our base learners: $\mathcal{F}_{Tree} = \{\vec{q}(\cdot) | \vec{q}(\vec{x}) = \sum_{l=1}^{L} \vec{\gamma}_l \mathbb{1}_{\vec{x} \in R_l}\}$.

[45] See Appendix C.4 for details on how the number of iterations K impacts STB's prescriptive performance.

Algorithm 4.1 Subgradient Tree Boosting

1: $\vec{q}_0(\vec{x}) = \arg\min_{\vec{q}} \sum_{n=1}^{N} L(\vec{q}, \boldsymbol{d}^n)$ ▷ Initialize hypothesis function.
2: **for** $k = 0...K-1$ **do**
3: $\pi_k(n) = \text{rand}(\{1...N\})$ for $n = 1...\lfloor \xi N \rfloor$ ▷ Draw random subset.
4: $\vec{s}_k(\vec{x}^{(\pi_k(n))}) \in \partial L(\vec{q}, \boldsymbol{d}^{(\pi_k(n))})|_{\vec{q}=\vec{q}_k(\vec{x}^{(\pi_k(n))})}$ ▷ Determine subgradient.
5: $\{R_{lk}\} = \arg\min_{\{R_l\},\{\vec{\gamma}_l\}} \sum_{n=1}^{\lfloor \xi N \rfloor} \left|\left| \vec{s}_k(\vec{x}^{(\pi_k(n))}) - \sum_{l=1}^{L} \vec{\gamma}_l \mathbb{1}_{\vec{x}^{(\pi_k(n))} \in R_l} \right|\right|^2$ ▷ Learn decision tree.
6: **for** $l = 1...L$ **do**
7: $\vec{\lambda}_{lk} = \arg\min_{\vec{\lambda}} \sum_{n=1}^{N} L(\vec{q}_k(\vec{x}) + \vec{\lambda} \mathbb{1}_{\vec{x}^n \in R_{lk}}, \boldsymbol{d}^n)$ ▷ Determine step size and direction.
8: **end for**
9: $\vec{q}_{k+1}(\vec{x}) = \vec{q}_k(\vec{x}) + \nu \sum_{l=1}^{L} \vec{\lambda}_{lk} \mathbb{1}_{\vec{x} \in R_{lk}}$ ▷ Update hypothesis function.
10: **end for**

In Step 7, we use the partitioning of the feature space $\mathcal{X}$ into the terminal nodes R_{lk} of the regression tree and solve the SAA problem of determining the optimal $\vec{\lambda}_{lk}$ for each terminal node. Because this SAA step considers all training data samples, in contrast to the random subset considered when learning the regression tree, it guarantees that $\vec{q}_k(\vec{x}^n) + \sum_{l=1}^{L} \vec{\lambda}_{lk} \mathbb{1}_{\vec{x} \in R_{lk}} \in \mathcal{Q}\ \forall k, n$. This feasibility guarantee of the prescriptions for the training data ensures that the loss function is defined, which is a requirement for estimating a subgradient. Finally, a new element of the sequence of hypothesis functions $\vec{q}_{k+1}(\cdot)$ is calculated by adding the regression tree function with $\vec{\lambda}_{lk}$ to the preceding element of the sequence $\vec{q}_k(\cdot)$ (Step 9). The parameter ν is a shrinkage factor with $0 < \nu \leq 1$, which determines the learning rate. After completion of K iterations, the STB prescription function $\vec{q}^{STB}(\cdot) = \vec{q}_K(\cdot)$ is determined and can be used to prescribe capacities for new feature vectors $\vec{x}$.

4.3.2 Structural Comparison of STB and kERM

This section addresses the structural similarity between the prescription functions determined by STB and kERM when using a random forest kernel. The aim of both kERM and STB is to determine a prescription function $\vec{q}(\cdot)$ that solves Problem 4.2 over a function space $\mathcal{F}$. STB uses an iterative greedy ap-

proach to determine the prescription function (Algorithm 4.1), while kERM solves Problem 4.2 directly for a kernel-based function space. One of the simplest function spaces for which Problem 4.2 can be solved directly is the space of linear functions, leading to a linear ERM solution. However, because the relationship between the feature vector $\vec{x}$ and the optimal decision $\vec{q}^*$ is often non-linear (Bertsimas and Kallus 2020), a decision-maker may achieve better results using non-linear function spaces. One way to incorporate non-linearity efficiently is kernelization based on a kernel function $K(\vec{x}_1, \vec{x}_2)$, which can be interpreted as a measure of similarity between the feature vectors $\vec{x}_1$ and $\vec{x}_2$. Using kernelization, the kERM approach solves Problem 4.2 over the non-linear reproducing kernel Hilbert space $\mathcal{H}_K$, providing a prescription function as:

$$\vec{q}^{\,\mathrm{kERM}}(\vec{x}) = \sum_{n=1}^{N} \vec{u}^n K\left(\vec{x}^n, \vec{x}\right) - \vec{b}. \tag{4.7}$$

This prescription function's dependency on $\vec{x}$ is defined by the kernel function $K(\vec{x}_1, \vec{x}_2)$ and, for example, for a linear kernel $K(\vec{x}_1, \vec{x}_2) = \langle \vec{x}_1, \vec{x}_2 \rangle$, the structure of $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ differs markedly from the elements of the function space $\mathcal{F}_{STB}$ that STB uses. However, the prescription functions $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ and $\vec{q}^{\,\mathrm{STB}}(\cdot)$ become structurally similar when the decision-maker chooses the random forest kernel proposed in Notz and Pibernik (2021) as:

$$K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2) = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathbb{1}_{\mathcal{R}^k(\vec{x_1})=\mathcal{R}^k(\vec{x}_2)}}{\sum_{j=1}^{N} \mathbb{1}_{\mathcal{R}^k(\vec{x_1})=\mathcal{R}^k(\vec{x}_j)}}, \tag{4.8}$$

consisting of K trees. In particular, we can state the kERM prescription function as:

$$\begin{aligned}
\vec{q}^{\,\mathrm{kERM}}(\vec{x}) &= \sum_{n=1}^{N} \vec{u}^n \left(\frac{1}{K} \sum_{k=1}^{K} \frac{\mathbb{1}_{\mathcal{R}^k(\vec{x})=\mathcal{R}^k(\vec{x}^n)}}{\sum_{j=1}^{N} \mathbb{1}_{\mathcal{R}^k(\vec{x})=\mathcal{R}^k(\vec{x}_j)}} \right) - \vec{b} \\
&= \sum_{k=1}^{K} \sum_{l=1}^{L_k} \mathbb{1}_{\vec{x} \in \mathcal{R}_l^k} \left(\frac{1}{K} \sum_{n=1}^{N} \vec{u}^n \frac{\mathbb{1}_{\vec{x}^n \in \mathcal{R}_l^k}}{\sum_{j=1}^{N} \mathbb{1}_{\vec{x}^j \in \mathcal{R}_l^k}} \right) - \vec{b} \\
&=: \sum_{k=1}^{K} \sum_{l=1}^{L_k} \mathbb{1}_{\vec{x} \in \mathcal{R}_l^k} \vec{\alpha}_l^k - \vec{b},
\end{aligned} \tag{4.9}$$

for an ensemble of K trees with L_k leaf nodes $\mathcal{R}_l^k$ with value $\vec{\alpha}_l^k$ in each node l of tree k. This expression is structurally similar to the prescription function determined by STB

$$\vec{q}^{\,\mathrm{STB}}(\vec{x}) = \vec{q}_0 + \sum_{k=1}^{K} \sum_{l=1}^{L} \nu \vec{\lambda}_{lk} \mathbb{1}_{\vec{x} \in R_{lk}}, \tag{4.10}$$

as the constant offset $-\vec{b}$ in (4.9) corresponds to the STB initialization $\vec{q}_0$ in (4.10), and the kERM leaf node values $\vec{\alpha}_l^k$ in (4.9) correspond to the STB leaf node values $\nu\vec{\lambda}_{lk}$ in (4.10). Despite the similarity in structure, the prescriptive approaches' training methodologies differ significantly: all decision trees of the random forest that defines the kernel function $K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2)$ are trained simultaneously using historical observations of demand, so the tree structure is independent of the problem-specific loss function. In contrast, STB's decision trees are trained sequentially, each depending on the subgradient of the loss function with respect to the hypothesis function of the previous iteration. In addition to these differences in the partitioning R_{lk} of the trees, kERM and STB also differ in terms of the number of free parameters that are optimized during training. In the case of kERM, the training procedure optimizes N parameters ($\vec{u}^n$), which equals the number of training data samples, while STB's iterative learning algorithm determines $K \times L$ parameters ($\vec{\lambda}_{lk}$), where K is the number of iterations and L is the number of each decision tree's leaf nodes.

4.3.3 Explaining STB Prescriptions using SHAP Values

Explainable prescriptions that allow decision-makers to understand decisions are increasingly required when prescriptive analytics approaches are applied in practice. Most prescriptive analytics approaches provide only a decision $\vec{q}(\vec{x})$ for each new feature vector $\vec{x}$ and no further justification, which may limit human decision-makers' ability to trust the results. Explanations that are provided in addition to a prescribed decision can overcome this limitation and support building trust in the prescriptions.

The well-known prescriptive analytics approaches, kERM and wSAA, are

based on potentially complex, non-linear kernel and weight functions. In addition, wSAA does not provide an explicit prescription function but relies on re-optimization for each new feature vector. Consequently, there is no general method with which to derive explanations for the prescriptions of kERM and wSAA directly, and the best approach to generating insights into the underlying mechanics is to analyze the importance of individual features in determining the similarity measure provided by the kernel and weight functions. However, such feature importances (e.g., as provided in Notz and Pibernik 2021) can only be assumed to apply similarly to the actual prescriptions and do not provide insights beyond which features generally play an important role in determining prescriptions.

In contrast, because its prescription function has an additive structure of simple decision trees that are inherently explainable, STB allows the decision-maker to derive detailed explanations for each individual prescription. In particular, with STB, a breakdown of individual prescriptions into individual features' impacts (e.g., increases or decreases of a capacity prescription by a certain amount) and SHAP dependence analyses that show how the value of an individual feature drives the prescription value can be derived. These explanations are derived after the iterative training of the STB prescription function has been completed.

Assuming feature vectors that consist of p real-valued features—that is, $\vec{x} \in \mathcal{X} \subseteq \mathbb{R}^p$—then an individual prescription $\vec{q}(\vec{x})$ is considered to be explainable if it can be written as the sum of the effects $\vec{\phi}_m(\vec{q}(\cdot), \vec{x})$ of the individual features $m = 1...p$ (following Lundberg et al. 2020):

$$\vec{q}(\vec{x}) = \vec{\phi}_0(\vec{q}(\cdot)) + \sum_{m=1}^{p} \vec{\phi}_m(\vec{q}(\cdot), \vec{x}), \tag{4.11}$$

where $\vec{\phi}_0(\vec{q}(\cdot))$ is a constant offset. Lundberg et al. (2020) term this the property *local accuracy* and additivity. In addition to local accuracy, Lundberg et al. (2020) propose that the feature effects $\vec{\phi}_m(\vec{q}(\cdot), \vec{x})$ should fulfill the properties of *consistency* between models and *missingness*, which ensures that the assigned effects of features that have no impact on a model vanish. As Lund-

berg and Lee (2017) show, these properties can be guaranteed only by defining

$$\vec{\phi}_m(\vec{q}(\cdot), \vec{x}) := \sum_{U \in \mathcal{U}} \frac{1}{p!} \left[\vec{q}_{\vec{x}}(V_m^U \cup m) - \vec{q}_{\vec{x}}(V_m^U) \right], \tag{4.12}$$

where $\mathcal{U}$ is the set of all orderings of the p features, V_m^U is the set of all features before feature m in ordering U, and $\vec{q}_{\vec{x}}(V_m^U)$ is the expected model prescription for $\vec{x}$ when only the features V_m^U are present. These effects $\vec{\phi}_m(\vec{q}(\cdot), \vec{x})$, which can be understood as individual impacts of features on the prescription value, are rooted in the game theoretic Shapley values, so they are termed *SHapley Additive exPlanation* (SHAP) values (Lundberg and Lee 2017).

For a single decision tree, a simple way to calculate $\vec{q}_{\vec{x}}(V_m^U)$ is to follow the decision path in the tree for all features $x_i \in V_m^U$ toward the respective leaf node and to average the prescription value over the child nodes when a decision feature is not an element of V_m^U. This approach, which is outlined in Algorithm 1 in Lundberg et al. (2020), can be applied to all K trees of the STB prescription function because of the SHAP values' additivity. However, this approach to calculating the SHAP values has a complexity of $O(KLp2^p)$ (Lundberg et al. 2020), making the approach impractical for use with a large number of features p. Lundberg et al. (2020) resolve this problem by proposing the *TreeSHAP* algorithm, which allows us to calculate the SHAP values in polynomial time (see Algorithm 2 in Lundberg et al. 2020). In Section 4.5.5 we demonstrate how STB prescriptions can be explained by deriving individual prescriptions' SHAP values in both the mail logistics provider and the aviation maintenance provider case studies. We also analyze numerically how an individual feature's value impacts the prescription.

4.4 Estimating Subgradients of OM Loss Functions

In Section 4.3.1 we proposed the STB approach for use under the assumption that a subgradient for the OM problem is known. However, it is often not trivial to obtain a subgradient when the OM loss function is complex, such as

in the case of stochastic problems with recourse. In this section we address this problem and derive subgradients for the most common OM problems, including the well-known newsvendor and complex (two-stage) stochastic problems with recourse. When STB is applied, it is sufficient to determine an arbitrary subgradient $\vec{s}_L(\vec{q}, \boldsymbol{d}) \in \partial L(\vec{q}, \boldsymbol{d})$ of the loss function $L(\vec{q}, \boldsymbol{d})$ at a specific point $\vec{q}$, which is commonly termed *weak* subgradient calculus. Like numerical differentiation, numerical approaches to estimating a subgradient have been proposed (e.g., Studniarski 1989), but the required calculations may be computationally expensive.[46] Therefore, we focus on deriving the required subgradients analytically.

4.4.1 Subgradients for Common OM Loss Functions

Many common convex OM loss functions are not differentiable at certain points; for example, the loss function of the well-known newsvendor model $L_{NV}(q, d) = c_o(q - d)^+ + c_u(d - q)^+$ with decision q, demand d, and overage and underage cost factors $c_o, c_u > 0$ is not differentiable at $q = d$. This property is rooted in the $(\cdot)^+$ operator, which, like the absolute value $|\cdot|$, is not differentiable when the argument vanishes. For such loss functions, Proposition 4.2 provides a piecewise-defined subgradient.

Proposition 4.2. *Let $L(q)$ be a convex loss function that is differentiable at $q \in \mathbb{R}/q_0$; then a subgradient $s_L(q) \in \partial L(q)$ of the loss function is given as*

$$s_L(q) = \begin{cases} L'(q) & \text{for } q \neq q_0 \\ s_0 & \text{for } q = q_0, \end{cases} \tag{4.13}$$

where $L'(q) = \frac{dL(q)}{dq}$ and $s_0 \in \mathbb{R}$ is a subgradient of $L(q)$ at q_0, such that $L(q) \geq L(q_0) + s_0(q - q_0)\ \forall q$.

For the newsvendor loss function, this result allows us to derive a subgra-

[46] For example, the algorithm presented in Studniarski (1989) may require about twenty iterations to estimate a single subgradient.

dient $s_{L_{NV}}(q,d) \in \partial_q L_{NV}(q,d)$ with respect to decision q as

$$s_{L_{NV}}(q,d) = \begin{cases} c_o & \text{for } q > d \\ 0 & \text{for } q = d \\ -c_u & \text{for } q < d. \end{cases} \tag{4.14}$$

Setting $s_{L_{NV}}(q,d) = 0$ for $q = d$ is arbitrary, as any value of $[-c_u, c_o]$ is sufficient for weak subgradient calculus. Similarly, we can use the results of Proposition 4.2 to derive a subgradient of the loss function $L_1(q) = (q)^+$:

$$s_{L_1}(q) = \begin{cases} 0 & \text{for } q < 0 \\ 0.5 & \text{for } q = 0 \\ 1 & \text{for } q > 0, \end{cases} \tag{4.15}$$

or of the loss function $L_2(q) = |q|$:

$$s_{L_2}(q) = \begin{cases} -1 & \text{for } q < 0 \\ 0 & \text{for } q = 0 \\ 1 & \text{for } q > 0. \end{cases} \tag{4.16}$$

Based on the results of Proposition 4.2 and these illustrative examples for common non-differentiable operators, one can derive a subgradient for many convex but non-differentiable OM loss functions.

4.4.2 Subgradients for Stochastic Problems with Recourse

Besides the OM loss functions that contain the non-differentiable operators presented in Section 4.4.1, many common OM problems follow the structure of (two-stage) stochastic problems with recourse: In the first stage, a decision-maker makes a decision under (demand) uncertainty; when demand is realized, a recourse action is taken in the second stage. Imagine the illustrative example of a car rental company that plans its (long-term) capacity in terms of compact and mid-sized cars under demand uncertainty (Netessine et al. 2002).

Each day, once demand is realized, the company has the option to upgrade by, for example, offering a mid-sized car to a customer who requests a compact car "to satisfy unexpectedly high demand for compact cars" (Netessine et al. 2002, p. 375). Netessine et al. (2002) mention other examples that follow this problem structure, including business class upgrades in commercial aviation, upgrades in time-shared executive jets, or experience-level upgrades when allocating hardware technicians in the telecommunication industry. A large class of two-stage shipment planning problems has a similar structure, such as the problem of allocating products in a warehouse network (Bertsimas and Kallus 2020).

All of these examples have in common a decision-maker who wants to minimize costs (or maximize profits) at both stages. Therefore, we define a general form of the loss function for such (two-stage) stochastic optimization problems with recourse as

$$\begin{aligned} L_{\boldsymbol{d}}(\vec{q}) &= f_{0,\boldsymbol{d}}(\vec{q}) + f^*_{1,\boldsymbol{d}}(\vec{q}) \\ \text{with } f^*_{1,\boldsymbol{d}}(\vec{q}) &= \min_{\boldsymbol{z}\in\mathcal{Z}} f_1(\boldsymbol{z},\boldsymbol{d}) \\ &\qquad s.t.\ g_j(\boldsymbol{z},\boldsymbol{d}) \leq q_j \ \forall j \\ &\qquad\quad h_{1,l}(\boldsymbol{z},\boldsymbol{d}) \leq u_l \ \forall l \\ &\qquad\quad h_{2,m}(\boldsymbol{z},\boldsymbol{d}) = v_m \ \forall m, \end{aligned} \tag{4.17}$$

where $f_{0,\boldsymbol{d}}(\vec{q})$ is convex and differentiable, $f^*_{1,\boldsymbol{d}}(\vec{q})$ is jointly convex in $\vec{q}$, and $u_l, v_m \in \mathbb{R}$ are constant. Assuming that $\vec{q}$ is a capacity decision made at the first stage (e.g., the number of compact and mid-sized cars to be purchased by the car rental company), we can interpret the second-stage variable $\boldsymbol{z}$ as the allocation of this capacity $\vec{q}$ to the realized demands $\boldsymbol{d}$. The functions $h_{1,l}(\boldsymbol{z},\boldsymbol{d})$ and $h_{2,m}(\boldsymbol{z},\boldsymbol{d})$ allow us to include additional constraints at the second stage (recourse action). Despite the generality of this formulation, which allows us to represent the most common two-stage stochastic OM problems, we can derive a subgradient that makes STB applicable. We use perturbation theory of convex optimization, as presented in Boyd and Vandenberghe (2004) and Boyd et al. (2018), to derive an arbitrary subgradient $\vec{s}_{L,\boldsymbol{d},\vec{q}}$ of $L_{\boldsymbol{d}}(\vec{q})$ with

respect to the decision $\vec{q}$.

Proposition 4.3. *(Rockafellar 1970) Assume a loss function $L_{\boldsymbol{d}}(\vec{q}) : \mathcal{Q} \to \mathbb{R}$ of the form defined in (4.17) with $\mathcal{Q} \subset \mathbb{R}^I$, $\operatorname{relint} \mathcal{Q} \neq \emptyset$, $f_{0,\boldsymbol{d}}(\vec{q})$ convex and differentiable, $f^*_{1,\boldsymbol{d}}(\vec{q})$ jointly convex in $\vec{q}$, and $u_l, v_m \in \mathbb{R}$. Assume also $f^*_{1,\boldsymbol{d}}(\vec{q}) < \infty$ for at least one $\vec{q} \in \mathcal{Q}$ and $f_{0,\boldsymbol{d}}(\vec{q}) < \infty$, $f_{0,\boldsymbol{d}}(\vec{q}), f^*_{1,\boldsymbol{d}}(\vec{q}) > -\infty$ for all $\vec{q} \in \mathcal{Q}$. Then the subdifferential $\partial L_{\boldsymbol{d}}(\vec{q})$ can be expressed as*

$$(\partial L_{\boldsymbol{d}}(\vec{q}))_j = \frac{\partial f_{0,\boldsymbol{d}}(\vec{q})}{\partial q_j} + (\partial f^*_{1,\boldsymbol{d}}(\vec{q}))_j. \tag{4.18}$$

The result of Proposition 4.3 allows us to express all subgradients as the sum of the gradient of the differential part $f_{0,\boldsymbol{d}}(\vec{q})$ and a subgradient of the non-differentiable part $f^*_{1,\boldsymbol{d}}(\vec{q})$ of the loss function.

Theorem 4.1. *Assume a loss function $L_{\boldsymbol{d}}(\vec{q}) : \mathcal{Q} \to \mathbb{R}$ of the form defined in (4.17) with $\mathcal{Q} \subset \mathbb{R}^I$, $\operatorname{relint} \mathcal{Q} \neq \emptyset$, $f_{0,\boldsymbol{d}}(\vec{q})$ convex and differentiable, $f^*_{1,\boldsymbol{d}}(\vec{q})$ jointly convex in $\vec{q}$, and $u_l, v_m \in \mathbb{R}$. Assume also $f^*_{1,\boldsymbol{d}}(\vec{q}) < \infty$ for at least one $\vec{q} \in \mathcal{Q}$ and $f_{0,\boldsymbol{d}}(\vec{q}) < \infty$, $f_{0,\boldsymbol{d}}(\vec{q}), f^*_{1,\boldsymbol{d}}(\vec{q}) > -\infty$ for all $\vec{q} \in \mathcal{Q}$, and that for the minimization defining $f^*_{1,\boldsymbol{d}}(\vec{q})$, strong duality holds with the dual optimum at $\vec{q}$ given as $(\vec{\alpha}^*_{\vec{q}}, \vec{\beta}^*_{1,\vec{q}}, \vec{\beta}^*_{2,\vec{q}})$. Then a subgradient of $L_{\boldsymbol{d}}(\vec{q})$ is given as*

$$(\vec{s}_{L,\boldsymbol{d},\vec{q}})_j = \frac{\partial f_{0,\boldsymbol{d}}(\vec{q})}{\partial q_j} - (\vec{\alpha}^*_{\vec{q}})_j. \tag{4.19}$$

Theorem 4.1 combines the result of Proposition 4.3 with perturbation theory (Boyd and Vandenberghe 2004, Boyd et al. 2018) and provides a subgradient of the loss function $L_{\boldsymbol{d}}(\vec{q})$ that can be used for STB. In the following sections we derive subgradients for the mail sorting and the aviation maintenance capacity planning problems introduced in Section 4.1.

4.4.3 Mail Sorting Capacity—A Two-Stage Capacity Planning Problem

In this section we present a mail logistics provider's two-stage capacity planning problem, which Notz and Pibernik (2021) study using prescriptive an-

alytics and was originally introduced (in a similar formulation) by Netessine et al. (2002) and Bassok et al. (1999). We revisit the loss function and use the results of Theorem 4.1 to derive a subgradient for STB.

Consider a company that provides I services using I types of capacities. At the first stage, the company plans the capacities q_j for a horizon of T periods at a reservation cost f_j, which are allocated to demands d_i on a period-by-period basis in the second stage. The company has the option of upgrading; that is, demand of type i can be fulfilled using capacity of type $j \leq i$, which achieves a marginal profit $a_{ij} = p_i - v_j + c_i$, where p_i is the revenue from fulfilling demand of type i, v_j is the cost of using capacity of type j, and c_i is the penalty cost for not fulfilling demand of type i (see Notz and Pibernik 2021 for a more detailed description).

This problem can be stated as a two-stage stochastic optimization problem:

$$\begin{aligned}
&\text{Stage 1: } \max_{\vec{q}, q_j \geq 0} \Pi(\vec{q}) = \max_{\vec{q}, q_j \geq 0} \left(\sum_{t=1}^{T} \mathbb{E}\left(\pi(\vec{D}^t, \vec{q})\right) - \sum_j f_j q_j \right) \\
&\text{Stage 2: } \pi(\vec{d}, \vec{q}) = \max_{\{y_{ij}\}} \sum_{i,j} a_{ij} y_{ij} - \sum_i c_i d_i \\
&\qquad\qquad \text{s.t.} \sum_j y_{ij} \leq d_i \ \forall i \\
&\qquad\qquad \sum_i y_{ij} \leq q_j \ \forall j \\
&\qquad\qquad y_{ij} \geq 0 \ \forall i, j \\
&\qquad\qquad y_{ij} = 0 \text{ if } i < j.
\end{aligned} \tag{4.20}$$

Because this planning problem aims at maximizing profit, we define the loss function as gap to optimal profit (Notz and Pibernik 2021), that is, as the difference between the achieved profit $\Pi(\vec{q}, \boldsymbol{d})$ and the optimal profit $\Pi^*(\boldsymbol{d}) := \max_{\vec{q}} \Pi(\vec{q}, \boldsymbol{d})$:

$$L(\vec{q}, \boldsymbol{d}) = \Pi^*(\boldsymbol{d}) - \Pi(\vec{q}, \boldsymbol{d}) = \Pi^*(\boldsymbol{d}) + \sum_{t=1}^{T} -\pi(\vec{d}^t, \vec{q}) + \sum_j f_j q_j, \tag{4.21}$$

where the negative allocation profit for each period is defined as:

$$\begin{aligned} -\pi(\vec{d},\vec{q}) = \min_{\{y_{ij}\}} \; & \sum_i c_i d_i - \sum_{i,j} a_{ij} y_{ij} \\ \text{s.t.} \; & \sum_j y_{ij} \leq d_i \; \forall i \\ & \sum_i y_{ij} \leq q_j \; \forall j \\ & y_{ij} \geq 0 \; \forall i,j \\ & y_{ij} = 0 \text{ if } i < j. \end{aligned} \tag{4.22}$$

Proposition 4.4. *(Notz and Pibernik 2021) The loss function $L(\vec{q},\boldsymbol{d})$, as defined in (4.21), is jointly convex in $\vec{q}$.*

Proposition 4.5. *Assume $\mathcal{Q} \subset \mathbb{R}_+^I$ convex, open, and non-empty; then for all $\vec{q} \in \mathcal{Q}$, the subdifferential $\partial L_{\boldsymbol{d}}(\vec{q})$ is non-empty and at least one subgradient exists.*

Because this loss function is convex (Proposition 4.4), we can prove the existence of a subgradient (Proposition 4.5). Using the result of Theorem 4.1, we derive in Proposition 4.6 a subgradient of the loss function that allows us to apply the STB approach to Problem 4.20.

Proposition 4.6. *Assume $\mathcal{Q} \subset \mathbb{R}_+^I$ convex, open, and non-empty; then for all $\vec{q} \in \mathcal{Q}$, a subgradient of the loss function (4.21) is given as:*

$$(\vec{s}_{L,\boldsymbol{d},\vec{q}})_j = f_j - \sum_{t=1}^{T} \beta^*_{j,\vec{d}^t,\vec{q}} \; , \tag{4.23}$$

where $\{\beta^*_{j,\vec{d},\vec{q}}\}$ is the solution to

$$\begin{aligned} \max_{\{\alpha_i\},\{\beta_j\}} & \sum_i (c_i - \alpha_i)\, d_i - \sum_j \beta_j q_j \\ s.t. \; & \alpha_i, \; \beta_j \geq 0 \; \forall i,j \\ & \alpha_i + \beta_j \geq a_{ij} \; \forall i \geq j. \end{aligned} \tag{4.24}$$

4.4.4 Aviation Maintenance Capacity—A Two-Shift Capacity Planning Problem

In this section we present the two-shift staffing problem originally studied by Notz et al. (2020) and use the results of Theorem 4.1 to derive a subgradient for the STB approach.

Consider a company that faces uncertain hourly demand arrivals in $T = 20$ time periods between 0 a.m. and 8 p.m. each day. The company wants to plan staff capacity for processing unknown demand in a two-shift structure. Capacity reservations incur a cost of c_q per hour, while the first (second) shift has a duration of $\tau_a = 8$ ($\tau_b = 6$) hours. Unfulfilled demand backlogged between the shifts incurs a cost c_3 per item, and leftover demand at the end of the second shift incurs an overtime cost of c_2 per item. See Notz et al. (2020) for additional details and the derivation of the following problem statement:

$$\begin{aligned} \min_{\vec{q}=(q_a,q_b)} C(\vec{q}) &:= c_q\left(\tau_a q_a + \tau_b q_b\right) + \mathbb{E}[c_2 M_{20} + c_3 M_{14}] \\ \text{s.t. } M_t &= \begin{cases} 0 \text{ for } t = 0 \\ (M_{t-1} + D_t)^+ \text{ for } 1 \le t \le 6 \\ (M_{t-1} + D_t - q_a)^+ \text{ for } 7 \le t \le 14 \\ (M_{t-1} + D_t - q_b)^+ \text{ for } 15 \le t \le 20, \end{cases} \end{aligned} \tag{4.25}$$

where M_t is the length of the queue at the end of time period t. Following Notz et al. (2020), the corresponding linearized loss function can be defined as:

$$L(\vec{q}, \vec{d}) = c_q\left(\tau_a q_a + \tau_b q_b\right) + C_{\text{backlog}}(\vec{q}, \vec{d}), \tag{4.26}$$

where the backlogging cost is given as

$$
\begin{aligned}
C_{\text{backlog}}(\vec{q}, \vec{d}) = \min_{\{m_t\} \in \mathbb{R}^{21}} \; & (c_2 m_{20} + c_3 m_{14}) \\
\text{s.t. } m_t \geq & \begin{cases} m_{t-1} + d_t \text{ for } 1 \leq t \leq 6 \\ m_{t-1} + d_t - q_a \text{ for } 7 \leq t \leq 14 \\ m_{t-1} + d_t - q_b \text{ for } 15 \leq t \leq 20 \end{cases} \\
& m_t \geq 0 \; \forall t \\
& m_0 = 0.
\end{aligned}
\tag{4.27}
$$

Proposition 4.7. *(Notz et al. 2020) The loss function $L(\vec{q}, \vec{d})$, as stated in (4.26), is jointly convex in $\vec{q}$.*

Proposition 4.8. *Assume $\mathcal{Q} \subset \mathbb{R}^2_+$ convex, open, and non-empty; then for all $\vec{q} \in \mathcal{Q}$, the subdifferential $\partial L_{\vec{d}}(\vec{q})$ is non-empty and at least one subgradient exists.*

Using the convexity of the loss function (Proposition 4.7), we can prove the existence of a subgradient (Proposition 4.8). Similar to Section 4.4.3, we derive in Proposition 4.9 a subgradient of the loss function (4.26) using Theorem 4.1, which allows us to apply STB.

Proposition 4.9. *Assume $\mathcal{Q} \subset \mathbb{R}^2_+$ convex, open, and non-empty; then for all $\vec{q} \in \mathcal{Q}$, a subgradient of the loss function (4.26) is given as:*

$$
\vec{s}_{L,\vec{d},\vec{q}} = \begin{pmatrix} c_q \tau_a - \sum_{t=7}^{14} \beta^*_{t,\vec{d},\vec{q}} \\ c_q \tau_b - \sum_{t=15}^{20} \beta^*_{t,\vec{d},\vec{q}} \end{pmatrix}, \tag{4.28}
$$

where $\{\beta^*_{t,\vec{d},\vec{q}}\}$ *is the solution to*

$$\begin{aligned} \max_{\{\beta_t\}} & \sum_{t=1}^{20} \beta_t d_t - \sum_{t=7}^{14} \beta_t q_a - \sum_{t=15}^{20} \beta_t q_b \\ s.t.\ & \beta_{t+1} \geq \beta_t \ \forall t \neq 14, 20 \\ & \beta_{15} \geq \beta_{14} - c_3 \\ & \beta_{20} \leq c_2. \end{aligned} \tag{4.29}$$

4.5 Numerical Evaluation

This section applies the STB approach to the mail sorting capacity planning problem (Section 4.4.3) and the aviation maintenance capacity planning problem (Section 4.4.4); benchmarks its performance against those of wSAA, kERM, and the traditional (feature-less) SAA approach; and demonstrates how its prescriptions can be explained using SHAP values. We study STB's performance using historical demand data from two case companies and cost parameters that lead to a variety of optimal service levels.

4.5.1 Problem Statement and Parameter Settings

Our analysis of the mail sorting capacity planning problem is inspired by our work with a German logistics provider that collects, sorts, and delivers mail. The provider has to plan the capacity $\vec{q}$ of $I = 3$ service lines, which is constant for a horizon of one week ($T = 5$). After the demand arrives on each day t of the horizon, the decision-maker allocates the planned capacities to the demand, giving consideration also to the upgrading option. This capacity planning problem is formally stated in Problem 4.20, and we use cost parameters (shown in Table 4.1) that induce a variation of the optimal service level[47] (Notz and Pibernik 2021).

Our second analysis is inspired by our work with a German aviation maintenance provider that maintains aircraft parts for various airlines. In particu-

[47] We define the approximate optimal service level as $\mathrm{SL}_i = \frac{C_{U,i}}{C_{U,i}+C_{O,i}}$, with $C_{O,i} \approx f_i/5$ and $C_{U,i} \approx p_i + c_i - v_i - f_i/5$.

Table 4.1: Mail Sorting Capacity—Parameter setting.

f	v	p	c	$a_{i,i}$
$\begin{pmatrix} 2700..142 \\ 1130..60 \\ 500..26 \end{pmatrix}$	$\begin{pmatrix} 40 \\ 30 \\ 20 \end{pmatrix}$	$\begin{pmatrix} 600 \\ 260 \\ 120 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 7.5 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 570 \\ 238 \\ 105 \end{pmatrix}$

lar, we study the problem of planning staff capacity for two shifts in the facility where the aircraft parts arrive and are processed. The first (second) shift has a duration of $\tau_a = 8$ ($\tau_b = 6$) hours, while parts arrive throughout the day. In addition to the capacity cost c_q, we assume a cost c_3 for backlogging demand between the two shifts (which is the intangible cost of congestion and can be tuned for load-balancing between the shifts) and an overtime cost c_2 (which is incurred to process remaining demand at the end of the day). This planning problem is formalized in Problem 4.25. We use cost parameters (shown in Table 4.2) that induce a variation of the optimal service level[48] (Notz et al. 2020).

Table 4.2: Aviation Maintenance Capacity—Parameter setting.

Capacity cost c_q	**Overtime cost** c_2	**Backlogging cost** c_3
5...0.3	5.25	0.6...0.06

4.5.2 Demand Data and Feature Engineering

All numerical analyses of both capacity planning problems are based on historical demand data that we received from the case companies and features (co-variates) that we constructed. In the mail logistics provider's case, we received historical demand data for the 2014-2017 period, based on which we constructed a data set $S_N^{Mail} = \{(\boldsymbol{d}^1, \vec{x}^1), ..., (\boldsymbol{d}^N, \vec{x}^N)\}$ with $N = 209$ weeks, feature vectors $\vec{x}^n \in \mathbb{R}^p$ consisting of $p = 162$ features, and demand matri-

[48]The approximate optimal service level is defined as $\text{SL} = \frac{C_U}{C_U + C_O} = 1 - \frac{c_q}{c_2}$.

ces $\boldsymbol{d}^n \in \mathbb{R}^{I \times T}$, which represent demand for each service line i on each day t of the week. The feature vectors $\vec{x}^n$ include date-based features (e.g., the year and the week number), lagged demand features (e.g., demand in the same week one year ago), and public holiday features (e.g., indicators of a public holiday that was a few days before or after the week in focus). A detailed description of all features can be found in Appendix C in Notz and Pibernik (2021).

For the analysis of the aviation maintenance capacity planning problem, we received historical demand data for the 2016-2017 period, based on which we constructed a data set $S_N^{Aviation} = \{(\vec{d}^1, \vec{x}^1), ..., (\vec{d}^N, \vec{x}^N)\}$ with $N = 532$ days, feature vectors $\vec{x}^n \in \mathbb{R}^p$ consisting of $p = 142$ features, and demand vectors $\vec{d}^n \in \mathbb{R}^T$, which represent the demand for each hour t of the day in focus. Similar to the mail sorting case, the feature vectors $\vec{x}^n$ include date-based features, lagged demand features, and public holiday features. In addition, the feature vectors contain business-related features like expected demand for the day in focus. A detailed description of all features used in our analyses can be found in Appendix B in Notz et al. (2020).

4.5.3 Evaluation Procedure

We split the data sets into training data ($N = 157$ weeks for the mail sorting case and $N = 425$ days for the aviation maintenance case) and test data ($N_{Test} = 52$ weeks for the mail sorting case and $N_{Test} = 107$ days for the aviation maintenance case) and evaluate STB and the benchmark approaches by comparing the performance of their prescriptions for the test data:

1. STB: Use Algorithm 4.1 (Section 4.3.1) with loss functions defined in (4.21) and (4.26) and subgradients (4.23) and (4.28) to determine the STB prescription function, which is used to prescribe capacities for each week/day of the test period.

2. wSAA: Apply wSAA with the random forest weight function to Problems 4.20 and 4.25, as presented in Notz and Pibernik (2021) and Notz et al. (2020).

3. kERM: Apply kERM with the random forest kernel to Problems 4.20 and 4.25, as presented in Notz and Pibernik (2021) and Notz et al. (2020).

4. SAA: Solve the wSAA approach with $w_n(\vec{x}) = 1/N$ for both capacity planning problems to obtain the SAA capacity prescription, which is constant for the test period.

A more detailed description of all approaches is provided in Appendix C.3. For the mail sorting case, all approaches' levels of performance are evaluated in terms of the gap to optimal profit: $\Delta_{\Pi,\text{abs}} = \Pi^*(\boldsymbol{d}) - \Pi(\vec{q}, \boldsymbol{d})$, where $\Pi^*(\boldsymbol{d})$ is the optimal profit for a given demand $\boldsymbol{d}$. For the aviation maintenance case, we use the gap to optimal cost: $\Delta_{C,\text{abs}} = L(\vec{q}, \vec{d}) - C^*(\vec{d})$, where $C^*(\vec{d})$ is the optimal cost for a given demand $\vec{d}$, to evaluate the approaches' levels of performance.

4.5.4 Performance Results

Figure 4.1 shows the absolute gap to optimal profit for the mail sorting case for optimal service levels in the range of 5-95 percent. Across the range of services levels, all prescriptive analytics approaches outperform SAA[49], which is the only approach that does not incorporate feature data, demonstrating the features' prescriptive value. Among the prescriptive approaches, STB leads to comparable or better results than kERM and wSAA for all service levels above 40 percent. Therefore, we conclude that, at least in this instance (and for service levels above 40 percent), STB is the preferable approach, because it provides at least a similar level of performance and is inherently explainable (Section 4.5.5).

However, the conclusion we draw for the mail sorting case does not apply to the aviation maintenance case. Figure 4.2 plots the absolute gap to optimal cost for all approaches across various service levels. While STB's performance is comparable to that of kERM, both approaches' level of performance is lower

[49]For very high service levels, kERM's performance is comparable to that of SAA because of the regularization effect (Notz and Pibernik 2021).

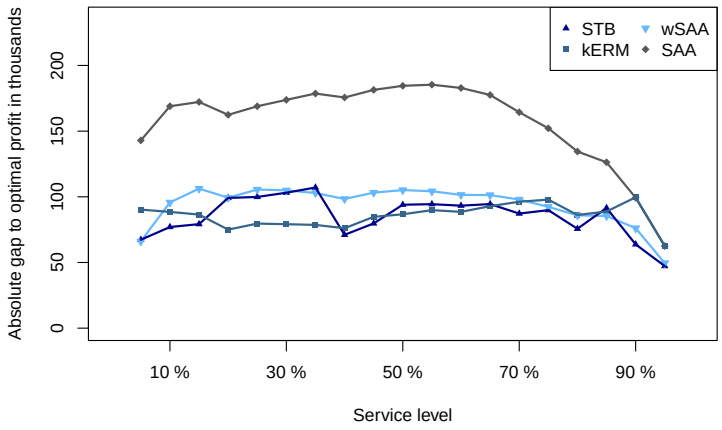


Figure 4.1: Absolute gap to optimal profit for the mail sorting case.

than that of wSAA. The performance gap between wSAA and STB is particularly significant for low or high service levels. We explain this result in terms of the regularization effect, which Notz et al. (2020) introduce to explain kERM's inferior performance compared to that of wSAA: for low or high service levels, it appears optimal to choose a higher regularization parameter, which leads to a higher (negative or positive) safety buffer and allows kERM to adapt its prescription according to the service-level regime. However, this increased regularization reduces the variance between the weekdays, which results in performance that is inferior to that of wSAA. This regularization effect applies similarly to STB, where stronger regularization corresponds to choosing a smaller number of iterations K and, therefore, a smaller number of base learners that define the prescription function. Because STB's performance level is lower than that of wSAA, we face in this instance a trade-off between prescription performance and explainability. While wSAA is the preferable approach in terms of performance, STB has the advantage of being inherently explainable.

The next section shows how the STB prescriptions can be explained using SHAP values.

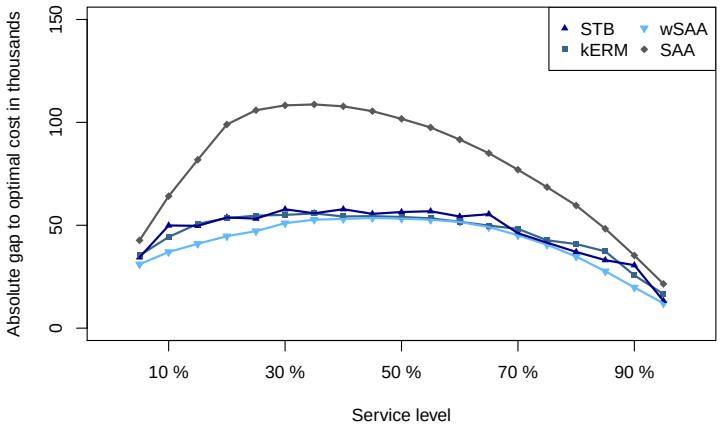
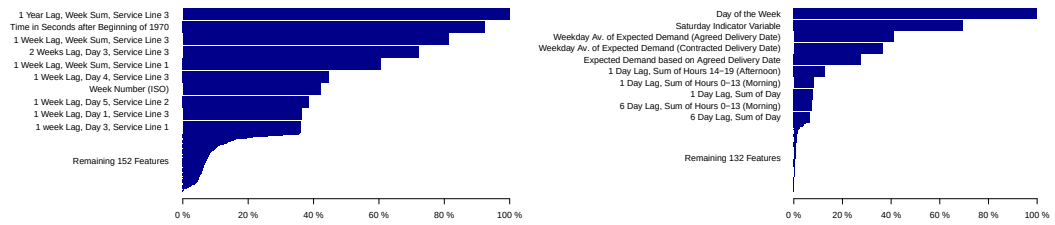


Figure 4.2: Absolute gap to optimal cost for the aviation maintenance case.

4.5.5 Explaining STB Prescriptions

In this section we derive the SHAP values for STB prescriptions and demonstrate these prescriptions' explainability by breaking capacity prescriptions down into individual features' impacts and SHAP dependence diagrams that show how a feature's value drives the prescription.

To contrast the STB explanations with the insights that can be generated for wSAA and kERM, Figure 4.3 plots feature importance analyses that are carried out based on the weight or kernel functions of wSAA and kERM, respectively.[50] These analyses show the degree to which an individual feature affects the prescription function but do not explain individual prescriptions.

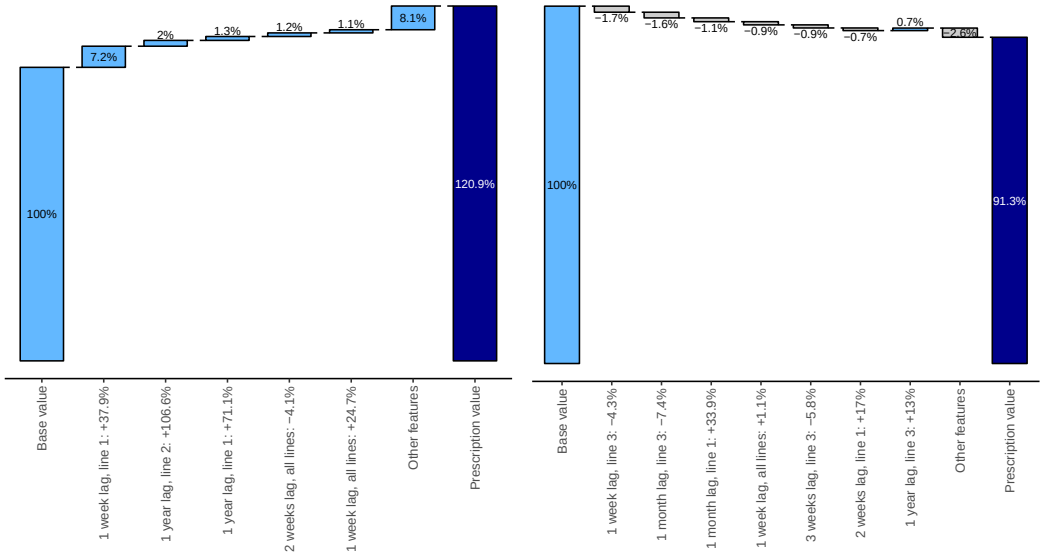


(a) Mail sorting case. (b) Aviation maintenance case.

Figure 4.3: Feature importance analyses for mail sorting and aviation maintenance cases.

[50] A feature's importance is measured as the decrease in the node impurity of the random forests that define the wSAA weight function and the kERM kernel function.

In contrast, the explanations that can be derived for the STB prescriptions shed light on which feature and feature value impact the prescription and by how much. We aggregate the impact of several similar features because using the large number of features (162 for the mail sorting case or 142 for the aviation maintenance case) leads to overly complex explanations where many similar features have a comparatively small impact.[51] Based on this aggregation, we present several illustrative examples of breaking the prescriptions down into a base value—the average prescription of the STB model with no features, mostly driven by the SAA initialization—and the impacts of individual aggregated features.



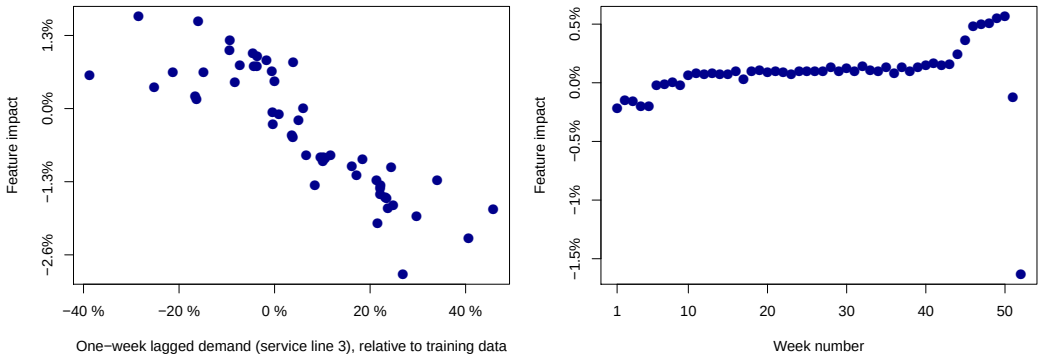
(a) Service line 1, week 42 (test period). (b) Service line 3, week 30 (test period).

Figure 4.4: Breakdown of prescriptions for the mail sorting case (SL=50%).

Figure 4.4 shows the breakdown of two exemplary prescriptions into aggregated feature impacts for the mail sorting case (week 42 in Figure 4.4a and week 30 in Figure 4.4b). The depiction of the breakdown in Figure 4.4a suggests that the prescription for service line 1 increases by 7.2 percent with

[51] We add the SHAP values of a group of features and average the percentage deviation of their feature values from the features' mean values (of the training data). A detailed description of the aggregated features can be found in Appendix C.2.

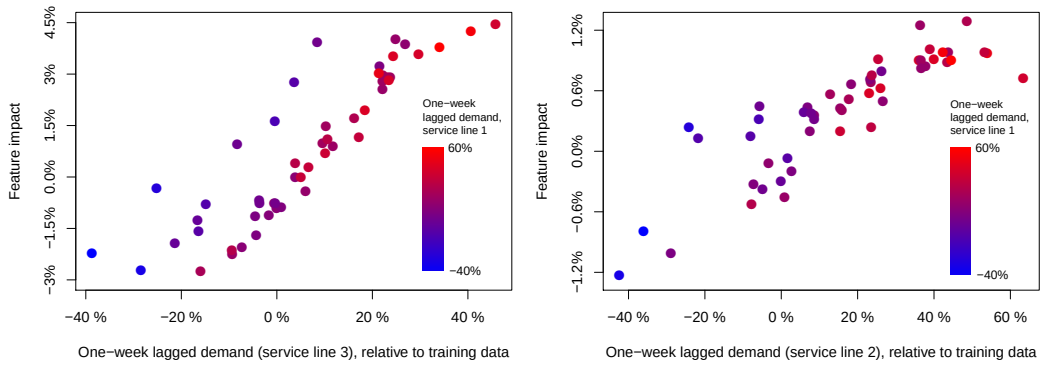
respect to the base value because the demand for the same service line in the preceding week was 37.9 percent above average. The prescription also increases because of, for example, an above-average demand for service lines 1 and 2 in the same week of the preceding year. Similarly, Figure 4.4b shows how the prescription for service line 3 decreases because of, for example, an average demand that is 4.3 percent below average in the preceding week for the same service line. These breakdowns also demonstrate the interdependency among the three service lines, as the prescription for service line 1 (Figure 4.4a) depends on lagged demand for service line 2, and the prescription for service line 3 (Figure 4.4b) depends on lagged demand for service line 1. This interdependency can be explained by correlations between the demands and the upgrading option (Notz and Pibernik 2021). These explanations help a decision-maker to understand the rationale behind a certain prescription, which is often a prerequisite of acceptance and trust.



(a) Prescription for service line 2. (b) Prescription for service line 3.

Figure 4.5: SHAP dependence plots: feature impact (relative to base value) depending on the feature value (relative to mean) for the mail sorting case.

In addition, the SHAP values allow us to elucidate how the individual features' values drive STB's prescriptions. Figure 4.5 plots the impact on prescriptions of the feature "One-week lagged demand for service line 3" and the week number. These depictions are *SHAP dependence plots* (Lundberg et al. 2020). Figure 4.5a shows that a higher demand for service line 3 in the preceding week leads to a lower prescription for service line 2 capacity,



(a) Prescription for service line 3. (b) Prescription for service line 1.

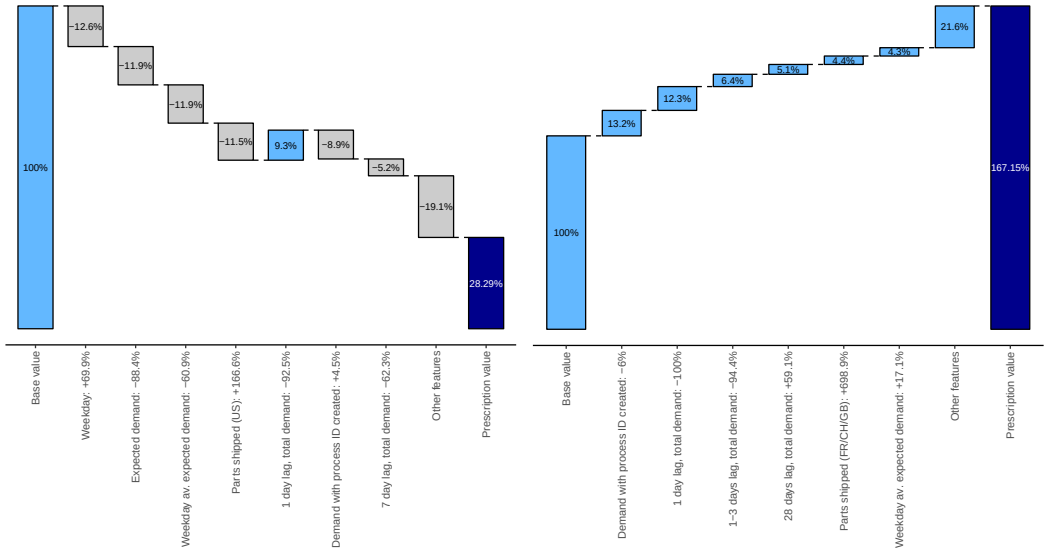
Figure 4.6: SHAP dependence plots: feature impact (relative to base value) depending on the feature value (relative to mean) with interaction for the mail sorting case.

suggesting a correlation between the demands for the two service lines. The impact of the week number on service line 3's capacity, shown in Figure 4.5b, suggests a non-linear relationship with slightly lower (negative) impact on the capacity prescription at the beginning of a year, almost zero impact between weeks 10 and 40, and a positive impact toward the end of each year.

The impact on service line 3's prescription, depicted in Figure 4.5b, appears to vary little for a given week number, but such is not always the case. For many SHAP dependence plots (e.g., Figure 4.5a), the variance along the vertical axis is significant, so the impact of a given feature's value varies widely. This variance can often be explained by a second feature, which suggests that interaction effects between features may be present. To explain such interaction effects, Figure 4.6 shows SHAP dependence plots that depict the value of a one-week lagged demand for service line 1 as a second feature (as in Lundberg et al. 2020).

The impact of a one-week lagged demand for service line 3 on the prescription for the same service line (Figure 4.6a) suggests that, in the medium range of -10 percent to $+10$ percent, the lagged demand for service line 3 has a higher (more positive) impact when the lagged demand for service line 1 is lower, and a lower (less positive) impact when the service line 1 lagged demand is higher. We explain this interaction between the two features by

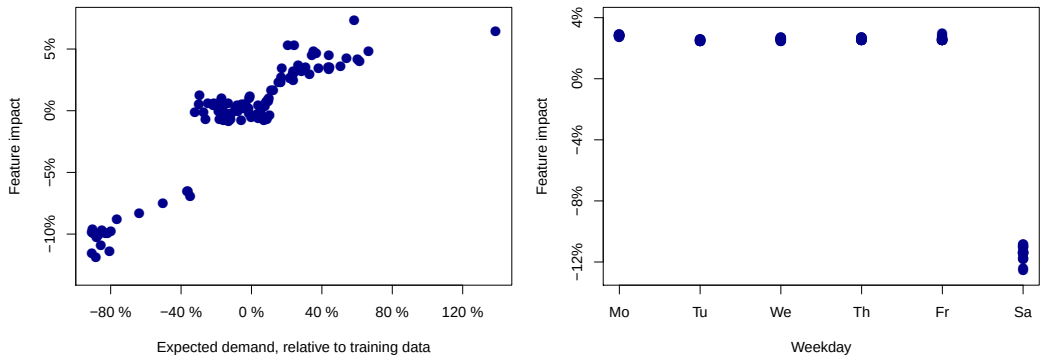
means of the upgrading option: A higher lagged demand for service line 1 may lead to higher capacity in service line 1, which can also satisfy the demand for service line 3 (through upgrading), so less dedicated service line 3 capacity is required. A similar interaction effect can be seen in the service line 1 prescription depicted in Figure 4.6b.



(a) Shift 2, day 7 (test period, Saturday). (b) Shift 2, day 8 (test period, Monday).

Figure 4.7: Breakdown of prescriptions for the aviation maintenance case (SL=50%).

One can obtain similarly structured explanations of the STB prescriptions in the aviation maintenance case. Figure 4.7 shows the breakdown of the capacity prescriptions for shift 2 for a Saturday (Figure 4.7a) and a Monday (Figure 4.7b). Several of the features that have large impacts on the prescription are related to the day of the week (e.g., the weekday feature, the weekday average expected demand, the one-day lagged demand feature), which is consistent with Notz et al.'s (2020) finding that the weekday is the most important feature. In addition to the weekday, features that describe the demand for which a process ID was created or that describe the parts that were shipped before the date in focus have significant impacts on the prescription.



(a) Feature: Expected demand.

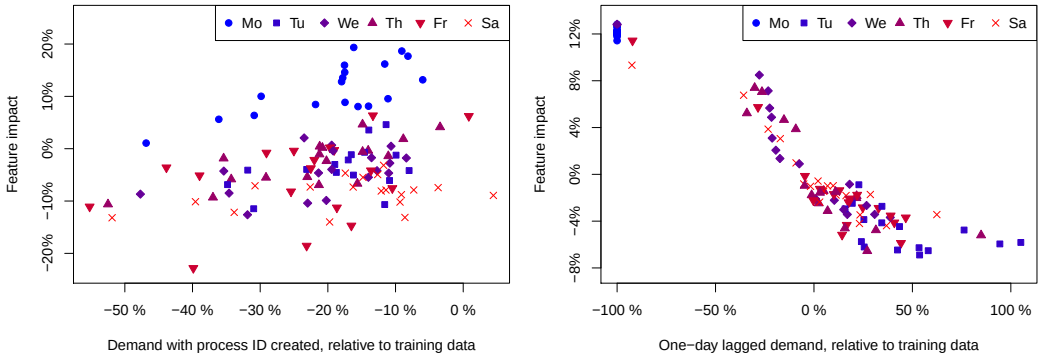
(b) Feature: Weekday.

Figure 4.8: SHAP dependence plots: feature impact (relative to base value) depending on the feature value (relative to mean) for shift 2's capacity prescription for the aviation maintenance case.

The SHAP dependence plot depicted in Figure 4.8a shows that higher expected demand (based on the estimated delivery date that was updated after shipping) leads to a larger capacity prescription; the plot shows a non-linear step at ≈ -30 percent, where smaller expected demand leads to a significantly smaller (more negative) impact on the capacity prescription. Figure 4.8b shows that the weekday feature points to a strongly reduced capacity on Saturdays, which we explain by the smaller average demand on Saturdays (Notz et al. 2020).

Figure 4.9a depicts, for the individual weekdays, the impact on the prescription of the feature describing demand for which a process ID was created. Without the weekday differentiation, the prescription impact of a feature value of -20 percent ranges between -20 percent and $+20$ percent, a high variance. This range can be reduced significantly by including the weekday feature, which suggests an interaction effect between both features: On Mondays, the feature value of -20 percent increases the prescription by about 15 percent, while on Saturdays, the same feature value decreases the prescription by about 15 percent. A similar interaction effect can be observed in Figure 4.9b, where the impact of the one-day lagged demand feature is typically higher (more positive) on Saturdays than it is on, for example, Tuesdays.

We conclude that the explanations derived for the STB prescriptions pro-



(a) Demand with process ID created.

(b) One-day lagged demand.

Figure 4.9: SHAP dependence plots: feature impact (relative to base value) depending on the feature value (relative to mean) with interaction for shift 2's capacity prescription for the aviation maintenance case.

vide valuable insights into the model's mechanics and allow the decision-maker to understand the rationale behind individual prescriptions, thus facilitating acceptance and building trust. In addition, SHAP dependence plots provide insights into the dependence between features' value and the impact on prescriptions and shed light on complex interaction effects between features.

In the case of the mail logistics provider, STB's performance is comparable to those of wSAA and kERM. In combining this observation with the explainability of its prescriptions—the derived explanations provide significantly more valuable insights than those that can be derived for wSAA and kERM using feature importance analyses—, we can conclude that STB is the preferable approach for this particular instance. However, in the case of the aviation maintenance service provider, wSAA performs better than STB across all service levels, which leads to a trade-off between prescription performance and explainability. Overall, then, STB leads to a gap to optimal cost that averages 13.6 percent higher than that of wSAA; however, depending on the service level, the increase can be as low as 2 percent, which is what it is at a service level of 70 percent. Even so, in contrast to wSAA, STB provides detailed explanations of its prescriptions. Which approach is preferable depends on the decision-maker's particular requirements regarding prescription performance and explainability.

4.6 Conclusion

This paper proposes a novel data-driven and explainable prescriptive analytics approach to solving complex OM problems. STB combines methods of (sub)gradient descent in function space with SAA, provides a function that prescribes decisions directly based on a data set of historical demand observations and predictive features, and allows a decision-maker to find detailed explanations for these prescriptions. We demonstrate how a subgradient can be derived for the most common OM problems, including the large class of two-stage stochastic problems with recourse, and we compute the SHAP values that allow decision-makers to explain individual prescriptions.

Using historical demand and feature data from two case companies—a mail logistics provider and an aviation maintenance provider—we demonstrate STB's applicability to two complex capacity planning problems and benchmark its performance against those of two other prescriptive approaches—wSAA and kERM—and SAA for a variety of cost parameter settings. We find detailed explanations for several exemplary prescriptions of both planning problems by determining the impact of individual features on the prescription, and use SHAP dependence plots to shed light on how individual features' value and the interaction effects between features drive the prescription. The results of the mail sorting case suggest that STB's performance is at least comparable to those of wSAA and kERM, while also providing explainable prescriptions. However, in the case of the aviation maintenance provider, wSAA's performance is superior, and a decision-maker faces the trade-off between prescription performance (wSAA) and explainability (STB).

Our study's limited scope—only two planning problems, two data sets, and a variety of cost parameters—demonstrates the difficulty of generalizing performance-related results of prescriptive analytics approaches beyond a particular instance: While STB's prescriptions provide detailed explanations in both settings, wSAA has superior performance in one of the settings. Therefore, additional studies that use a variety of OM problems and data sets are needed.

5 Summary and Conclusion

The rise of AI and machine learning, which impacts almost all areas in business and society, and the recent availability of large amounts of data, such as historical observations of demand, have led to the development of novel prescriptive analytics approaches to solving classical OM planning problems. These new approaches represent a significant opportunity to improve OM decision-making, so they constitute a new field of research within OM. With a focus on capacity planning, which entails solving complex OM planning problems, this dissertation addresses the guiding research question, introduced in Chapter 1, on how prescriptive analytics approaches to solving complex capacity planning problems can improve decision-making. This dissertation, which consists of three independent research articles, develops novel prescriptive analytics approaches (kERM, wSAA, OP, STB) to solving two realistic capacity planning problems (those of a mail logistics provider and an aviation maintenance provider) and derives analytical properties, including out-of-sample performance guarantees and the universal approximation property for kERM. The dissertation's comprehensive numerical studies benchmark the prescriptive approaches against traditional contenders, including two-step approaches and SAA; shed light on the underlying performance drivers; and demonstrate how explainable prescriptions help decision-makers understand the causality behind the decisions.

The first article (Chapter 2) addresses Research Question 1, introduced in Chapter 1, by developing two prescriptive analytics approaches, wSAA and kERM, to solving the complex two-stage capacity planning problem of a mail logistics service provider that observes multivariate demand and makes vector-valued capacity decisions. On the theoretical side, the article provides solutions for the kERM approach for non-linear function spaces, derives out-of-sample performance guarantees for kERM when using various kernels, and

shows kERM's universal approximation property when using a universal kernel. On the numerical side, comprehensive performance analyses are conducted using data from the logistics service provider and realistic cost parameters and parameter settings that vary the service level. The results suggest that the prescriptive approaches can lead to significant performance improvements compared to those of traditional contenders like SAA and two-step approaches. Our analyses shed light on the two prescriptive approaches' underlying performance drivers, and while the performance improvements depend on the service level, the prescriptive approaches appear to be much more robust to variations in the exogenous cost parameters than the traditional approaches are. This robustness is an important property, and prescriptive analytics approaches are attractive choices for solving a capacity planning problem.

Research Question 2 is addressed in Chapter 3 by using prescriptive analytics approaches to solve the MSSP of a maintenance service provider in the aviation industry, a complex queuing-type capacity planning problem with uncertain, time-varying rates of demand arrival without abandonment. The article first derives an approximated MSSP (AMSSP) by applying fluid and stationary approximations before solving the AMSSP using the prescriptive approaches wSAA and kERM. In addition, a novel prescriptive analytics approach, termed OP approach, is proposed, which has the advantage of simplicity because it requires the decision-maker only to solve a deterministic optimization problem and to use standard predictive machine learning tools to derive a prescription function. The results of numerical experiments conducted using historical observations of demand suggest that all three prescriptive approaches (wSAA, kERM, OP) overcome both the time-structure and feature effects that are likely to be associated with traditional approaches. The analyses shed light on the differences between the approaches' performance and show that wSAA appears to be the most suitable method for solving the AMSSP. This research demonstrates how a queuing-type capacity planning problem can be solved using prescriptive analytics, so it provides a foundation for connecting the "worlds" of queuing theory and prescriptive analytics.

The third article (Chapter 4) proposes a novel prescriptive analytics approach that provides explainable decisions, termed subgradient tree boosting

(STB), in answering Research Question 3. The STB approach is motivated by the success of gradient boosting methods in machine learning and the need for explanations for prescribed decisions. The paper's main methodological contributions are the proposed STB approach, which combines (sub)gradient descent in function space with SAA, and the demonstration of ways to derive a subgradient for the most common OM problems, including the large class of two-stage stochastic problems with recourse. The STB approach is used in numerical experiments to solve the capacity planning problems introduced in Chapters 2 and 3 (those of a mail logistics provider and an aviation maintenance provider), and its performance is benchmarked against those of wSAA and kERM. In addition, detailed explanations for several exemplary prescriptions are derived that help decision-makers to understand the causality behind the decisions. The results for the case of the mail logistics provider suggest that STB is the preferable approach because it leads to a performance that is at least comparable to those of wSAA and kERM while also providing explainable prescriptions. However, this observation does not hold for the case of the aviation maintenance provider, where wSAA had the best performance, resulting in a trade-off between performance and explainability.

In conclusion, prescriptive analytics approaches—in particular, wSAA, kERM and STB—provide promising new ways to solve complex capacity management problems. While the research presented in this dissertation makes an effort to characterize and study these approaches both theoretically and numerically, opportunities for future research remain.

The two case studies demonstrate that prescriptive approaches can lead to significant performance improvements compared to traditional approaches' performance. As in the domain of machine learning, which focuses on classification and regression tasks, the generalizability of these results remains an issue, as whether the approaches lead to the same relative levels of performance for different sets of historical observations in the same problem domain remains unclear. In addition, because the prescriptive analytics approaches rely on non-standard (asymmetric) cost functions, the results can be sensitive to variations in the (exogenous) parameters, making generalization even more difficult. Since the numerical results cannot be generalized beyond the

boundaries of the specific case studies presented here, one avenue for future research is to apply prescriptive analytics approaches to other sets of demand and feature data in the same problem domain.

A second direction for future research is to apply prescriptive analytics approaches to even more complex OM problems, such as the multi-period capacity management problems Shumsky and Zhang (2009) and Yu et al. (2015) addressed, multi-period inventory management problems, and more complex queuing-type problems, such as queuing models that include abandonment. Future research is needed that explores how prescriptive analytics approaches can be applied to these more complex classes of OM problems, and that studies the performance improvements relative to traditional approaches for these problem classes.

A third direction for future research lies in the areas of modern machine learning methods like deep reinforcement learning (DRL) (Mnih et al. 2013), which combines deep neural networks and reinforcement learning, and generative adversarial networks (GAN) (Goodfellow et al. 2014), which consist of two opposing neural networks. Such research may fuel the development of additional prescriptive analytics approaches to OM problems. For example, DRL has been applied to intractable inventory management problems in a stationary setting, where a large number of demand samples could be drawn from a known distribution (Gijsbrechts et al. 2019). However, when the demand distribution is unknown, the large amounts of training data required for DRL are typically not available because of a limited number of relevant historical observations. Therefore, the question concerning how to use such a DRL approach in non-stationary settings with feature-dependent demand and where only a limited number of historical observations is available remains unanswered. Generative models like GANs may help decision-makers generate realistic data samples from an unknown joint distribution of features and demand based on a training data set, but how best to exploit such enriched data sets for decision-making also remains an open research question. Consequently, new prescriptive analytics approaches driven by, for example, DRL and GAN methodologies or combinations of both mark a promising avenue for future research.

A Appendix of Chapter 2

A.1 Proofs

Proof of Proposition 2.1

The proof of Proposition 2.1 follows a similar structure as the proof of Proposition 1 in Netessine et al. (2002). The profit $\pi(\vec{d}, \vec{q})$ is determined by solving a linear program and therefore concave. Because taking the expectation preserves concavity, $\Pi(\vec{q})$ is a sum of concave and linear functions, and therefore jointly concave in $\vec{q}$.

Proof of Proposition 2.2

$\pi(\vec{d}, \vec{q})$ is jointly concave in $\vec{q}$ (Proposition 2.1), therefore $\Pi(\vec{q}, \boldsymbol{d})$ is concave as sum of concave and linear functions. With $-\Pi(\vec{q}, \boldsymbol{d})$ convex and $\Pi^*(\boldsymbol{d})$ constant with respect to $\vec{q}$, the loss function $L(\vec{q}, \boldsymbol{d})$ is a sum of convex and constant functions, and therefore jointly convex in $\vec{q}$.

Proof of Proposition 2.3

Because $L(\vec{q}, \boldsymbol{d})$ is jointly convex in $\vec{q}$ (Proposition 2.2) and the weights $w_n(\vec{x})$ are non-negative by definition, the objective function of wSAA is a weighted sum of convex functions, and therefore jointly convex in $\vec{q}$.

Proof of Proposition 2.4

To prove that $N^{-\frac{2l+1}{2l+p}}$ is an individual lower rate of convergence we apply the general results of Theorem 3.3 in Györfi et al. (2002) to our loss function and wSAA.

Following Györfi et al. (2002), we define the class of distributions $(\vec{X}, D)$ with an (l, C)-smooth function mapping from features to mean demand and the notion of the lower individual rate of convergence for an approach.

Definition A.1. *(Following Definition 3.3 in Györfi et al. 2002) A function $f : \mathbb{R}^p \to \mathbb{R}$ is called (l, C)-smooth with $C > 0$ if for every $\alpha = (\alpha_1, ..., \alpha_p)$, $\alpha_i \in \mathbb{N}_0$, $\sum_{i=1}^{p} \alpha_i = k$ with some $k \in \mathbb{N}_0$, such that $l = k + \beta$ with $0 < \beta \leq 1$, the partial derivative $\frac{\partial^k f}{\partial x_1^{\alpha_1} ... \partial x_p^{\alpha_p}}$ exists and satisfies*

$$\left| \frac{\partial^k f}{\partial x_1^{\alpha_1} ... \partial x_p^{\alpha_p}}(\vec{x}) - \frac{\partial^k f}{\partial x_1^{\alpha_1} ... \partial x_p^{\alpha_p}}(\vec{z}) \right| \leq C||\vec{x} - \vec{z}||^{\beta} \ \forall \vec{x}, \vec{z} \in \mathbb{R}^p. \tag{A.1}$$

Definition A.2. *(Following Definition 3.4 in Györfi et al. 2002) Let $\mathcal{P}^{(l,C)}$ denote the class of distributions $(\vec{X}, D)$ with $\vec{X} \sim \mathcal{U}[0, 1]^p$ and $D = m(\vec{X}) + \mathcal{E}_{std}$, where $\mathcal{E}_{std} \sim \mathcal{N}(0, 1)$ is noise independent of $\vec{X}$, and $m(\vec{X})$ is an (l, C)-smooth function.*

Definition A.3. *(Following Definition 3.5 in Györfi et al. 2002) A sequence $a_N \geq 0$ is called an individual lower rate of convergence of q^{wSAA} for a class $\mathcal{P}$ of distributions if*

$$\inf_{\{w_n\}} \sup_{(\vec{X},D)\in\mathcal{P}} \limsup_{N\to\infty} \frac{\mathbb{E}_{S_N}[||q^{wSAA} - q^*||^2]}{a_N} > 0 \tag{A.2}$$

with

$$||q^{wSAA} - q^*||^2 = \mathbb{E}_{\vec{X}} \left[\left(q^{wSAA}(\vec{X}) - q^*(\vec{X}) \right)^2 \right].$$

Definition A.3 introduces the concept of an individual lower rate of convergence, which is the fastest rate with which an approach can converge to the optimal solution over all possible joint distributions of $\vec{X} \times D$ within the class of distributions $\mathcal{P}$. While such a rate a_N does not prevent an approach to converge faster to the optimal solution for a fixed distribution of $\vec{X} \times D$, it states that for each approach there is at least one distribution within $\mathcal{P}$ such that the approach does converge faster than a_N.

Theorem 3.3 in Györfi et al. (2002) states that an individual lower rate of convergence of any approach q_N for the class $\mathcal{P}^{(l,C)}$ is given as $b_N N^{-\frac{2l}{2l+p}}$ for

an arbitrary positive sequence b_N with $\lim_{N\to\infty} b_N = 0$:

$$\inf_{\{q_N\}} \sup_{(\vec{X},D)\in\mathcal{P}^{(l,C)}} \limsup_{N\to\infty} \frac{\mathbb{E}_{S_N}[||q_N - m||^2]}{b_N N^{-\frac{2l}{2l+p}}} > 0. \tag{A.3}$$

We need to prove that $m(\vec{x}) = q^*(\vec{x})$, then we define $b_N = N^{-\frac{1}{2l+p}}$ and choose the subset $\{q^{\text{wSAA}}\} \subseteq \{q_N\}$, with q^{wSAA} being parameterized through $\{w_n\}$, to obtain the expression to be proven.
Because the loss function $L(q,d) = |q-d|^2$ is strictly convex in q, its minimum is unique at $q = d$ and the optimal prescription function is given by

$$q^*(\vec{x}) = \arg\min_{q\in\mathcal{Q}} \mathbb{E}[L(q,D)|\vec{X} = \vec{x}] = \mathbb{E}[D|\vec{X} = \vec{x}]. \tag{A.4}$$

By definition, $\mathbb{E}[D|\vec{X} = \vec{x}] = m(\vec{x})$, therefore we obtain

$$\inf_{\{w_n\}} \sup_{(\vec{X},D)\in\mathcal{P}^{(l,C)}} \limsup_{N\to\infty} \frac{\mathbb{E}_{S_N}[||q^{\text{wSAA}} - q^*||^2]}{N^{-\frac{2l+1}{2l+p}}} > 0, \tag{A.5}$$

which concludes the proof.

Proof of Proposition 2.5

The expression $||\vec{q}(\cdot)||^2_{\mathcal{F}}$ is jointly convex in $\vec{q}(\cdot)$, because the norm is by definition convex and positive, and $(\cdot)^2$ is non-decreasing for positive numbers. Therefore, as the loss function is convex (Proposition 2.2), the objective function of the ERM approach is a weighted sum of convex functions with positive weights $\lambda, 1/N \geq 0$, and therefore jointly convex in $\vec{q}(\cdot)$.

Proof of Theorem 2.1

The derivation of the kernelized ERM approach is based on the ERM solution for a linear function space (see Appendix A.7.4). This linear solution is *kernelized* by mapping the feature vectors $\vec{x}$ into $\mathcal{H}_K$ and showing that the scalar product between the mapped feature vectors can be expressed using the kernel function K.

Let $\mathbb{R}^{\mathcal{X}} = \{f(\cdot) : \mathcal{X} \to \mathbb{R}\}$ be a space of functions and let $\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ be a feature map with $\Phi(\vec{x})(\cdot) = K(\cdot, \vec{x})$, then one can construct a vector space $\mathcal{V}_K \supseteq \{\Phi(\vec{x})(\cdot) : \vec{x} \in \mathcal{X}\}$ and define a scalar product such that (see Section 2.2.2 in Schölkopf and Smola 2002):

$$\begin{aligned} \langle \Phi(\vec{x}_1), \Phi(\vec{x}_2) \rangle &= \langle K(\cdot, \vec{x}_1), K(\cdot, \vec{x}_2) \rangle \\ &= K(\vec{x}_1, \vec{x}_2). \end{aligned} \tag{A.6}$$

The vector space $\mathcal{V}_K$ can be turned into a Hilbert space $\mathcal{H}_K$ by defining a norm based on the scalar product, as presented in Section 2.2.3 in Schölkopf and Smola (2002).

We kernelize the linear solution to the capacity planning problem (Theorem A.2) by projecting each feature vector $\vec{x}$ into $\mathcal{H}_K$ using the feature map $\Phi(\vec{x})$. As both the dual Lagrangian L_{dual} (A.81) and the solution function (A.80) only depend on the scalar product of feature vectors, the feature map effectively replaces these scalar products by the kernel function as

$$\begin{aligned} \langle \vec{x}^p, \vec{x}^q \rangle \to \langle \Phi(\vec{x}^p), \Phi(\vec{x}^q) \rangle \\ = \; K(\vec{x}^p, \vec{x}^q), \end{aligned} \tag{A.7}$$

from which we obtain the kernelized solution. This application of the implicit feature map is usually referred to as "kernel trick" (see Remark 2.8 in Schölkopf and Smola 2002, p. 34).

Proof of Corollary 2.1

The primal Lagrangian L_{primal} with feature vectors $\vec{x}$ mapped into $\mathcal{H}_K$ as $\Phi(\vec{x})$ is by definition affine and therefore concave in the Lagrangian multipliers α_i^{tn}, β_j^{tn}, ϵ_j^n. The dual Lagrangian is the point-wise infimum of a collection of concave functions:

$$L_{\text{dual}}(\{\alpha_i^{tn}\}, \{\beta_j^{tn}\}, \{\epsilon_j^n\}) = \inf_{\mathbf{W}, \vec{b}, \{y_{ij}^{tn}\}} L_{\text{primal}}(\mathbf{W}, \vec{b}, \{y_{ij}^{tn}\}, \{\alpha_i^{tn}\}, \{\beta_j^{tn}\}, \{\epsilon_j^n\}), \tag{A.8}$$

and therefore concave in α_i^{tn}, β_j^{tn}, ϵ_j^n.

Proof of Proposition 2.6

To prove that K^{RF} defines a reproducing kernel Hilbert space, we first show symmetry and positive semi-definiteness of K^{RF}, and then apply the Moore-Aronszajn Theorem. Let

$$k_{\mathcal{R}^l}(\vec{x}_1, \vec{x}_2) := \mathbb{1}[\mathcal{R}^l(\vec{x}_1) = \mathcal{R}^l(\vec{x}_2)] \tag{A.9}$$

indicate if $\vec{x}_1$ and $\vec{x}_2$ are assigned to the same terminal node of tree l. Then, by definition, $k_{\mathcal{R}^l}(\vec{x}_1, \vec{x}_2)$ is symmetric, and also positive semi-definite as shown in the proof of Lemma 3.1 in Davies and Ghahramani (2014). Therefore,

$$K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2) = \sum_{l=1}^{L} \frac{1}{L \sum_{j=1}^{N} \mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_j)]} k_{\mathcal{R}^l}(\vec{x}_1, \vec{x}_2) \tag{A.10}$$

is a weighted sum of positive semi-definite functions with non-negative weights, and therefore also positive semi-definite (see Observation 7.1.3 in Horn and Johnson 2013).
Because $\forall \vec{x}_1, \vec{x}_2, k_{\mathcal{R}^l}(\vec{x}_1, \vec{x}_2) > 0 : \; \mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_j)] = \mathbb{1}[\mathcal{R}^l(\vec{x_2}) = \mathcal{R}^l(\vec{x}_j)] \; \forall j$ the random forest kernel $K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2)$ is symmetric, and therefore a kernel by Definition 2.1.
Based on the Moore-Aronszajn Theorem (see Part I Section 2 (4) in Aronszajn 1950, p. 344), which states that "to every positive matrix $K(x, y)$ there corresponds one and only one class of functions [...], forming a Hilbert space and admitting $K(x, y)$ as a reproducing kernel", we conclude that $K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2)$ is a reproducing kernel with reproducing kernel Hilbert space $\mathcal{H}_{K^{\mathrm{RF}}}$.

Proof of Lemma 2.1

To prove that $L(\vec{q}, \boldsymbol{d})$ is bounded (part a of Lemma 2.1), we first derive bounds on $\pi(\vec{d}, \vec{q})$, which then allows to bound $\Pi(\vec{q}, \boldsymbol{d})$ and $\Pi^*(\boldsymbol{d})$.
Because $a_{ii} \geq a_{ij} \; \forall j < i$ (see Section 2.3), we obtain

$$\pi(\vec{d}, \vec{q}) = \max_{\{y_{ij}\}} \sum_{i,j} a_{ij} y_{ij} - \sum_i c_i d_i \leq \sum_i a_{ii} d_i. \tag{A.11}$$

Because $\max_{\{y_{ij}\}} \sum_{i,j} a_{ij} y_{ij} \geq 0$, as $y_{ij} = 0$ is always a feasible solution, we derive as lower bound:

$$\pi(\vec{d}, \vec{q}) = \max_{\{y_{ij}\}} \sum_{i,j} a_{ij} y_{ij} - \sum_i c_i d_i \geq -\sum_i c_i d_i. \tag{A.12}$$

Using these results, we obtain bounds for the optimal profit:

$$\Pi^*(\boldsymbol{d}) = \max_{\vec{q}} \left(\sum_{t=1}^{T} \pi(\vec{d^t}, \vec{q}) - \sum_j f_j q_j \right) \leq T \bar{d} \sum_i a_{ii}, \tag{A.13}$$

and for the profit:

$$\Pi(\vec{q}, \boldsymbol{d}) = \sum_{t=1}^{T} \pi(\vec{d^t}, \vec{q}) - \sum_j f_j q_j \geq -T \bar{d} \sum_i c_i - \bar{q} \sum_j f_j. \tag{A.14}$$

Therefore, we obtain as bound on the loss function:

$$\begin{aligned} L(\vec{q}, \boldsymbol{d}) &= \Pi^*(\boldsymbol{d}) - \Pi(\vec{q}, \boldsymbol{d}) \\ &\leq T \bar{d} \sum_i (a_{ii} + c_i) + \bar{q} \sum_j f_j =: \bar{l} < \infty. \end{aligned} \tag{A.15}$$

To show that the loss function is equi-Lipschitz (part b of Lemma 2.1), we first derive a bound on the change in profit π caused by an increase in capacity $\vec{q}$, and then show that this bounds the change in loss. To bound the change in π, we consider, in a first step, a change in capacity q_j for a single service line j, while keeping all other capacities constant. In a second step we generalize the bound to arbitrary changes in capacity.

Assume $q'_j \geq q_j$, without loss of generality, so that the capacity increase from q_j to q'_j relaxes the constraint $\sum_i y_{ij} \leq q_j$ on stage 2 of the planning problem. Therefore, the allocation y_{ij} that yields $\pi(\vec{d^t}, \vec{q})$ is also a feasible solution to $\pi(\vec{d^t}, \vec{q'}) = \max_{y'_{ij}} \sum_{ij} a_{ij} y'_{ij} - \sum_i c_i d_i^t$, hence

$$\pi(\vec{d^t}, \vec{q}) \leq \pi(\vec{d^t}, \vec{q'}). \tag{A.16}$$

Because $a_{ii} \geq a_{ij}$ $\forall j < i$ (see Section 2.3), the maximum achievable profit for

the additional capacity $q'_j - q_j$ is given as $a_{jj}(q'_j - q_j)$, which yields

$$|\pi(\vec{d^t}, \vec{q'}) - \pi(\vec{d^t}, \vec{q})| \leq a_{jj}|q'_j - q_j|, \tag{A.17}$$

which also holds for $q'_j \leq q_j$, as $\vec{q}$ and $\vec{q'}$ are interchangeable in (A.17). Because $\vec{q}$ and $\vec{q'}$ may differ in several of the I dimensions, we generalize the bound by defining a finite sequence $\vec{q_l}$ with $l = 0..I$ and $\vec{q_0} = \vec{q}$, $\vec{q_I} = \vec{q'}$, such that $\vec{q_l}$ and $\vec{q_{l-1}}$ only differ in dimension l. Using this sequence, we bound the change in π as

$$\begin{aligned} |\pi(\vec{d^t}, \vec{q'}) - \pi(\vec{d^t}, \vec{q})| &= \left| \sum_{l=1}^{I} \left(\pi(\vec{d^t}, \vec{q_l}) - \pi(\vec{d^t}, \vec{q_{l-1}}) \right) \right| \\ &\leq \sum_{l=1}^{I} a_{ll}|q'_l - q_l| \leq I a_{\max} ||\vec{q'} - \vec{q}||_\infty, \end{aligned} \tag{A.18}$$

where $a_{\max} := \max_j a_{jj}$.

Based on this bound, we show that $L(\vec{q}, \boldsymbol{d})$ is equi-Lipschitz:

$$\begin{aligned} |L(\vec{q'}, \boldsymbol{d}) - L(\vec{q}, \boldsymbol{d})| &= \left| \sum_{t=1}^{T} \left(\pi(\vec{d^t}, \vec{q'}) - \pi(\vec{d^t}, \vec{q}) \right) - \sum_j f_j \left(q'_j - q_j \right) \right| \\ &\leq \sum_{t=1}^{T} \left| \pi(\vec{d^t}, \vec{q'}) - \pi(\vec{d^t}, \vec{q}) \right| + \sum_j f_j \left| q'_j - q_j \right| \\ &\leq \left(TIa_{\max} + \sum_j f_j \right) ||\vec{q'} - \vec{q}||_\infty \\ &=: M_{Lip} ||\vec{q'} - \vec{q}||_\infty, \end{aligned} \tag{A.19}$$

which concludes the proof.

Proof of Theorem 2.2

To prove the stated out-of-sample performance guarantee, we use a general out-of-sample performance guarantee provided by BK and bound the Rademacher complexity. To be able to bound the Rademacher complexity, we assume a bounded function space $\mathcal{F}$ where $||\vec{b}||_\infty \leq B_C$ and $||q_{U,j}||_K \leq B_U \ \forall j$. Such

a bound is necessary for learning (e.g., an unbounded RKHS would lead to overfitting and the prescription function would not generalize), and standard in statistical learning when deriving out-of-sample performance bounds, see, e.g., Bartlett and Mendelson (2002).

Because out-of-sample performance guarantees depend on the richness of the function space $\mathcal{F}$ over which the kERM approach is solved, we use the empirical Rademacher complexity of $\mathcal{F}$, given S_N, to quantify this richness.

Definition A.4. *(Following Shalev-Shwartz and Ben-David 2014) A random variable $\sigma \in \{\pm 1\}$ is called Rademacher random variable if*

$$P(\sigma = +1) = P(\sigma = -1) = 1/2. \tag{A.20}$$

Definition A.5. *(BK) The empirical multivariate Rademacher complexity for $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}^I\}$ over a training data set S_N is defined as*

$$\mathrm{Rad}_N(\mathcal{F}, S_N) = \mathbb{E}\left(\frac{2}{N} \sup_{\vec{q} \in \mathcal{F}} \sum_{n=1}^{N} \sum_{j=1}^{I} \sigma_{jn} q_j(\vec{x}^n) \middle| \vec{x}^1, ..., \vec{x}^N \right), \tag{A.21}$$

where σ_{jn} are independent Rademacher variables.

Theorem A.1. *(Following BK) Assume S_N, generated by iid sampling from a joint distribution of $\vec{X} \times \boldsymbol{D}$, $L(\vec{q}, \boldsymbol{d})$, as defined in (2.2), and let $\delta > 0$. Then, with probability of at least $1 - \delta$ for any function $\vec{q}(\cdot) \in \mathcal{F}$, the true risk is bounded as*

$$R(\vec{q}(\cdot)) \leq R_N(\vec{q}(\cdot)) + 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} + M_{Lip}\,\mathrm{Rad}_N(\mathcal{F}, S_N), \tag{A.22}$$

where $\bar{l}$ is the bound and M_{Lip} is the Lipschitz constant of $L(\vec{q}, \boldsymbol{d})$.

Proof of Theorem A.1: Because $L(\vec{q}, \boldsymbol{d})$ is bounded and equi-Lipschitz (Lemma 2.1), the results of Theorem 8 in BK apply, which proves the stated out-of-sample performance guarantee.

Theorem A.1 bounds the true risk by expressions that can be evaluated using the data set S_N and the function space $\mathcal{F}$, so (A.22) can be considered an out-of-sample performance guarantee.

In the following we derive bounds on the Rademacher complexity for the function spaces used by kERM.

Definition A.6. *Let the sum of two function spaces be the element-wise sum:*

$$\mathcal{F} + \mathcal{G} := \{(f+g)(\cdot) = f(\cdot) + g(\cdot) | f \in \mathcal{F}, g \in \mathcal{G}\}. \tag{A.23}$$

Lemma A.1. *Assume function spaces $\mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$. Then the empirical multivariate Rademacher complexity over S_N of $\mathcal{F}$ is given as:*

$$\mathrm{Rad}_N(\mathcal{F}, S_N) = \mathrm{Rad}_N(\mathcal{F}_\mathcal{U}, S_N) + \mathrm{Rad}_N(\mathcal{F}_\mathcal{C}, S_N). \tag{A.24}$$

Proof of Lemma A.1: Using Definition A.5 of the empirical multivariate Rademacher complexity and Definition A.6 of the sum of function spaces, we obtain

$$\begin{aligned}
\mathrm{Rad}_N(\mathcal{F}, S_N) &= \mathrm{Rad}_N(\mathcal{F}_U + \mathcal{F}_C, S_N) \\
&= \mathbb{E}\left(\frac{2}{N} \sup_{\vec{f}^U \in \mathcal{F}_U, \vec{f}^C \in \mathcal{F}_C} \sum_{n=1}^{N} \sum_{j=1}^{I} \sigma_{jn} \left(f_j^U(\vec{x}^n) + f_j^C(\vec{x}^n) \right) \middle| \vec{x}^1, ..., \vec{x}^N \right) \\
&= \mathbb{E}\left(\frac{2}{N} \sup_{\vec{f}^U \in \mathcal{F}_U} \sum_{n=1}^{N} \sum_{j=1}^{I} \sigma_{jn} f_j^U(\vec{x}^n) \middle| \vec{x}^1, ..., \vec{x}^N \right) \\
&\quad + \mathbb{E}\left(\frac{2}{N} \sup_{\vec{f}^C \in \mathcal{F}_C} \sum_{n=1}^{N} \sum_{j=1}^{I} \sigma_{jn} f_j^C(\vec{x}^n) \middle| \vec{x}^1, ..., \vec{x}^N \right) \\
&= \mathrm{Rad}_N(\mathcal{F}_U, S_N) + \mathrm{Rad}_N(\mathcal{F}_C, S_N),
\end{aligned} \tag{A.25}$$

which concludes the proof of Lemma A.1.

The result of Lemma A.1 allows us to separate the Rademacher complexity of the function space $\mathcal{F}$ (defined in Section 2.5.1) into the complexities of $\mathcal{F}_U$ and $\mathcal{F}_C$, for which we present bounds in the following.

Lemma A.2. *The empirical Rademacher complexity of $\mathcal{F}_C$ inside a ball of radius B_C, with $||\vec{b}||_\infty \leq B_C$, is bounded as*

$$\mathrm{Rad}_N(\mathcal{F}_C, S_N) \leq \frac{2\sqrt{2} I B_C e}{\sqrt{\pi}\sqrt{N}} \tag{A.26}$$

with Euler constant e.

Proof of Lemma A.2: To bound the empirical multivariate Rademacher complexity of $\mathcal{F}_C$, we use Definition A.5 to obtain:

$$\begin{aligned}\mathrm{Rad}_N(\mathcal{F}_C, S_N) &= \mathbb{E}\left(\frac{2}{N}\sup_{\vec{q}_C\in\mathcal{Q}_C}\sum_{n=1}^{N}\sum_{j=1}^{I}\sigma_{jn}q_{C,j}(\vec{x}^n)\middle|\vec{x}^1,...,\vec{x}^N\right)\\ &\leq \frac{2}{N}\sum_{j=1}^{I}\mathbb{E}\left(\sup_{|b_j|\leq B_C}(-b_j)\sum_{n=1}^{N}\sigma_{jn}\right)\\ &= \frac{2I}{N}\mathbb{E}\left(\sup_{|b_0|\leq B_C}(-b_0)\sum_{n=1}^{N}\sigma_{0,n}\right).\end{aligned} \tag{A.27}$$

Because $\sum_{n=1}^{N}\sigma_{0,n}$ follows a binomial distribution, we bound the Rademacher complexity as:

$$\begin{aligned}\mathrm{Rad}_N(\mathcal{F}_C, S_N) &\leq \frac{2I}{N2^N}\sum_{k=0}^{N}\sup_{|b_0|\leq B_C}(-b_0)\binom{N}{k}\Big[(+1)\cdot k+(-1)\cdot(N-k)\Big]\\ &= \frac{2IB_C}{N2^N}\sum_{k=0}^{N}\binom{N}{k}|2k-N|.\end{aligned} \tag{A.28}$$

To further bound $\sum_{k=0}^{N}\binom{N}{k}|2k-N|$, we assume N even, without loss of generality:

$$\begin{aligned}\sum_{k=0}^{N}\binom{N}{k}|2k-N| &= \sum_{k=0}^{N/2}\binom{N}{k}(N-2k)+\sum_{k=N/2}^{N}\binom{N}{k}(2k-N)\\ &= -\sum_{k=0}^{N/2}\binom{N}{k}2k+\sum_{k=N/2}^{N}\binom{N}{k}2k\\ &\quad+\sum_{k=0}^{N/2}\binom{N}{k}N-\sum_{k=N/2}^{N}\binom{N}{k}N\\ &= 2\left(-\sum_{k=1}^{N/2}\binom{N}{k}k+\sum_{k=N/2}^{N}\binom{N}{k}k\right).\end{aligned} \tag{A.29}$$

Using $\binom{N}{k}k = \binom{N-1}{k-1}N$, we obtain

$$\begin{aligned}\sum_{k=0}^{N}\binom{N}{k}|2k-N| &= 2N\left(-\sum_{k=1}^{N/2}\binom{N-1}{k-1}+\sum_{k=N/2}^{N}\binom{N-1}{k-1}\right)\\ &= 2N\left(-\sum_{k=0}^{N/2-1}\binom{N-1}{k}+\sum_{k=N/2}^{N-1}\binom{N-1}{k}+\binom{N-1}{N/2-1}\right)\\ &= N\binom{N}{N/2} \leq N\frac{e}{\sqrt{2\pi}\sqrt{N}}2^{N+1},\end{aligned} \tag{A.30}$$

where we used the bounds on $N!$ for all $N \in \mathbb{N}$ based on Stirling's formula as presented in Robbins (1955):

$$\sqrt{2\pi}N^{N+1/2}e^{-N}e^{1/(12N+1)} \leq N! \leq \sqrt{2\pi}N^{N+1/2}e^{-N}e^{1/(12N)}. \tag{A.31}$$

The Rademacher complexity of $\mathcal{F}_C$ is therefore bounded as

$$\mathrm{Rad}_N(\mathcal{F}_C, S_N) \leq \frac{2\sqrt{2}IB_Ce}{\sqrt{\pi}\sqrt{N}}, \tag{A.32}$$

which concludes the proof of Lemma A.2.

For a data-independent kernel $K(\vec{x}_1, \vec{x}_2)$ we can bound the Rademacher complexity of $\mathcal{F}_U$ as presented in Lemma A.3.

Lemma A.3. *The empirical Rademacher complexity of $\mathcal{F}_U$ with all components $q_{U,j}$ inside a ball of radius B_U, with $||q_{U,j}||_K \leq B_U$, is bounded as*

$$\mathrm{Rad}_N(\mathcal{F}_U, S_N) \leq \frac{2IB_U}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{n=1}^{N}K(\vec{x}^n, \vec{x}^n)}. \tag{A.33}$$

Proof of Lemma A.3 : To bound the empirical multivariate Rademacher

complexity of $\mathcal{F}_U$, we use Definition A.5 to obtain:

$$\begin{aligned}\mathrm{Rad}_N(\mathcal{F}_U, S_N) &= \mathbb{E}\left(\frac{2}{N}\sup_{\vec{q}_U\in\mathcal{F}_U:||q_{U,j}||_K\leq B_U}\sum_{n=1}^{N}\sum_{j=1}^{I}\sigma_{jn}q_{U,j}(\vec{x}^n)\middle|\vec{x}^1,...,\vec{x}^N\right)\\ &= \mathbb{E}\left(I\frac{2}{N}\sup_{q_{U,1}\in\mathcal{H}_K:||q_{U,1}||_K\leq B_U}\sum_{n=1}^{N}\sigma_{1,n}q_{U,1}(\vec{x}^n)\middle|\vec{x}^1,...,\vec{x}^N\right)\\ &= I\,\mathrm{Rad}_N(\mathcal{F}_{U,1}, S_N),\end{aligned} \tag{A.34}$$

where

$$\mathcal{F}_{U,1} = \{q_{U,1}(\cdot) : \mathcal{X}\to\mathbb{R} : q_{U,1}(\cdot)\in\mathcal{H}_K,\ ||q_{U,1}||_K \leq B_U\}. \tag{A.35}$$

The result of Lemma 22 in Bartlett and Mendelson (2002) allows us to bound the Rademacher complexity of this ball of radius B_U in the reproducing kernel Hilbert space as

$$\mathrm{Rad}_N(\mathcal{F}_{U,1}, S_N) \leq \frac{2B_U}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{n=1}^{N}K(\vec{x}^n,\vec{x}^n)}, \tag{A.36}$$

and the bound to be proven for Lemma A.3 directly follows.

We combine the general out-of-sample guarantee of Theorem A.1 with the results for the empirical Rademacher complexity presented in Lemmas A.1, A.2 and A.3 to obtain the out-of-sample guarantee stated in Theorem 2.2, which concludes the proof.

Proof of Proposition 2.7

To prove that the stated function space $\mathcal{F}_{K_{\mathrm{RBFG}}}$ is dense in $C(\mathcal{X},\mathbb{R}^I)$, we first establish universality of the RBF Gauss kernel (following Steinwart and Christmann 2008), then show that $\mathcal{F}_{K_{\mathrm{RBFG}}}$ is equivalent to the reproducing kernel Hilbert space of a vector-valued kernel and use the results of Theorem 12 in Caponnetto et al. (2008) to prove universality.

Definition A.7. *(Following Steinwart and Christmann 2008 and Caponnetto et al. 2008) A continuous kernel $K(\vec{x}_1,\vec{x}_2)$ on a compact space $\mathcal{X}$ is called*

universal if the corresponding reproducing kernel Hilbert space $\mathcal{H}_K$ is dense in the space $C(\mathcal{X}, \mathbb{R})$ of all continuous, real-valued functions over $\mathcal{X}$, that is

$$\forall g \in C(\mathcal{X}, \mathbb{R}), \epsilon > 0 \; \exists f \in \mathcal{H}_K : \; ||f - g||_\infty \leq \epsilon. \tag{A.37}$$

In the following lemma we introduce the most popular universal kernel, the Gaussian RBF kernel.

Lemma A.4. *(Following Steinwart and Christmann 2008) The RBF Gauss kernel*

$$K_{RBFG}(\vec{x}_1, \vec{x}_2) := \exp\left(-\gamma^2 |\vec{x}_1 - \vec{x}_2|^2\right) \tag{A.38}$$

is universal for any $\gamma > 0$, $\vec{x}_1, \vec{x}_2 \in \mathcal{X}$ and compact $\mathcal{X} \subset \mathbb{R}^p$, and bounded as $K_{RBFG}(\vec{x}_1, \vec{x}_2) \leq 1$.

Proof of Lemma A.4: The universality of the RBF Gauss kernel has been shown in Micchelli et al. (2006) and in Corollary 4.58 in Steinwart and Christmann (2008). The bound on the kernel function $K_{\text{RBFG}}(\vec{x}_1, \vec{x}_2) \leq 1$ directly follows from $-\gamma^2 |\vec{x}_1 - \vec{x}_2|^2 \leq 0 \; \forall \vec{x}_1, \vec{x}_2$.

The function space

$$\mathcal{F}_0 = \left\{ \vec{q}(\cdot) : \mathcal{X} \to \mathbb{R}^I : q_j(\vec{x}) = q_{\mathcal{H},j}(\vec{x}) = \sum_{n=1}^{\infty} u_j^n K\left(\vec{x}^n, \vec{x}\right) \right\} \tag{A.39}$$

is equivalent to the reproducing kernel Hilbert space corresponding to the vector-valued kernel $\boldsymbol{K}(\vec{x}_1, \vec{x}_2) := K(\vec{x}_1, \vec{x}_2)\boldsymbol{I}$ with $\boldsymbol{I}$ the identity matrix (see in Álvarez et al. 2012):

$$\begin{aligned} \mathcal{F}_1 = \Big\{ \vec{q}(\cdot) : \mathcal{X} \to \mathbb{R}^I : \vec{q}(\vec{x}) &= \sum_{n=1}^{\infty} \boldsymbol{K}\left(\vec{x}^n, \vec{x}\right) \vec{u}^n = \sum_{n=1}^{\infty} K\left(\vec{x}^n, \vec{x}\right) \boldsymbol{I} \vec{u}^n \\ &= \sum_{n=1}^{\infty} K\left(\vec{x}^n, \vec{x}\right) \vec{u}^n \Big\}. \end{aligned} \tag{A.40}$$

Because the identity matrix $\boldsymbol{I}$ is positive definite, and the scalar-valued RBF Gauss kernel is universal (Lemma A.4), the assumptions of Theorem 12 in Caponnetto et al. (2008) are fulfilled, the vector-valued kernel is universal and its reproducing kernel Hilbert space is dense in $C(\mathcal{X}, \mathbb{R}^I)$.

Because the function space $\mathcal{F}_{K_{\text{RBFG}}}$ equals the reproducing kernel Hilbert space of $K(\vec{x}_1, \vec{x}_2)\boldsymbol{I}$ when setting $b_j = 0$, it is also dense in $C(\mathcal{X}, \mathbb{R}^I)$, which concludes the proof.

Proof of Proposition 2.8

To prove the convergence of kERM, we show that the deviation in risk from the optimal prescription function $\vec{q}^*_{\mathcal{F}}(\cdot)$ within each function space $\mathcal{F}$ approaches zero for $N \to \infty$ and that the function space $\mathcal{F}$ equals $\mathcal{F}_{K_{\text{RBFG}}}$ in the limit of $N \to \infty$. Similar as in Theorem 2.2, we assume a bounded function space $\mathcal{F}$ where $||\vec{b}||_\infty \leq B_{C,N}$ and $||q_{U,j}||_K \leq B_{U,N}$ $\forall j$, which allows us to bound the Rademacher complexity (see Proof of Theorem 2.2 for details). To show the universal approximation property, however, we need to expand these bounds slowly with the number of data samples N, such that a) in the limit of $N \to \infty$ the function space is unbounded ($B_{U,N}, B_{C,N} = \infty$) and therefore dense in $C(\mathcal{X}, \mathbb{R}^I)$, and b) our out-of-sample performance guarantees, which contain $B_{U,N}/\sqrt{N}$ and $B_{C,N}/\sqrt{N}$, still converge ($\lim_{N\to\infty} B_{U,N}/\sqrt{N}, B_{C,N}/\sqrt{N} = 0$). One example for such sequences is $B_{U,N} = B_{C,N} = (N)^{\frac{1}{3}}$.

The RBF Gauss kernel is data-independent and bounded (Lemma A.4) as $K_{\text{RBFG}}(\vec{x}_1, \vec{x}_2) \leq 1$, therefore we obtain an out-of-sample performance guarantee (using Theorem 2.2) as

$$R(\vec{q}(\cdot)) \leq R_N(\vec{q}(\cdot)) + 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} + M_{Lip}\left(\frac{2\sqrt{2}IB_{C,N}e}{\sqrt{\pi}\sqrt{N}} + \frac{2IB_{U,N}}{\sqrt{N}}\right). \quad \text{(A.41)}$$

Let $\Delta_N(\vec{q}^{\text{kERM}}_{\mathcal{F}}(\cdot))$ denote the deviation of $\vec{q}^{\text{kERM}}_{\mathcal{F}}(\cdot)$ from the optimal solution $\vec{q}^*_{\mathcal{F}}(\cdot)$ within function space $\mathcal{F}$ in risk such that

$$\Delta_N(\vec{q}^{\text{kERM}}_{\mathcal{F}}(\cdot)) := R(\vec{q}^{\text{kERM}}_{\mathcal{F}}(\cdot)) - R(\vec{q}^*_{\mathcal{F}}(\cdot)). \quad \text{(A.42)}$$

Because $R_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) = \min_{\vec{q}(\cdot)\in\mathcal{F}} R_N(\vec{q}(\cdot)) \leq R_N(\vec{q}_{\mathcal{F}}^{*}(\cdot))$, we obtain

$$\begin{aligned}\Delta_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) &= R(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) - R_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) + R_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) - R(\vec{q}_{\mathcal{F}}^{*}(\cdot)) \\ &\leq \left[R(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) - R_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot))\right] + \left[R_N(\vec{q}_{\mathcal{F}}^{*}(\cdot)) - R(\vec{q}_{\mathcal{F}}^{*}(\cdot))\right].\end{aligned} \tag{A.43}$$

While $R(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) - R_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot))$ is bounded by (A.41), we derive a bound on the term $R_N(\vec{q}_{\mathcal{F}}^{*}(\cdot)) - R(\vec{q}_{\mathcal{F}}^{*}(\cdot))$ in the following.
Because the loss function is bounded (Lemma 2.1), we can apply Hoeffding's inequality for $\epsilon > 0$ (similar as in Expressions 5.7 and 5.8 in Vapnik 1998, p. 186) and obtain:

$$P\left(R_N(\vec{q}_{\mathcal{F}}^{*}(\cdot)) - R(\vec{q}_{\mathcal{F}}^{*}(\cdot)) > \epsilon\right) \leq exp\left(\frac{-2\epsilon^2 N}{\bar{l}^2}\right), \tag{A.44}$$

from which we derive, by defining $\delta := exp\left(\frac{-2\epsilon^2 N}{\bar{l}^2}\right)$:

$$P\left(R_N(\vec{q}_{\mathcal{F}}^{*}(\cdot)) - R(\vec{q}_{\mathcal{F}}^{*}(\cdot)) \leq \bar{l}\sqrt{\frac{\log(1/\delta)}{2N}}\right) \geq 1-\delta. \tag{A.45}$$

Therefore, we obtain for the deviation in risk with probability of at least $1-2\delta$:

$$\begin{aligned}\Delta_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) &\leq 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} + M_{Lip}\left(\frac{2\sqrt{2}IB_{C,N}e}{\sqrt{\pi}\sqrt{N}} + \frac{2IB_{U,N}}{\sqrt{N}}\right) + \bar{l}\sqrt{\frac{\log(1/\delta)}{2N}} \\ &=: \frac{C_\delta}{\sqrt{N}} + \frac{C_2 B_{C,N}}{\sqrt{N}} + \frac{C_3 B_{U,N}}{\sqrt{N}},\end{aligned} \tag{A.46}$$

where C_δ, C_2, C_3 are constant with respect to N.

Because $\lim_{N\to\infty} B_{U,N}/\sqrt{N}, B_{C,N}/\sqrt{N} = 0$ by assumption, we obtain that

$$\lim_{N\to\infty} \Delta_N(\vec{q}_{\mathcal{F}}^{\text{kERM}}(\cdot)) = 0 \tag{A.47}$$

with probability of at least $1-2\delta$.

In addition, because $\lim_{N\to\infty} B_{U,N}, B_{C,N} = \infty$, we obtain for the function space

$$\lim_{N\to\infty} \mathcal{F} = \mathcal{F}_{K_{\text{RBFG}}}, \tag{A.48}$$

and because $\mathcal{F}_{K_{\text{RBFG}}}$ is dense in $C(\mathcal{X}, \mathbb{R}^I)$ (Proposition 2.7), we obtain that the risk of kERM converges in probability toward the risk of $\vec{q}^*(\cdot)$, which concludes the proof.

Proof of Proposition 2.9

The proof of Proposition 2.9 follows a similar structure as the proof of Proposition 2.4—we apply the general results of Theorem 3.3 in Györfi et al. (2002) to the stated loss function and the kERM approach. Similar to Proposition 2.4 we assume a class of distributions $\mathcal{P}^{(l,C)}$ where the relationship between features and mean demand is an (l, C)-smooth function (see Definition A.2 and Györfi et al. 2002 for details).

Theorem 3.3 in Györfi et al. (2002) states that an individual lower rate of convergence of any approach q_N for the class $\mathcal{P}^{(l,C)}$ is given as $b_N N^{-\frac{2l}{2l+p}}$ for an arbitrary positive sequence b_N with $\lim_{N\to\infty} b_N = 0$; we define $b_N = N^{-\frac{1}{2l+p}}$ to obtain:

$$\inf_{\{q_N\}} \sup_{(\vec{X},D)\in\mathcal{P}^{(l,C)}} \limsup_{N\to\infty} \frac{\mathbb{E}_{S_N}[||q_N - m||^2]}{N^{-\frac{2l+1}{2l+p}}} > 0. \tag{A.49}$$

Because $m(\vec{x}) = q^*(\vec{x})$ (Proposition 2.4), we obtain, for choosing the subset $\{q^{\text{kERM}}\} \subseteq \{q_N\}$, with q^{kERM} being parameterized through kernel $K(\cdot,\cdot)$:

$$\inf_{K(\cdot,\cdot)} \sup_{(\vec{X},D)\in\mathcal{P}^{(l,C)}} \limsup_{N\to\infty} \frac{\mathbb{E}_{S_N}[||q^{\text{kERM}} - q^*||^2]}{N^{-\frac{2l+1}{2l+p}}} > 0, \tag{A.50}$$

which concludes the proof.

Proof of Theorem 2.3

To prove the out-of-sample performance guarantee stated in Theorem 2.3, we show that the kernel function $K^{\text{RF}}(\vec{x}_1, \vec{x}_2)$ as defined in (2.13) is bounded for any data set $S_{N_{RF}}$ and any random forest trained on $S_{N_{RF}}$, and then apply the results of Theorem 2.2 for training kERM on the data set $S_{N-N_{RF}}$. Similar as in Theorem 2.2, we assume a bounded function space $\mathcal{F}$ where $||\vec{b}||_\infty \leq B_C$ and $||q_{U,j}||_K \leq B_U \ \forall j$.

Because every leaf node $\mathcal{R}^l$ of every tree l of the random forest contains

at least one data sample of $S_{N_{RF}}$, we obtain:

$$\sum_{j=1}^{N_{RF}} \mathbb{1}[\mathcal{R}^l(\vec{x}) = \mathcal{R}^l(\vec{x}_j)] \geq 1 \ \forall l, \vec{x} \in \mathcal{X}. \tag{A.51}$$

There we obtain, following (2.13):

$$\begin{aligned} K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2) &= \frac{1}{L} \sum_{l=1}^{L} \frac{\mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_2)]}{\sum_{j=1}^{N_{RF}} \mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_j)]} \\ &\leq \frac{1}{L} \sum_{l=1}^{L} \mathbb{1}[\mathcal{R}^l(\vec{x_1}) = \mathcal{R}^l(\vec{x}_2)] \\ &\leq 1. \end{aligned} \tag{A.52}$$

Applying this bound to the result of Theorem 2.2, we obtain the expression to be proven.

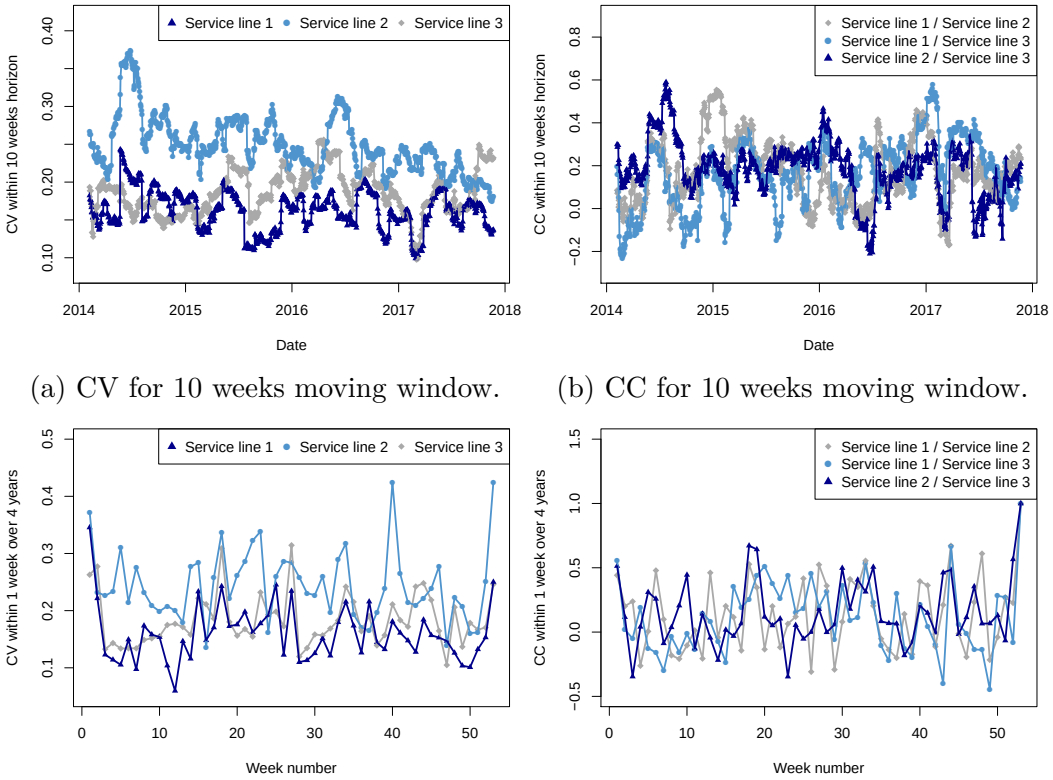
A.2 Characteristics of the Historical Demand Data

In this section we present a descriptive analysis of the historical demand data provided by the logistics service provider with whom we collaborated.

Figure A.1 illustrates the variations of CV and CC by plotting each for a moving time window of ten weeks (panels (a) and (b)) and for the same week in each year (e.g., for all observations in the first week in 2014-2017, see panels (c) and (d)). As stated in Section 2.6.1, we observe periods with higher and lower CVs and CCs, which may be explained by certain features related to the time series.

Because interviews with experts suggested that demand may contain complex seasonalities with different frequencies, we use a TBATS[52] model, which was introduced by De Livera et al. (2011) as an extension to the common BATS model for time series modelling with multiple seasonalities.

[52] The acronym TBATS stands for Trigonometric (for multiple seasonalities), Box-Cox transformation, ARMA errors, Trend, Seasonality.



(a) CV for 10 weeks moving window. (b) CC for 10 weeks moving window.

(c) CV for single week across 2014-2017. (d) CC for single week across 2014-2017.

Figure A.1: CV and CC of daily demand, estimated for 10 weeks moving window (top) or a single week (bottom).

The time series of daily demands is prepared such that all days (Monday to Friday) of the period 2014-2017 are single entries, resulting in a time series of 1043 entries. The seasonal periods are set to 5 (week), 21 (month), 65 (quarter year), 130 (half year) and 260 (year), and a TBATS model is fitted using the Akaike information criterion (AIC, see Akaike 1974) for model selection. Figure A.2 displays the components identified by the TBATS model, which shows strongest seasonality for quarter-yearly and yearly frequencies.

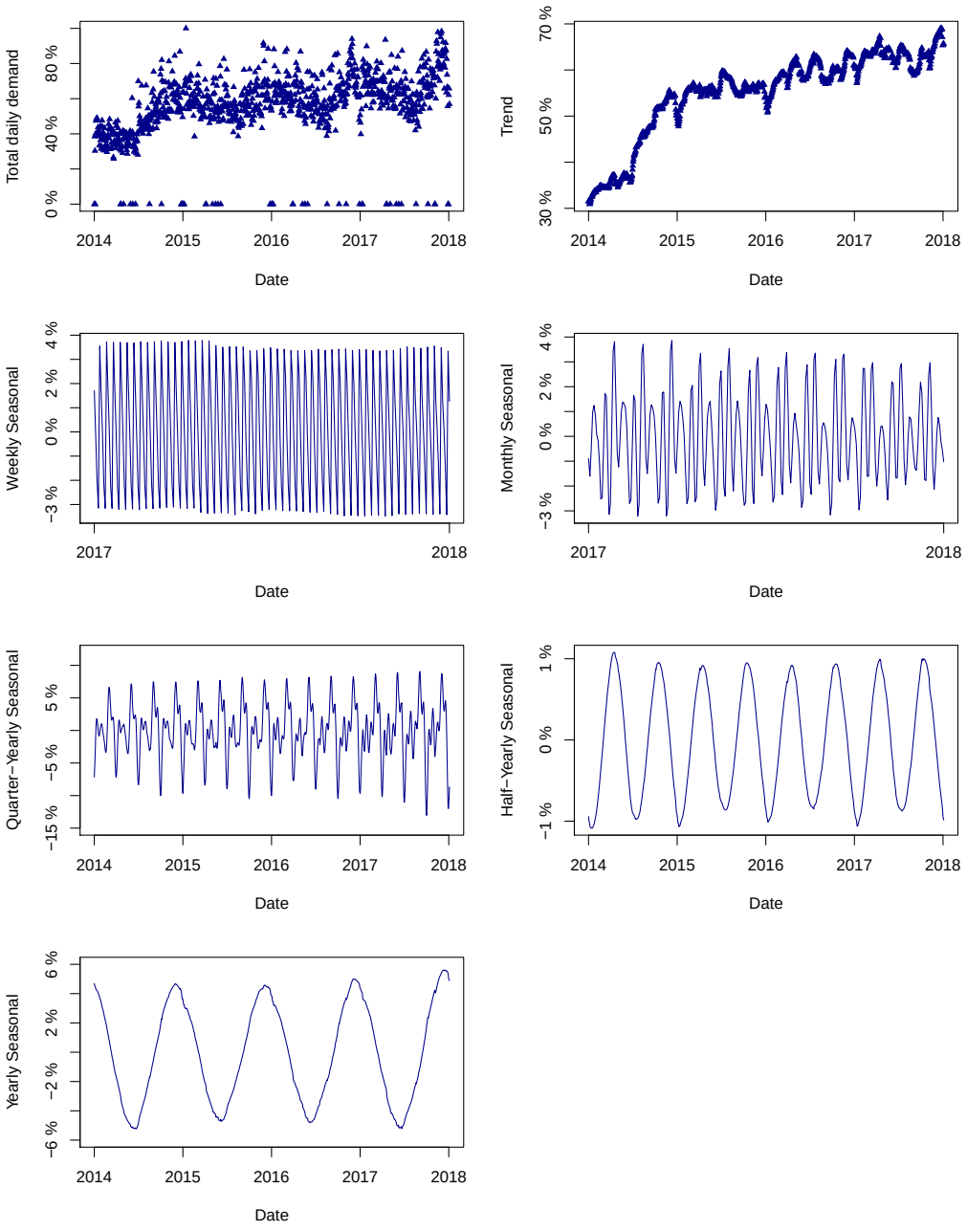


Figure A.2: Decomposition of total daily demand using a TBATS model. Note that the depicted time frame is reduced for weekly and monthly seasonal components due to the high frequency.

A.3 Details on the Features used for Numerical Experiments in Section 2.6

This section provides further details on the elements of the feature vector $\vec{x}^n \in \mathbb{R}^{162}$. As mentioned in Section 2.6.2, features are constructed in three groups, date-based, lag-demand-based, and public-holiday based. In the first group we construct features describing

- the year (2014-2017),
- the half of the year (1-2),
- the quarter of the year (1-4),
- the month of the year (1-12),
- the week number (1-53) by US and ISO standards,
- the week number within the month (1-5),
- the week number modulo 2, 3, or 4 (0-1, 0-2, 0-3), and
- the time, in terms of a continuously increasing index (number of seconds after the beginning of 1970, in the range 1388620800-1514505600).

All of these 11 date-based features are constructed using the *timetk* package in R.
In the second group, we construct 140 features representing lagged demand as

- 1-3 weeks lag, by service line and day,
- 1 month lag, by service line and day,
- 1 year lag, by service line and day,
- 1-3 weeks lag, summed across service lines, by day,
- 1 month lag, summed across service lines, by day,
- 1 year lag, summed across service lines, by day,

- 1-3 weeks lag, summed across days of the week, by service line,
- 1 month lag, summed across days of the week, by service line,
- 1 year lag, summed across days of the week, by service line,
- 1-3 weeks lag, summed across service lines and days of the week,
- 1 month lag, summed across service lines and days of the week,
- 1 year lag, summed across service lines and days of the week,
- sum of 1-2 weeks lag, summed across service lines and days of the week,
- sum of 1-3 weeks lag, summed across service lines and days of the week,
- difference between 1 and 2 weeks lag, by service line and day
- difference between 1 and 2 weeks lag, summed across service lines and days of the week,
- difference between 1 and 3 weeks lag, summed across service lines and days of the week, and
- difference between 2 and 3 weeks lag, summed across service lines and days of the week.

For the third group, based on interviews with experts, we construct 11 indicator variables representing the relation to public holidays, indicating if

- a day of the week is a public holiday (0-1),
- there is a public holiday 1-3 days before the week of interest (0-1), or
- there is a public holiday 1-3 days after the week of interest (0-1).

This yields a total of 162 features.

Because we use a large number of features in our analyses, it is important to understand which features drive the prescription performance the most, and how the approaches react to an overly large number of features. Because both

of our prescriptive approaches use a random forest kernel or weight function, which is based on a random forest model, we analyse the feature importance by means of the decrease in node impurity (residual sum of squares) achieved by splitting using this variable, averaged over the three service lines.

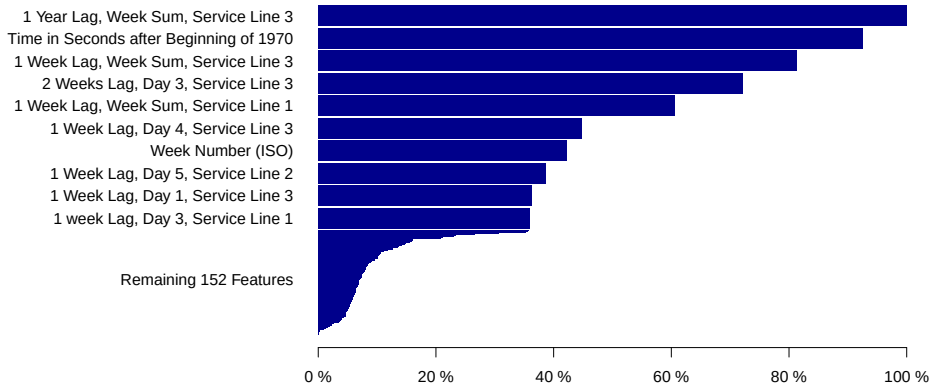


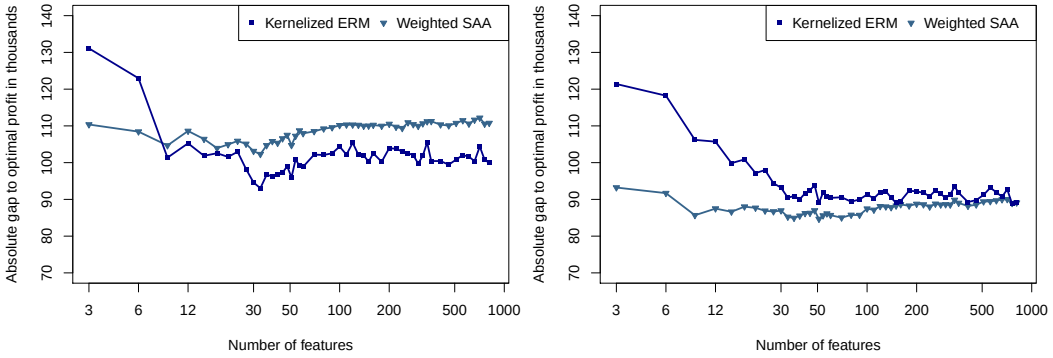
Figure A.3: Feature importance.

The results are depicted in Figure A.3. While the importance of the week number and lagged demands was expected (due to the yearly structure of demand which exhibits similarity from year to year), the importance of the time in seconds (which is a continuously increasing number) mainly stems from the demand structure of the first service line. In the majority of cases the random forest splits using this feature in June 2014, separating the first six months of training data from the rest—which can be understood by the differing demand structure visible in Figure 2.1a.

To study the impact of using only few (the most important features) up to a very large number of features (including higher-order features, e.g. the square of existing features), we plot the performance of wSAA and kERM, in Figure A.4, for varying numbers of features.[53]

Clearly, neither of the approaches suffers greatly from a large number of features—both approaches lead to a comparably stable performance even

[53]For simplicity we use a kernel based on three scalar random forest models that predict the average demand over one week for each service line (trained using the *caret* R-package as described in Kuhn (2008) with cross validation parameter tuning, varying the number of variables to select for each tree between 5 and 140). Using these three random forest models we calculate the kernel matrix by averaging over all three random forest models.



(a) Service level of 50%. (b) Service level of 80%.

Figure A.4: Performance of wSAA and kERM for a varying number of features.

in the high-dimensional regime of $p = 810 \gg N = 157$. This observation is not surprising, because the random forest includes feature selection by design (see Section 2.4.2). The results are further consistent with similar numerical experiments conducted by BK (BK, p. 1028).

We further observe that wSAA already performs well with a very small number of features (the most important features), while kERM—at least for higher service levels—appears to require more features, and that the optimal number of features appears to depend on the optimal service level (and therefore on the planning problem itself).

A.4 Detailed Description of Approaches for Numerical Evaluation

In this section we describe the implementation of each of the approaches presented in Section 2.6.3 in more detail.

A.4.1 Kernelized ERM

The implementation of kERM (2.11) using the random forest kernel (2.13) consists of two parts: the computation of the kernel matrix $\mathbf{K}^{\mathrm{RF}}_{pq}$, and the estimation of the prescription function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ by solving (2.12). To calcu-

late $\mathbf{K}^{\mathrm{RF}}_{pq}$, we train a multivariate random forest model to predict the demands for each service line and each day of one week using the *RandomForestRegressor* class of the *scikit-learn* package for python.[54] Using this random forest model we calculate the kernel matrix for training $\mathbf{K}^{\mathrm{RF}}_{\mathrm{train},pq}$ with $p, q = 1...N$ for N training samples by implementing (2.13). Similarly, we calculate the kernel matrix for evaluation $\mathbf{K}^{\mathrm{RF}}_{\mathrm{eval},pq}$ with $p = 1...N$ and $q = 1...N_{\mathrm{test}}$ for N_{test} samples of the evaluation period. Then, in the second step, we solve Problem 2.12 using *Gurobi Optimizer*. Based on the estimated β_j^{tn} and ϵ_j^n we derive the optimal values for u_j^n, and from these we derive the prescription function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ as stated in (2.11). To estimate the remaining parameters b_j we solve Problem 2.10, which is linear in b_j when applying (2.11), using *Gurobi Optimizer*. Based on the function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ we estimate the capacity prescriptions for the evaluation period.

The prescription functions are optimized by tuning the cost parameters λ_j, which is accomplished using simple cross validation, with 2/3 of the training data used to train the model, and the remaining 1/3 of the training data to evaluate the achieved profit. As the cost parameters need to balance the variation of the prescription function and the achieved profit, while the latter scales approximately with a_{jj}, we set $\lambda_j = c_0 \cdot a_{jj}$ and tune $c_0 = 10^{-4}...10^4$ using the procedure described above. For each set of exogenous parameters, the cost parameter with the highest validation profit is chosen.

A.4.2 Weighted SAA

wSAA is implemented in two steps, with the first step being identical to kERM: we use the same matrix $\mathbf{K}^{\mathrm{RF}}_{\mathrm{eval},pq}$ (see Appendix A.4.1) as weight function, which implements $w_p^{\mathrm{RF}}(\vec{x}^q)$ as defined in (2.9) with $p = 1...N$ for N training samples and $q = 1...N_{\mathrm{test}}$ for N_{test} samples of the evaluation period. Based on this weight function, we solve Problem 2.7 using *Gurobi Optimizer*.

[54] Because demand is multivariate, a multivariate random forest is more appropriate than a combination of scalar random forest models.

A.4.3 SAA

The SAA approach is a simplified version of wSAA, and can be derived by setting $w_n(\vec{x}) := 1/N$ in Problem 2.7:

$$\begin{aligned}\vec{q}^{\,\mathrm{SAA}}(\vec{x}) = \arg\min_{\vec{q}\in\mathcal{Q}} \min_{\{y_{ij}^{tn}\}} \sum_{n=1}^{N} \frac{1}{N} \left(\sum_j f_j q_j - \sum_{t=1}^{T} \left(\sum_{i,j} a_{ij} y_{ij}^{tn} - \sum_i c_i d_i^{tn} \right) \right) \\ \text{s.t.} \quad \sum_j y_{ij}^{tn} \leq d_i^{tn} \ \forall i,n,t \\ \sum_i y_{ij}^{tn} \leq q_j \ \forall j,n,t \\ y_{ij}^{tn} \geq 0 \ \forall i,j,n,t \\ y_{ij}^{tn} = 0 \text{ if } i < j \ \forall n,t.\end{aligned} \tag{A.53}$$

This problem is solved using *Gurobi Optimizer*.

A.4.4 SVR-SEO

The SVR-SEO approach represents the class of traditional parametric approaches, and has been selected as benchmark approach due to its structural similarity to kERM. Because support vector regression is kernel-based, we use $\mathbf{K}^{\mathrm{RF}}_{\mathrm{train},pq}$ as calculated for kERM (see Appendix A.4.1) to train 15 epsilon-SVR models using the *kernlab* R-package with simple cross validation parameter tuning, similar to kERM.[55] We further use $\mathbf{K}^{\mathrm{RF}}_{\mathrm{eval},pq}$ to predict demand values for all service lines and all days of the evaluation period and calculate the in-sample residuals to evaluate the covariance matrix $\mathbf{\Sigma}$, from which we derive the coefficients of variation and correlation, which are assumed to be constant. Combining the predicted demand values with these coefficients we obtain a set of multivariate normal distributions of demand for each day of the evaluation period. We approach the stochastic optimization problem (2.1) using Monte Carlo simulation and sample average approximation by taking $N_{MC} = 300$ samples of the multivariate normal distributions for each day and solving the

[55] The cost parameter is tuned across the range $C = 10^{-4}...10^{4}$, while the epsilon parameter is estimated as $\epsilon = 3\sigma\sqrt{\ln N/N}$ (see Cherkassky and Ma 2002) with σ being the demand noise level, estimated using random forest models.

resulting problem using *Gurobi Optimizer*.

A.4.5 ARIMA-SEO

The ARIMA-SEO approach represents the class of traditional parametric, time series-based approaches. We selected ARIMA because it lead to a higher performance than other popular approaches (see Appendix A.6 for a performance comparison of ARIMA and ETS models). We train three ARIMA models to the demand values for the three service lines, with the model parameters being estimated using the automatic fitting approach as described in Hyndman and Khandakar (2008) and implemented in the R-package *forecast*. Similar to the SVR-SEO approach we use the in-sample residuals to estimate the covariance matrix $\mathbf{\Sigma}$ and the coefficients of variation and correlation. Combining the forecasted demand values with these coefficients we obtain a set of multivariate normal distributions of demand for each day of the evaluation period. From here we follow the exact same procedure as for SVR-SEO to solve Problem 2.1.

A.5 Definition and Details of Alternative Kernel Functions

In Section 2.6.6 we analyze the performance of kERM when using alternative kernels, including a linear kernel

$$K_{Lin}(\vec{x}_1, \vec{x}_2) = \langle \vec{x}_1, \vec{x}_2 \rangle, \tag{A.54}$$

and a polynomial kernel (homogeneous 3rd degree)

$$K_{Poly3}(\vec{x}_1, \vec{x}_2) = \langle \vec{x}_1, \vec{x}_2 \rangle^3, \tag{A.55}$$

which are both data-independent kernels. The performance analysis further includes the universal RBF Gauss kernel as defined in Section 2.5.2, and the random forest kernel presented in Section 2.5.3.

We conjectured, in Section 2.6.6, that the higher performance of kERM

at very high service levels when using an RBF Gauss kernel is rooted in the more homogeneous similarity values that induce a higher implicit uncertainty such that kERM does not exhibit a significant regularization effect. To further elucidate on this property, we plot, in Figure A.5, the standard deviation of the normalized kernel values K_{ij} with $\sum_{j=1}^{N} K_{ij} = 1$ for each data sample i.

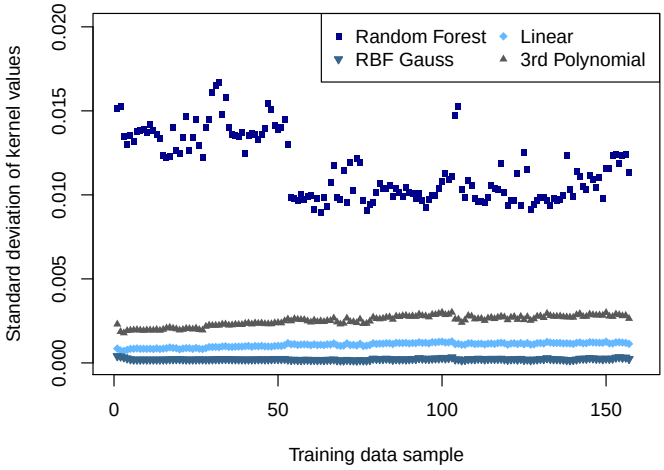


Figure A.5: Standard deviation of kernel values for all training data samples.

The standard deviation of the kernel values is clearly higher for the random forest kernel than for all other kernels, which are data-independent. This suggests that the random forest kernel assigns a higher similarity to a few samples, while most others exhibit a low similarity value.

A.6 Performance Comparison of ARIMA and ETS Models

To compare the prescription performance of ARIMA and ETS models in terms of gap to optimal profit, we follow the same procedure as described in Section 2.6. Figure A.6 shows that ARIMA models outperform ETS models for the data set at hand over the full range of service levels. We therefore choose an ARIMA model as time series-based benchmark for the numerical evaluation.

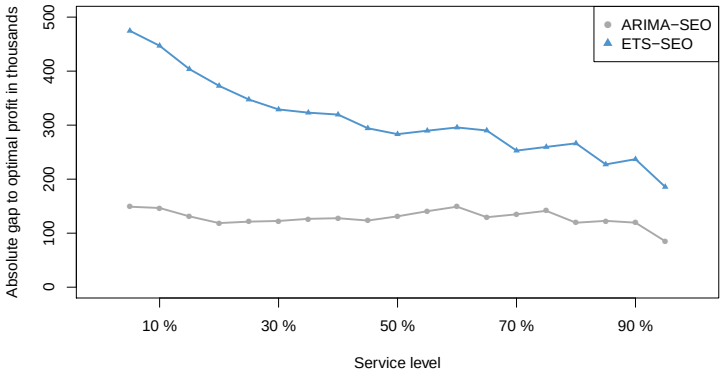


Figure A.6: Absolute gap to optimal profit across service levels.

A.7 Additional Theoretical Insights

A.7.1 Analytical Results for Weighted SAA

This section presents two analytical results for wSAA—asymptotic optimality for a number of weight functions and, for scalar-valued decisions, the restriction of the prescriptions to convex combinations of decisions that would have been optimal for past observations.

Proposition A.1. *(Following BK) $\vec{q}^{wSAA}(\vec{x})$ is asymptotically optimal over closed, bounded, non-empty decision space $\tilde{\mathcal{Q}} \subset \mathbb{R}_+^I$ and bounded, non-empty demand space $\tilde{\mathcal{D}} \subset \mathbb{R}_+^{I \times T}$ for S_N generated by iid sampling and any of the following weight functions:*

a) Based on kNN:

$$w_n^{kNN}(\vec{x}) = \frac{1}{k} \mathbb{1}[\vec{x}^n \text{ is a kNN of } \vec{x}], \tag{A.56}$$

with $k = \min(\lceil CN^\delta \rceil, N-1)$ for some $C > 0$, $0 < \delta < 1$.

b) Based on kernel methods:

$$w_n^K(\vec{x}) = \frac{K((\vec{x}^n - \vec{x})/h_N)}{\sum_{k=1}^N K((\vec{x}^k - \vec{x})/h_N)}, \tag{A.57}$$

with $h_N = CN^{-\delta}$ for some $C > 0$, $0 < \delta < 1/p$ with $\vec{x} \in \mathbb{R}^p$, and K being

one of the following kernels: naïve $K(\vec{x}) = \mathbb{1}[||\vec{x}|| \leq 1]$, *Epanechnikov* $K(\vec{x}) = (1-||\vec{x}||^2)\mathbb{1}[||\vec{x}|| \leq 1]$, *Tri-cubic* $K(\vec{x}) = (1 - ||\vec{x}||^3)^3\mathbb{1}[||\vec{x}|| \leq 1]$, *or Gaussian* $K(\vec{x}) = \exp(-||\vec{x}||^2/2)$.

c) *Based on recursive kernel methods:*

$$w_n^{rK}(\vec{x}) = \frac{K((\vec{x}^n - \vec{x})/h_n)}{\sum_{k=1}^N K((\vec{x}^k - \vec{x})/h_k)}, \tag{A.58}$$

with $h_n = Cn^{-\delta}$ *for some* $C > 0$, $0 < \delta < 1/(2p)$ *with* $\vec{x} \in \mathbb{R}^p$, *and* K *being the naïve kernel* $K(\vec{x}) = \mathbb{1}[||\vec{x}|| \leq 1]$.

d) *Based on local linear methods:*

$$w_n^{LL}(\vec{x}) = \frac{\tilde{w}_n(\vec{x})}{\sum_{k=1}^N \tilde{w}_k(\vec{x})}, \tag{A.59}$$

with

$$\begin{aligned} \tilde{w}_n(\vec{x}) &= k_n(\vec{x}) \left(1 - \sum_{l=1}^N k_l(\vec{x})(\vec{x}^l - \vec{x})^T \Xi(\vec{x})^{-1}(\vec{x}^n - \vec{x})\right), \\ \Xi(\vec{x}) &= \sum_{n=1}^N k_n(\vec{x})(\vec{x}^n - \vec{x})(\vec{x}^n - \vec{x})^T, \\ k_n(\vec{x}) &= K((\vec{x}^n - \vec{x})/h_N), \end{aligned} \tag{A.60}$$

and $h_N = CN^{-\delta}$ *for some* $C > 0$, $0 < \delta < 1/p$ *with* $\vec{x} \in \mathbb{R}^p$, *and* K *being one of the following kernels: naïve* $K(\vec{x}) = \mathbb{1}[||\vec{x}|| \leq 1]$, *Epanechnikov* $K(\vec{x}) = (1-||\vec{x}||^2)\mathbb{1}[||\vec{x}|| \leq 1]$, *Tri-cubic* $K(\vec{x}) = (1 - ||\vec{x}||^3)^3\mathbb{1}[||\vec{x}|| \leq 1]$, *or Gaussian* $K(\vec{x}) = \exp(-||\vec{x}||^2/2)$, *if the distribution of feature vectors* $\vec{x}$ *is absolutely continuous and the probability density* $f(\vec{x})$ *is bounded away from 0 and* ∞ *over* $\mathcal{X}$.

Proof of Proposition A.1: To prove asymptotic optimality of $\vec{q}^{\text{wSAA}}(\vec{x})$, we apply Theorems 2-5 presented in BK to our loss function. Because any bounded $\tilde{\mathcal{Q}}$ and $\tilde{\mathcal{D}}$ are subsets of $\mathcal{Q}$ and $\mathcal{D}$ (Definition 2.2), we obtain the following properties for our loss function L:

- Existence (Assumption 1 in BK): The loss function is bounded (see Lemma 2.1), therefore $\mathbb{E}(|L(\vec{q}, \boldsymbol{D})|) \leq \bar{l} < \infty$, and the feasible region is $\mathcal{Q} \neq \varnothing$.
- Continuity (Assumption 2 in BK): The loss function is equi-Lipschitz (Lemma 2.1).
- Regularity (Assumption 3 in BK): $\mathcal{Q} \neq \varnothing$ is closed, bounded and non-empty by definition.

In addition, because $0 \leq L(\vec{q}, \boldsymbol{d}) \leq \bar{l}$ (Lemma 2.1), the following holds:

$$\mathbb{E}[|L(\vec{q}, \boldsymbol{D})| \max(\log |L(\vec{q}, \boldsymbol{D})|, 0)] \leq \bar{l} \log \bar{l} < \infty.$$

Therefore, the results of Theorems 2-5 (BK) can be applied to our loss function, proving asymptotic optimality of wSAA for any of the weight functions presented in Proposition A.1.

Proposition A.1 states that wSAA is asymptotically optimal for a number of classes of weight functions and certain conditions, e.g., for kNN the parameter k needs to be chosen accordingly.

Proposition A.2. *Assume $\mathcal{Q} = \mathcal{D} = \mathbb{R}$, and a loss function $L(q, d)$ strictly convex in q. Let*

$$\mathcal{Q}_{int} = \{q_n := \arg\min_{q \in \mathcal{Q}} L(q, d^n); n = 1, ..., N\} \tag{A.61}$$

denote the set of optimal solutions for individual demand realizations d_n and $\text{conv}(\mathcal{Q}_{int})$ *its convex hull. Then*

$$q^{wSAA}(\vec{x}) \in \text{conv}(\mathcal{Q}_{int}) \subset \mathcal{Q} \; \forall \vec{x} \in \mathcal{X}. \tag{A.62}$$

Proof of Proposition A.2: To prove $q^{\text{wSAA}}(\vec{x}) \in \text{conv}(\mathcal{Q}_{\text{int}}) \; \forall \vec{x} \in \mathcal{X}$, we first define $g_{\vec{x}}(q) := \sum_{n=1}^{N} w_n(\vec{x}) L(q, d^n)$. Because $q^{\text{wSAA}}(\vec{x}) = \arg\min_{q \in \mathcal{Q}} g_{\vec{x}}(q)$, we need to show that

$$\forall q \in \mathcal{Q}, q \notin \text{conv}(\mathcal{Q}_{\text{int}}) \; \exists q' \in \text{conv}(\mathcal{Q}_{\text{int}}) : \; g_{\vec{x}}(q') < g_{\vec{x}}(q), \tag{A.63}$$

meaning that the minimizer of $g_{\vec{x}}(q)$ is in $\text{conv}(\mathcal{Q}_{\text{int}})$.

Assume, without loss of generality, that the minimizers of the loss function $q_n = \arg\min_{q \in \mathcal{Q}} L(q, d^n)$ for the individual demand realizations d^n are ordered as $q_1 \leq q_2 \leq ... \leq q_N$. Consequently, $q_1 \leq q \leq q_N \ \forall q \in \text{conv}(\mathcal{Q}_{\text{int}})$. Let $\Delta > 0$, so that

$$q_1 - \Delta < q_1 \leq q_n \ \forall q_n. \tag{A.64}$$

Then there exists a $0 \leq \lambda_n < 1$ such that $q_1 = \lambda_n(q_1 - \Delta) + (1 - \lambda_n)q_n$. Because $L(q, d)$ is convex, we know that

$$L(q_1, d^n) \leq \lambda_n L(q_1 - \Delta, d^n) + (1 - \lambda_n)L(q_n, d^n) \tag{A.65}$$

for all $n = 1...N$. Because the optimal solutions q_n are unique ($L(q, d)$ strictly convex), we know that $L(q_n, d^n) < L(q_1 - \Delta, d^n)$, and

$$L(q_1, d^n) < \lambda_n L(q_1 - \Delta, d^n) + (1 - \lambda_n)L(q_1 - \Delta, d^n) = L(q_1 - \Delta, d^n). \tag{A.66}$$

Therefore,

$$g_{\vec{x}}(q_1 - \Delta) = \sum_{n=1}^{N} w_n(\vec{x})L(q_1 - \Delta, d^n) > \sum_{n=1}^{N} w_n(\vec{x})L(q_1, d^n) = g_{\vec{x}}(q_1). \tag{A.67}$$

In the same way we can show $g_{\vec{x}}(q_N + \Delta) > g_{\vec{x}}(q_N)$, therefore,

$$g_{\vec{x}}(q) > g_{\vec{x}}(q_1) \ \forall q < q_1 \text{ and } g_{\vec{x}}(q) > g_{\vec{x}}(q_N) \ \forall q > q_N, \tag{A.68}$$

which was to be proven.

Proposition A.2 demonstrates that wSAA, assuming scalar-valued decisions and an l_2 loss function, is restricted to interpolations between optimal solutions for historical observations of demand. In the following section we study how this may lead to sub-optimal decisions in case of strong trends.

A.7.2 Characteristics of kERM and wSAA for a Linear Demand Model

We illustrate the characteristics of kERM and wSAA concerning extrapolation and guaranteed feasibility of prescriptions by considering a strongly simplified version of our capacity planning problem. In particular, we consider the problem of planning the capacity for a single service line and a decision horizon of $T = 1$ day, which reduces the problem to a newsvendor-like situation. Let C_O and C_U denote the overage and under cost factors, and assume a feature-driven linear demand model, as provided in BR:

$$D|(X = x) = \beta x + \epsilon, \tag{A.69}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ is independent noise with cdf F_ϵ.

Proposition A.3. *Assume a linear demand model as in (A.69) with $\beta > 0$, a service level $C_U/(C_U+C_O) > 0.5$, and a training data set $S_N = \{(x^n, d^n)\}$ with $x^1 \leq x^2 \leq ... \leq x^{N-1} \leq x^N$. Let k_0 denote the number of nearest neighbors, let $0 < \delta < 1 - (1/2)^{k_0}$, and let $\Delta \geq 2\beta^{-1}F_\epsilon^{-1}\left((1-\delta)^{1/k_0}\right)$.*
Then

a) the prescription of wSAA, using a k_0 nearest neighbors weight function, for a new feature $x^{N+1} = x^N + \Delta$ will deviate from the optimal decision $q^*(x^{N+1})$ as

$$\left| q^{wSAA}(x^{N+1}) - q^*(x^{N+1}) \right| \geq \beta \frac{\Delta}{2} \tag{A.70}$$

with probability of at least $1 - \delta$,

b) kERM, using a linear kernel (corresponding to a linear function space), converges asymptotically to the optimal solution

$$q^{kERM}(x) \to q^*(x) = \beta x + F_\epsilon^{-1}\left(\frac{C_U}{C_U + C_O}\right). \tag{A.71}$$

Proof of Proposition A.3: The convergence of $q^{\mathrm{kERM}}(x)$ toward the op-

timal solution has been shown in Theorem 3 in BR. To prove the deviation of $q^{\text{wSAA}}(x^{N+1})$ from the optimal solution, we show that wSAA cannot prescribe capacities larger than its largest training demand value, while demand increases according to the linear model.
Because $q^{\text{wSAA}}(x) \in \text{conv}(\mathcal{Q}_{\text{int}})$ (see Proposition A.2) and the single-instance solution $q^n = \arg\min_{q\in\mathcal{Q}} L(q, d^n) = d^n$ for the newsvendor problem, we obtain

$$q^{\text{wSAA}}(x) \leq \max_{n\leq N} d^n \leq \max_{n\leq N}(\beta x^n + \epsilon^n) \ \forall x \tag{A.72}$$

with ϵ^n being the nth realization of the random variable ϵ.

Let $\bar{\epsilon} := F_\epsilon^{-1}((1-\delta)^{1/k_0})$, then $F_\epsilon(\bar{\epsilon}) = (1-\delta)^{1/k_0}$, which means that $\epsilon^n \leq \bar{\epsilon}$ with probability $(1-\delta)^{1/k_0}$. Let K_0 be the set of the k_0 nearest neighbors of x^{N+1}, then with probability of at least $1-\delta$:

$$q^{\text{wSAA}}(x^{N+1}) \leq \max_{n\in K_0}(\beta x^n + \epsilon^n) \leq \beta x^N + \bar{\epsilon} \leq \beta\left(x^N + \Delta/2\right). \tag{A.73}$$

Therefore, with probability of at least $1-\delta$, we obtain for the deviation of the wSAA prescription:

$$q^*(x^{N+1}) - q^{\text{wSAA}}(x^{N+1}) \geq \beta\left(x^N + \Delta\right) - \beta\left(x^N + \Delta/2\right) = \beta\frac{\Delta}{2} > 0, \tag{A.74}$$

from which we derive the statement to be proven.

Proposition A.3 shows that for the case of linear demand with trend, the restriction of wSAA to the convex hull of optimal decisions for past observations may lead to deviations from the optimal decision.

To analyze the feasibility of the prescriptions obtained from kERM and wSAA, assume a feasible region for our simplified model that is restricted as $\hat{\mathcal{Q}} = \{q \in \mathbb{R} : q \leq q_{\max}\}$, for example because of limitations of working space or machine capacity. In this setting, wSAA will obey the restriction on $\hat{\mathcal{Q}}$ and prescribe $q^{\text{wSAA}}(x) \leq q_{\max}$, even for $x > q_{\max}/\beta$. The kERM approach, in contrast, will only obey this restriction when learning the prescription function $q^{\text{kERM}}(\cdot)$. Assuming $d^n \leq q_{\max}\ \forall n \leq N$ and $C_U/(C_U + C_O) = 0.5$, kERM will therefore still converge toward $q^*(x) = \beta x + F_\epsilon^{-1}\left(\frac{C_U}{C_U+C_O}\right) = \beta x$. How-

ever, this function may prescribe infeasible solutions for $x^{N+1} > q_{\max}/\beta$ as $q^{\text{kERM}}(x^{N+1}) > q_{\max}$, which need to be corrected by post-processing.

A.7.3 Analysis of Intra-Week Variation Structure

While the capacity planning problem as stated in (2.1) is based on a set of T daily demands $\vec{D}^t$, each of which follows a distribution with density f_t, we can restate the relevant part of the problem using the average of the distribution densities $\tilde{f} = \frac{1}{T}\sum_t f_t$ as follows:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left(\pi(\vec{D}^t, \vec{q})\right) &= \sum_{t=1}^{T} \int_{-\infty}^{\infty} \pi(\vec{x}, \vec{q}) f_t(\vec{x}) d\vec{x} \\ &= T \int_{-\infty}^{\infty} \pi(\vec{x}, \vec{q}) \tilde{f}(\vec{x}) d\vec{x} = T\mathbb{E}\left(\pi(\tilde{\vec{D}}, \vec{q})\right).\end{aligned} \tag{A.75}$$

To characterize the distribution of $\tilde{\vec{D}}$ with density function $\tilde{f}$, we estimate mean $\tilde{\mu}$ and variation $\tilde{\sigma}^2$ for the simplified case of $\tilde{D} \in \mathbb{R}$. With mean μ_t and variation σ_t^2 of the individual daily demands $D^t \in \mathbb{R}$ we obtain:

$$\tilde{\mu} = \int_{-\infty}^{\infty} x\tilde{f}(x)dx = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty} x f_t(x)dx = \frac{1}{T}\sum_{t=1}^{T}\mu_t. \tag{A.76}$$

The mean $\tilde{\mu}$ of $\tilde{D}$ is therefore the mean of the individual μ_t for each day $t = 1...T$. We further obtain that the variation $\tilde{\sigma}^2$ consists of two components:

$$\begin{aligned}\tilde{\sigma}^2 &= \int_{-\infty}^{\infty}(x-\tilde{\mu})^2\tilde{f}(x)dx = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}\left((x-\mu_t)+(\mu_t-\tilde{\mu})\right)^2 f_t(x)dx \\ &= \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2 + \frac{1}{T}\sum_{t=1}^{T}(\mu_t-\tilde{\mu})^2 = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2 + \sigma_w^2.\end{aligned} \tag{A.77}$$

We call σ_w^2 the intra-week variation, which is the variance of the individual daily mean demands μ_t around $\tilde{\mu}$. For a constant daily demand variation $\sigma_t^2 = \sigma_d^2$ for all days t, we have $\tilde{\sigma} = \sqrt{\sigma_d^2 + \sigma_w^2}$.

A.7.4 ERM Solution for a Linear Function Space

For solving the ERM approach over a linear function space, we define

$$\mathcal{F} = \{\vec{q}(\cdot) : \vec{q}(\vec{x}) = \mathbf{W}\vec{x} - \vec{b}\}, \tag{A.78}$$

where $\mathbf{W} = (\vec{w}_1, \vec{w}_2, ..., \vec{w}_I)$ is the matrix representing the linear mapping of $\vec{x}$, and $\vec{b} = (b_1, b_2, ..., b_I)$ represents a constant offset[56].

For this particular function space, Problem 2.10 can be stated as:

$$\begin{aligned}
\min_{\mathbf{W},\vec{b},\{y_{ij}^{tn}\}} \; & \sum_{j=1}^{I} \lambda_j \|\vec{w}_j\|^2 + \sum_{n=1}^{N} \left(\sum_j f_j(\vec{w}_j \cdot \vec{x}^n - b_j) - \sum_{t=1}^{T} \left(\sum_{i,j} a_{ij} y_{ij}^{tn} - \sum_i c_i d_i^{tn} \right) \right) \\
\text{s.t.} \; & \sum_j y_{ij}^{tn} \leq d_i^{tn} \; \forall i, n, t \\
& \sum_i y_{ij}^{tn} \leq (\vec{w}_j \cdot \vec{x}^n - b_j) \; \forall j, n, t \\
& y_{ij}^{tn} \geq 0 \; \forall i, j, n, t \\
& y_{ij}^{tn} = 0 \text{ if } i < j \; \forall n, t \\
& \vec{w}_j \cdot \vec{x}^n - b_j \geq 0 \; \forall j, n.
\end{aligned} \tag{A.79}$$

In (A.79) we use the norm $\sum_{j=1}^{I} ||\vec{w}_j||^2$ for regularization so that we do not penalize shifts in the mean by the constant offset $\vec{b}$.[57] We also included a regularization parameter λ_j for each capacity (service line) j, which becomes important when profitabilities $a_{i,j}$ vary widely across capacities j. In such situations, a constant regularization parameter could lead to high variance for highly profitable capacities and high bias for capacities with low profitability at the same time.

Proposition A.4. *The objective function of the linear ERM approach is jointly convex in* $\{\vec{w}_j\}$ *and* $\vec{b}$.

[56] We introduce a constant offset to allow for a feature-independent term (e.g., allowing functions with non-zero mean for $\mathcal{X}$ centered around 0). Alternatively, one could add a constant entry $x_0 = 1$ to the feature vector, which would integrate the constant offset b_j as the first entry of each $\vec{w}_j$ (e.g., as in Section 2.3.1 in BR).

[57] The approach of not penalizing shifts in the mean is commonly chosen in regression problems. See for example Evgeniou et al. (2000) or Smola and Schölkopf (2004).

Proof of Proposition A.4: Because the objective function of the ERM approach (2.10) is convex (Proposition 2.5) and convexity is invariant under affine maps, such as $\vec{q}(\mathbf{W}, \vec{b}, \vec{x}) = \mathbf{W}\vec{x} - \vec{b}$, the objective function of the linear ERM approach is jointly convex in $\{\vec{w}_j\}$ and $\vec{b}$.

Theorem A.2. *The optimal solution to Problem A.79 is:*

$$\vec{q}^{ERM}(\vec{x}) = \sum_{n=1}^{N} \vec{u}^n \left(\vec{x}^n \cdot \vec{x}\right) - \vec{b}, \tag{A.80}$$

where the components of $\vec{u}^n$ *are defined as* $u_j^n = \frac{1}{2\lambda_j}\left(\sum_{t=1}^{T}(\beta_j^{tn}) + \epsilon_j^n - f_j\right)$, *and* β_j^{tn}, ϵ_j^n *is the optimal solution to the dual problem of (A.79):*

$$\begin{aligned}
\max_{\{\alpha_i^{tn}\},\{\beta_j^{tn}\},\{\epsilon_j^n\}} L_{dual} := & -\sum_{j=1}^{I}\lambda_j \sum_{p,q=1}^{N}\left(u_j^p u_j^q (\vec{x}^p \cdot \vec{x}^q)\right) + \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{t=1}^{T}(c_i - \alpha_i^{tn})d_i^{tn} \\
s.t.\ & \alpha_i^{tn},\ \beta_j^{tn},\ \epsilon_j^n \geq 0\ \forall i, n, t \\
& \alpha_i^{tn} + \beta_j^{tn} \geq a_{ij}\ \forall i \geq j,\ \forall n, t \\
& \sum_{n=1}^{N} u_j^n = 0\ \forall j.
\end{aligned} \tag{A.81}$$

Proof of Theorem A.2: To prove the solution stated in Theorem A.2 we derive the primal Lagrange function and prove strong duality, which allows to express the solution to Problem A.79 using the Lagrange dual function.

Let α_i^{tn}, β_j^{tn}, γ_{ij}^{tn}, δ_{ij}^{tn} and ϵ_j^n be Lagrange multipliers, then we obtain for the primal Lagrangian:

$$\begin{aligned}
L_{\text{primal}} = & \sum_{j=1}^{I}\lambda_j\|\vec{w}_j\|^2 + \sum_{n=1}^{N}\left(\sum_j f_j(\vec{w}_j \cdot \vec{x}^n - b_j) - \sum_{t=1}^{T}\left(\sum_{i,j} a_{ij}y_{ij}^{tn} - \sum_i c_i d_i^{tn}\right)\right) \\
& + \sum_{i,n,t}\alpha_i^{tn}\left(\sum_j y_{ij}^{tn} - d_i^{tn}\right) + \sum_{j,n,t}\beta_j^{tn}\left(\sum_i y_{ij}^{tn} - (\vec{w}_j \cdot \vec{x}^n - b_j)\right) \\
& - \sum_{i,j,n,t}\gamma_{ij}^{tn}y_{ij}^{tn} + \sum_{n,t,i<j}\delta_{ij}^{tn}y_{ij}^{tn} - \sum_{j,n}\epsilon_j^n\left(\vec{w}_j \cdot \vec{x}^n - b_j\right),
\end{aligned} \tag{A.82}$$

where $i, j = 1...I$, $t = 1...T$, and $n = 1...N$.

Because the objective function of the linear ERM approach is convex (Proposition A.4), all constraints of Problem A.79 are affine in the primal variables $\mathbf{W}, \vec{b}, \{y_{ij}^{tn}\}$, and $\mathbf{W}, \vec{b}, \{y_{ij}^{tn}\} = 0$ is a feasible solution with 0 being a relative interior point of the domain of definition $\mathbb{R}^{\nu}$ with dimension $\nu = I \times \dim \mathcal{X} + I + I \times I \times N \times T$, the Slater condition is fulfilled and we obtain that strong duality holds (see Section 5.2.3 in Boyd and Vandenberghe 2004). Therefore, the Karush-Kuhn-Tucker (KKT) conditions, which state that the partial derivatives of L_{primal} with respect to the primal variables $\vec{w}_j$, b_j, and y_{ij}^{tn} equal zero, provide necessary and sufficient conditions for optimality (Boyd and Vandenberghe 2004, p. 244):

$$\begin{aligned}
\frac{\partial L_{\text{primal}}}{\partial \vec{w}_j} &= 2\lambda_j \vec{w}_j + \sum_{n=1}^{N} \left(f_j - \sum_t \beta_j^{tn} - \epsilon_j^n \right) \vec{x}^n = 0 \\
\frac{\partial L_{\text{primal}}}{\partial b_j} &= \sum_{n=1}^{N} \left(-f_j + \sum_t \beta_j^{tn} + \epsilon_j^n \right) = 0 \\
\text{for } i < j\text{:} \quad \frac{\partial L_{\text{primal}}}{\partial y_{ij}^{tn}} &= -a_{ij} + \alpha_i^{tn} + \beta_j^{tn} - \gamma_{ij}^{tn} + \delta_{ij}^{tn} = 0 \\
\text{for } i \geq j\text{:} \quad \frac{\partial L_{\text{primal}}}{\partial y_{ij}^{tn}} &= -a_{ij} + \alpha_i^{tn} + \beta_j^{tn} - \gamma_{ij}^{tn} = 0.
\end{aligned} \tag{A.83}$$

Defining

$$u_j^n := \frac{1}{2\lambda_j} \left(\sum_t (\beta_j^{tn}) + \epsilon_j^n - f_j \right), \tag{A.84}$$

we obtain

$$\begin{aligned}
&\vec{w}_j = \sum_{n=1}^{N} u_j^n \vec{x}^n, \; 0 = \sum_{n=1}^{N} u_j^n \; \forall j, \text{ and} \\
&\gamma_{ij}^{tn} = -a_{ij} + \alpha_i^{tn} + \beta_j^{tn} + \begin{cases} \delta_{ij}^{tn} \text{ for } i < j \\ 0 \text{ otherwise,} \end{cases}
\end{aligned} \tag{A.85}$$

which allows us to state the primal Lagrangian as

$$L_{\text{primal}} = -\sum_{j=1}^{I} \lambda_j \sum_{p=1}^{N} \sum_{q=1}^{N} \left(u_j^p u_j^q (\vec{x}^p \cdot \vec{x}^q) \right) + \sum_{i,n,t} (c_i - \alpha_i^{tn}) d_i^{tn}. \tag{A.86}$$

Because L_{primal} is independent of the primal variables, we obtain for the dual

Lagrangian:

$$L_{\text{dual}} = \inf_{\mathbf{W},\vec{b},\{y_{ij}^{tn}\}} L_{\text{primal}} = L_{\text{primal}}, \tag{A.87}$$

from which we derive for the dual problem:

$$\begin{aligned} \max_{\{\alpha_i^{tn}\},\{\beta_j^{tn}\},\{\epsilon_j^n\}} & L_{\text{dual}} \\ \text{s.t. } & \alpha_i^{tn},\ \beta_j^{tn},\ \epsilon_j^n \geq 0\ \forall i,n,t \\ & \alpha_i^{tn} + \beta_j^{tn} \geq a_{ij}\ \forall i \geq j,\ \forall n,t \\ & \sum_{n=1}^{N} u_j^n = 0\ \forall j, \end{aligned} \tag{A.88}$$

which concludes the proof.

Corollary A.1. *The objective function L_{dual} of the dual problem (A.81) is concave in the Lagrange multipliers α_i^{tn}, β_j^{tn}, ϵ_j^n.*

Proof of Corollary A.1: Similarly as in the proof of Corollary 2.1, the primal Lagrangian L_{primal} is by definition affine and therefore concave in the Lagrangian multipliers α_i^{tn}, β_j^{tn}, ϵ_j^n. The dual Lagrangian is the point-wise infimum of a collection of concave functions:

$$L_{\text{dual}}(\{\alpha_i^{tn}\},\{\beta_j^{tn}\},\{\epsilon_j^n\}) = \inf_{\mathbf{W},\vec{b},\{y_{ij}^{tn}\}} L_{\text{primal}}(\mathbf{W},\vec{b},\{y_{ij}^{tn}\},\{\alpha_i^{tn}\},\{\beta_j^{tn}\},\{\epsilon_j^n\}), \tag{A.89}$$

and therefore concave in α_i^{tn}, β_j^{tn}, ϵ_j^n.

Problem A.81 is a quadratic optimization problem that can be solved efficiently using standard non-linear programming techniques. Given the optimal $\vec{u}^n$, the offset $\vec{b}$ can be determined by solving Problem A.79, which is linear in b_j.

A.7.5 Non-universality of the Random Forest Kernel

While our numerical experiments suggest that the random forest kernel may lead to superior performance in practical settings, a natural question is whether the random forest kernel is also universal. To shed light on this aspect, we

consider Simon-Gabriel and Schölkopf (2018), who showed that universality of a kernel is equivalent to a generalized definition of strictly positive definiteness (s.p.d.) for kernels (Theorem 6 in Simon-Gabriel and Schölkopf 2018). Because this generalized notion of s.p.d. contains (and therefore requires) the classical notion of s.p.d. of matrices and kernel matrices (see Definition 5 in Simon-Gabriel and Schölkopf 2018)—that is $\vec{\gamma}^T \boldsymbol{K} \vec{\gamma} > 0 \; \forall \vec{\gamma} \neq 0$—we can show by contradiction that the random forest kernel is not generally universal, and that the universal approximation property of kERM does not apply. The intuition behind the non-universality of the random forest kernel is that, because it is trained on a particular data set, a random forest may not distinguish between different feature vectors $\vec{x}$ across some region $\mathcal{R} \subset \mathcal{X}$, given that the demand does not differ across $\mathcal{R}$. Consequently none of the functions of the corresponding reproducing kernel Hilbert space distinguish between these differing feature vectors $\vec{x}$ and the function space is therefore not dense in the space of all continuous functions $C(\mathcal{X}, \mathcal{Q})$.

Proposition A.5. *The random forest kernel as defined in (2.13) is not generally universal.*

Proof of Proposition A.5: To prove the stated result, we show that the random forest kernel may lead to a kernel matrix that is not s.p.d., while being s.p.d. is a requirement for universality (Simon-Gabriel and Schölkopf 2018). Assume $\mathcal{D} = \mathcal{X} = \mathbb{R}$ and two feature vectors $x_1 < x_2$, which define the set $\mathcal{R} = \{x \in \mathcal{X} | x_1 \leq x \leq x_2\} \subset \mathcal{X}$. Further assume a joint distribution of $X \times D$ and that there is no difference in D over $\mathcal{R}$, such that the conditional probability distributions of D are $f(d, x) = f(d, x_1) = f(d, x_2) \; \forall x \in \mathcal{R}$. Let m_0 be the mean of D over $\mathcal{R}$. Further assume the sets $\mathcal{R}_a = \{x \in \mathcal{X} | x < x_1\} \subset \mathcal{X}$ and $\mathcal{R}_b = \{x \in \mathcal{X} | x > x_2\} \subset \mathcal{X}$ with m_a, m_b the means of D over $\mathcal{R}_a$ and $\mathcal{R}_b$.

When learning the random forest model for a data set $S_N = \{(x^n, d^n)\}$, each tree will determine the optimal splitting point s for each split to minimize the remaining sum of residual squares of the demand values d^n of the data set:

$$\min_s \sum_{\mathcal{R}_1 := \{n | x^n \leq s\}} (d^n - m_1)^2 + \sum_{\mathcal{R}_2 := \{n | x^n > s\}} (d^n - m_2)^2, \tag{A.90}$$

where m_1, m_2 are the mean values of the left and right child nodes $\mathcal{R}_1$ and $\mathcal{R}_2$ of the tree. When considering a split within $\mathcal{R}$, that is $x_1 \leq s \leq x_2$, we show in the following that splitting at either $s = x_1$ or $s = x_2$ is optimal, and the tree therefore does not split within $\mathcal{R}$.

Assume that $\mathcal{R}_a$ contains a data points x^n, that $\mathcal{R}_b$ contains b data points, and that $\mathcal{R}$ contains c data points and let j be the number of data points $x^n \in \mathcal{R}$ assigned to $\mathcal{R}_1$.

Then, for large N, the means m_1 and m_2 are given as:[58]

$$m_1 = \frac{m_a a + m_0 j}{a + j} \tag{A.91}$$

$$m_2 = \frac{m_b b + m_0 (c - j)}{b + c - j} \tag{A.92}$$

and the optimal splitting j is determined by solving:

$$\begin{aligned} \min_j M(j) := \min_j \Big[& a(m_a - m_1)^2 + j(m_0 - m_1)^2 + \\ & b(m_b - m_2)^2 + (c - j)(m_0 - m_2)^2 \Big]. \end{aligned} \tag{A.93}$$

Because the second derivative of $M(j)$ is negative:

$$\frac{\partial^2}{\partial j^2} M(j) = -\frac{2a^2 (m_a - m_0)^2}{(a + j)^3} - \frac{2b^2 (m_b - m_0)^2}{(b + c - j)^3} \tag{A.94}$$

the function $M(j)$ is concave and takes its minimum at the boundaries of the domain, that is either at $j = 0$ (corresponding to $s = x_1$) or at $j = c$ (corresponding to $s = x_2$).[59] Therefore, the trees of the random forest do not split within $\mathcal{R}$. Consequently, the random forest model assigns two feature vectors $x_a \neq x_b \in \mathcal{R}$ to the same leaf node in all trees, the corresponding columns of the kernel matrix $\boldsymbol{K} = \{K(x^p, x^q)\}$ are identical, and at least one eigenvalue of the kernel matrix is zero, implying that the random forest

[58] In the limit of $N \to \infty$, the mean demand of the data points in $\mathcal{R}_a$ equals m_a, the mean demand of the data points in $\mathcal{R}_b$ equals m_b, and each subset of the data points in $\mathcal{R}$ has the mean demand m_0.

[59] In case $\frac{\partial^2}{\partial j^2} M(j) = 0$, we obtain $m_1 = m_2 = m_0$ independent of j. Consequently, the sum of residual squares cannot be reduced by splitting within $\mathcal{R}$, and the tree would not split.

kernel is not s.p.d. Therefore, the random forest kernel is not universal, which concludes the proof.

A.7.6 Consistency and Rate of Convergence of kERM

Based on the out-of-sample performance guarantees for kERM presented in Theorems 2.2 and 2.3 for data-independent kernels and the data-dependent random forest kernel, we show that kERM is consistent and we bound the rate at which the risk of $\vec{q}^{\text{kERM}}(\cdot)$ converges to the risk of optimal solution of the function space $\mathcal{F}$.

Definition A.8. *(Following the definition presented in Vapnik 1998, p. 80) An ERM approach is called consistent for a function space $\mathcal{F}$ if $R(\vec{q}_N(\cdot))$ and $R_N(\vec{q}_N(\cdot))$, with $\vec{q}_N(\cdot) = \arg\min_{\vec{q}(\cdot)\in\mathcal{F}} R_N(\vec{q}(\cdot))$, converge to $\inf_{\vec{q}(\cdot)\in\mathcal{F}} R(\vec{q}(\cdot))$ in probability; that is $\forall\epsilon > 0$:*

$$\begin{aligned} \lim_{N\to\infty} P\left(\left|R(\vec{q}_N(\cdot)) - \inf_{\vec{q}(\cdot)\in\mathcal{F}} R(\vec{q}(\cdot))\right| > \epsilon\right) = 0 \\ \lim_{N\to\infty} P\left(\left|R_N(\vec{q}_N(\cdot)) - \inf_{\vec{q}(\cdot)\in\mathcal{F}} R(\vec{q}(\cdot))\right| > \epsilon\right) = 0. \end{aligned} \tag{A.95}$$

Proposition A.6. *Assume a function space $\mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$ with a data-independent bounded kernel $K(\vec{x},\vec{x}) \leq \bar{K}\ \forall\vec{x} \in \mathcal{X}$, or $\mathcal{F} = \mathcal{F}^{RF}$ with the random forest kernel, and $||\vec{b}||_\infty \leq B_C$ and $||q_{U,j}||_K \leq B_U\ \forall j$. Then the kERM approach defined by Problem 2.10 is consistent for the function space $\mathcal{F}$.*

Proof of Proposition A.6: To prove consistency of kERM we derive a bound on the probability that the empirical risk R_N deviates by more than ϵ from the true risk R, prove one-sided convergence and apply Theorem 3.1 of Vapnik (1998). To derive a bound on the probability of deviation in risk, we restate the out-of-sample performance guarantee presented in Theorem 2.2 for all $\vec{q}(\cdot) \in \mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$ as:

$$P\Big(R(\vec{q}(\cdot)) - R_N(\vec{q}(\cdot)) > C_{N,\delta}\Big) \leq \delta, \tag{A.96}$$

where, setting $\delta = 1/N$,

$$C_{N,\delta} \leq 3\bar{l}\sqrt{\frac{\log(2N)}{2N}} + M_{Lip}\left(\frac{2\sqrt{2}IB_C e^2}{\sqrt{\pi}\sqrt{N}} + \frac{2IB_U}{\sqrt{N}}\sqrt{\bar{K}}\right) =: C_N, \tag{A.97}$$

using the bound on the kernel function K. In case of the random forest kernel, we restate the out-of-sample performance guarantee presented in Theorem 2.3 as:

$$P\Big(R(\vec{q}(\cdot)) - R_{N-N_{RF}}(\vec{q}(\cdot)) > C_{N,\delta}\Big) \leq \delta \tag{A.98}$$

and set $\delta = 1/(N - N_{RF})$ to obtain

$$C_{N,\delta} = 3\bar{l}\sqrt{\frac{\log(2(N-N_{RF}))}{2(N-N_{RF})}} + M_{Lip}\left(\frac{2\sqrt{2}IB_C e}{\sqrt{\pi}\sqrt{N-N_{RF}}} + \frac{2IB_U}{\sqrt{N-N_{RF}}}\right) =: C_N \tag{A.99}$$

for all $\vec{q}(\cdot) \in \mathcal{F} = \mathcal{F}^{\text{RF}}$ with N_{RF} constant. Observe that, in both cases, $\lim_{N\to\infty} C_N = 0$ and C_N decreases monotonically for $N \geq 2$ or $N - N_{RF} \geq 2$ respectively. Consequently,

$$\forall \epsilon > 0 \; \exists N_0 \text{ s.t. } \forall N \geq N_0 : C_N < \epsilon. \tag{A.100}$$

Therefore, and because (A.96) holds for all $\vec{q}(\cdot) \in \mathcal{F}$, we obtain

$$\forall \epsilon > 0 \; \exists N_0 \text{ s.t. } P\left(\sup_{\vec{q}(\cdot)\in\mathcal{F}}\Big[R(\vec{q}(\cdot)) - R_N(\vec{q}(\cdot))\Big] > \epsilon\right) \leq \frac{1}{N} \; \forall N \geq N_0, \tag{A.101}$$

and therefore

$$\lim_{N\to\infty, N\geq N_0} P\left(\sup_{\vec{q}(\cdot)\in\mathcal{F}}\Big[R(\vec{q}(\cdot)) - R_N(\vec{q}(\cdot))\Big] > \epsilon\right) = 0 \tag{A.102}$$

for data-independent kernels, and similarly for the random forest kernel. This proves uniform one-sided convergence following Expression 3.14 in Vapnik (1998, p. 90). Therefore, and because $R(\vec{q}(\cdot))$ is bounded as

$$0 \leq R(\vec{q}(\cdot)) \leq \bar{l}, \tag{A.103}$$

based on the bound of the loss function (Lemma 2.1), the results of Theorem 3.1 of Vapnik (1998) apply, stating that the kERM approach is consistent, which concludes the proof.

Proposition A.7 states that the risk of the kernelized solution presented in Theorem 2.1 with a data-independent kernel or with the random forest kernel (2.13) converges at least as fast as $1/\sqrt{N}$ or $1/\sqrt{N - N_{RF}}$ respectively toward the risk of the optimal solution, with probability of at least $1 - 2\delta$.

Proposition A.7. *Assume a function space $\mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$ with a data-independent bounded kernel $K(\vec{x}, \vec{x}) \leq \bar{K}\ \forall \vec{x} \in \mathcal{X}$, or $\mathcal{F} = \mathcal{F}^{RF}$ with the random forest kernel, $||\vec{b}||_\infty \leq B_C$ and $||q_{U,j}||_K \leq B_U\ \forall j$, and let $\delta > 0$. Then the risk of the kERM approach converges with probability of at least $1 - 2\delta$ at least as fast as $1/\sqrt{N}$ or $1/\sqrt{N - N_{RF}}$ respectively toward the risk of the optimal solution $\vec{q}^*(\cdot) = \arg\min_{\vec{q}(\cdot) \in \mathcal{F}} R(\vec{q}(\cdot))$, and there is some $C_\delta < \infty$ such that*

$$\sup_{N>0} \frac{\left|R(\vec{q}^{kERM}(\cdot)) - R(\vec{q}^*(\cdot))\right|}{\frac{1}{\sqrt{N}}} \leq C_\delta \tag{A.104}$$

for data-independent kernels and

$$\sup_{N - N_{RF} > 0} \frac{\left|R(\vec{q}^{kERM}(\cdot)) - R(\vec{q}^*(\cdot))\right|}{\frac{1}{\sqrt{N - N_{RF}}}} \leq C_\delta \tag{A.105}$$

for the random forest kernel.

Proof of Proposition A.7: To bound the rate of convergence of kERM, we bound the deviation in risk of $\vec{q}^{\text{kERM}}(\cdot)$ from the optimal solution and show that it converges with $1/\sqrt{N}$ or $1/\sqrt{N - N_{RF}}$ respectively. In the following we only provide the derivation for the data-independent kernel, where kERM is trained using N data samples—the proof works similarly for the random forest kernel, where kERM is trained on $(N - N_{RF})$ data samples.

Let $\Delta(\vec{q}^{\text{kERM}}(\cdot))$ denote the deviation of $\vec{q}^{\text{kERM}}(\cdot)$ from the optimal solution $\vec{q}^*(\cdot)$ in risk such that

$$\Delta(\vec{q}^{\text{kERM}}(\cdot)) := R(\vec{q}^{\text{kERM}}(\cdot)) - R(\vec{q}^*(\cdot)). \tag{A.106}$$

Because $R_N(\vec{q}^{\,\mathrm{kERM}}(\cdot)) = \min_{\vec{q}(\cdot)\in\mathcal{F}} R_N(\vec{q}(\cdot)) \leq R_N(\vec{q}^{\,*}(\cdot))$, we obtain

$$\begin{aligned}\Delta(\vec{q}^{\,\mathrm{kERM}}(\cdot)) &= R(\vec{q}^{\,\mathrm{kERM}}(\cdot)) - R_N(\vec{q}^{\,\mathrm{kERM}}(\cdot)) + R_N(\vec{q}^{\,\mathrm{kERM}}(\cdot)) - R(\vec{q}^{\,*}(\cdot)) \\ &\leq \left[R(\vec{q}^{\,\mathrm{kERM}}(\cdot)) - R_N(\vec{q}^{\,\mathrm{kERM}}(\cdot))\right] + \left[R_N(\vec{q}^{\,*}(\cdot)) - R(\vec{q}^{\,*}(\cdot))\right].\end{aligned} \tag{A.107}$$

While $\left[R(\vec{q}^{\,\mathrm{kERM}}(\cdot)) - R_N(\vec{q}^{\,\mathrm{kERM}}(\cdot))\right]$ is bounded by some $C_{N,\delta}$ (Theorems 2.2 and 2.3), we derive a bound on the term $[R_N(\vec{q}^{\,*}(\cdot)) - R(\vec{q}^{\,*}(\cdot))]$ in the following.

Because the loss function is bounded (Lemma 2.1), we can apply Hoeffding's inequality for $\epsilon > 0$ (similar as in Expressions 5.7 and 5.8 in Vapnik 1998, p. 186) and obtain:

$$P\left(R_N(\vec{q}^{\,*}(\cdot)) - R(\vec{q}^{\,*}(\cdot)) > \epsilon\right) \leq exp\left(\frac{-2\epsilon^2 N}{\bar{l}^2}\right), \tag{A.108}$$

from which we derive, by defining $\delta := exp\left(\frac{-2\epsilon^2 N}{\bar{l}^2}\right)$:

$$P\left(R_N(\vec{q}^{\,*}(\cdot)) - R(\vec{q}^{\,*}(\cdot)) \leq \bar{l}\sqrt{\frac{\log(1/\delta)}{2N}}\right) \geq 1-\delta. \tag{A.109}$$

Therefore, we obtain for the deviation in risk with probability of at least $1-2\delta$:

$$\Delta(\vec{q}^{\,\mathrm{kERM}}(\cdot)) \leq C_{N,\delta} + \bar{l}\sqrt{\frac{\log(1/\delta)}{2N}} \leq \frac{C_\delta}{\sqrt{N}}, \tag{A.110}$$

for some C_δ independent of N, where we used that

$$\begin{aligned}C_{N,\delta} &= 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} + M_{Lip}\left(\frac{2\sqrt{2}IB_C e^2}{\sqrt{\pi}\sqrt{N}} + \frac{2IB_U}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{n=1}^{N} K(\vec{x}^n, \vec{x}^n)}\right) \\ &\leq 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} + M_{Lip}\left(\frac{2\sqrt{2}IB_C e^2}{\sqrt{\pi}\sqrt{N}} + \frac{2IB_U}{\sqrt{N}}\sqrt{\bar{K}}\right)\end{aligned} \tag{A.111}$$

for all $\vec{q}(\cdot) \in \mathcal{F} = \mathcal{F}_U + \mathcal{F}_C$ (Theorem 2.2), using the bound on the kernel function K.

Because $\vec{q}^{\,*}(\cdot) = \arg\min_{\vec{q}(\cdot)\in\mathcal{F}} R(\vec{q}(\cdot))$, we know $\Delta(\vec{q}^{\,\mathrm{kERM}}(\cdot)) \geq 0$. There-

fore, and because C_δ is independent of N, we obtain for all $N > 0$:

$$\frac{|\Delta(\vec{q}^{\,\mathrm{kERM}}(\cdot))|}{\frac{1}{\sqrt{N}}} \leq C_\delta < \infty \tag{A.112}$$

with probability of at least $1 - 2\delta$, which concludes the proof.

A.8 Additional Numerical Analyses

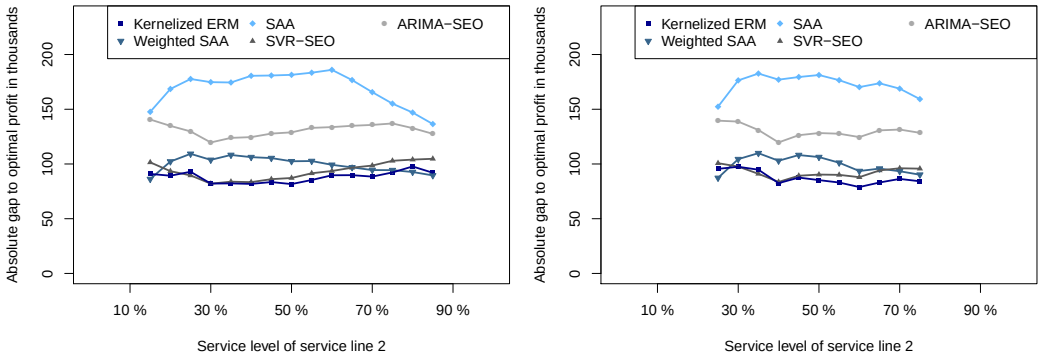
A.8.1 Variation of the Service Level under Heterogeneous Service Levels

Building on the numerical experiments presented in Section 2.6.5, in which we study the performance of the prescriptive approaches for various homogeneous service level settings, this section presents an analysis for varying heterogeneous service levels. The cost parameters used to induce the service level variations are depicted in Table A.1, while demand and feature data are the same as used in Section 2.6.

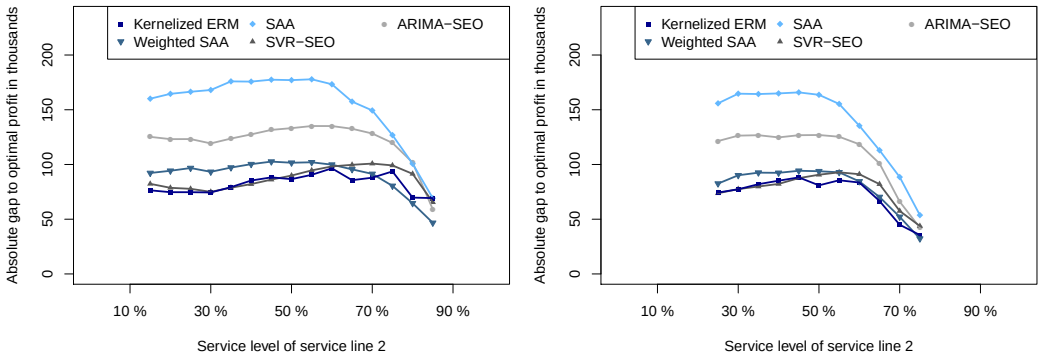
Table A.1: Parameter settings for the service level variation with heterogeneous service levels.

Figure	f	v	p	c	$a_{i,i}$	**SL**	α
A.7a	2700..710 1010..178 390..26	40 30 20	600 260 120	10 7.5 5	570 238 105	5%..75% 15%..85% 25%..95%	40%
A.7b	2700..1280 890..297 289..26	40 30 20	600 260 120	10 7.5 5	570 238 105	5%..55% 25%..75% 45%..95%	40%
A.7c	2130..142 1010..178 500..131	40 30 20	600 260 120	10 7.5 5	570 238 105	25%..95% 15%..85% 5%..75%	40%
A.7d	1560..142 890..297 500..236	40 30 20	600 260 120	10 7.5 5	570 238 105	45%..95% 25%..75% 5%..55%	40%

Figure A.7 depicts the absolute gab to optimal profit for settings with service line 1 having the lowest service level (top line, Figures A.7a and A.7b), and settings in which service line 1 has the highest service level (bottom line, Figures A.7c and A.7d).



(a) $SL_1 = SL_2 - 10\%$, $SL_3 = SL_2 + 10\%$ (b) $SL_1 = SL_2 - 20\%$, $SL_3 = SL_2 + 20\%$



(c) $SL_1 = SL_2 + 10\%$, $SL_3 = SL_2 - 10\%$ (d) $SL_1 = SL_2 + 20\%$, $SL_3 = SL_2 - 20\%$

Figure A.7: Variation of the service level under heterogeneous service levels.

Overall we observe similar effects as in Section 2.6.5, e.g., that the performance of SAA is similar to the performance of the prescriptive approaches for very high service levels—especially when service line 1 exhibits a high service level, due to the upgrading structure (service line 1 capacity can be employed to provide service on all service lines).

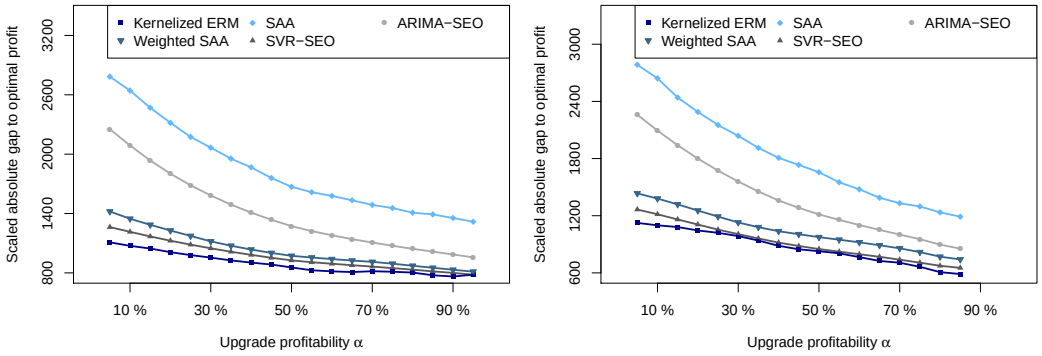
A.8.2 Variation of the Upgrade Profitability under Heterogeneous Service Levels

In this section we study the performance of the prescriptive approaches for varying upgrade profitabilities under heterogeneous service levels. The cost parameters used to induce the upgrade profitability variations are depicted in Table A.2, while demand and feature data are, again, the same as used in Section 2.6.

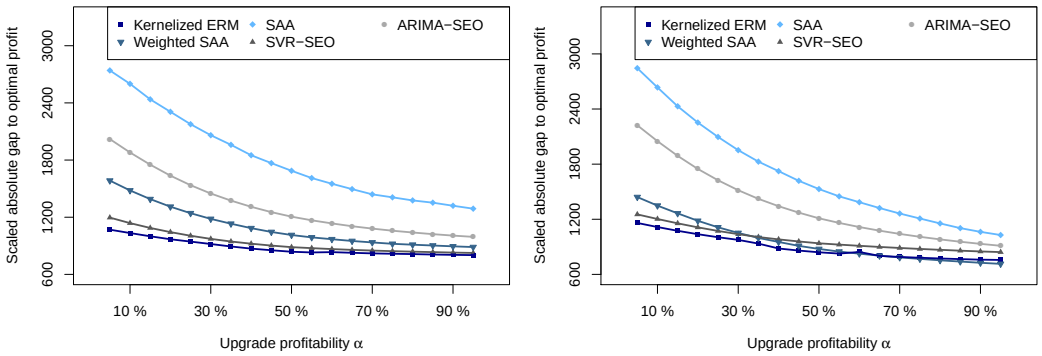
Table A.2: Parameter settings for the upgrade profitability variation with heterogeneous service levels.

Figure	f	v	p	c	$a_{i,i}$	**SL**	α
A.8a	75600..189 3800..200 210	40 30 20	37830..125 1920..123 120	10 7.5 5	37800..95 1900..100 105	60% 60% 60%	5% .. 95%
A.8b	75600..240 4750..280 315	40 30 20	37830..150 1920..134 120	10 7.5 5	37800..120 1900..112 105	60% 50% 40%	5% .. 85%
A.8c	132300..332 4750..250 158	40 30 20	37830..125 1920..123 120	10 7.5 5	37800..95 1900..100 105	30% 50% 70%	5% .. 95%
A.8d	94500..237 2850..150 53	40 30 20	37830..125 1920..123 120	10 7.5 5	37800..95 1900..100 105	50% 70% 90%	5% .. 95%

Because the marginal profits vary significantly across α, Figure A.8 plots the scaled absolute gap to optimal profit, similar as in Section 2.6.5. Comparing all plots of Figure A.8, we observe that the impact of the upgrade profitability on the performance of the approaches remains structurally similar for all service level regimes. Only the ordering of the approaches changes dependent on the service level regime, as seen before: for medium-range service levels, kERM outperforms wSAA independent of the upgrade profitability and for high service levels (Figure A.8d) (and a high upgrade profitability) wSAA leads to better performance.



(a) $SL_1 = 60\%$, $SL_2 = 60\%$, $SL_3 = 60\%$ (b) $SL_1 = 60\%$, $SL_2 = 50\%$, $SL_3 = 40\%$



(c) $SL_1 = 30\%$, $SL_2 = 50\%$, $SL_3 = 70\%$ (d) $SL_1 = 50\%$, $SL_2 = 70\%$, $SL_3 = 90\%$

Figure A.8: Variation of the upgrade profitability α in various service level settings.

A.8.3 Statistical Confidence of Prescription Performance

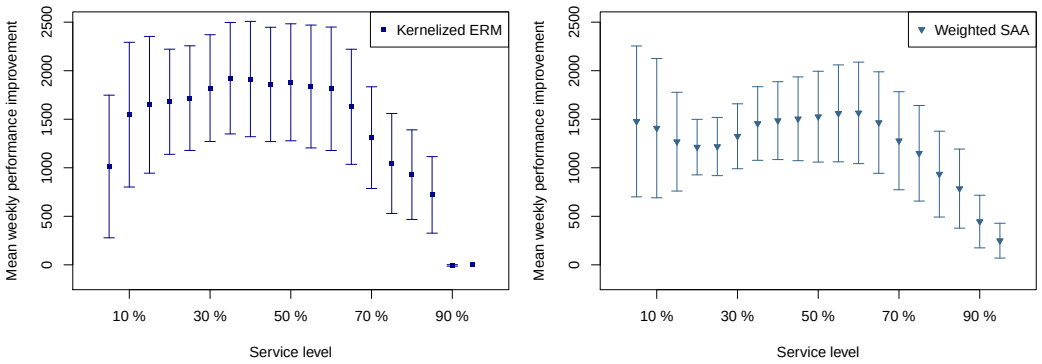
While our numerical evaluation shows a significant performance improvement from the feature-less approach SAA to the prescriptive approaches wSAA and kERM, a natural question is how certain we are of this improvement.

In our experiments we evaluate the prescriptive approaches on a test set containing 52 data points, based on which we can calculate a performance difference. While the underlying distribution of the performance improvement is unknown, the sample mean follows a normal distribution for large enough sample sizes, e.g. $N_{test} > 30$ (see Hogg et al. 2015, p. 303), and we can use the Student's t-distribution, introduced by W. S. Gosset, to estimate the approximate confidence interval of the true mean of the improvement distribution

(see Section 5.6 in Precht et al. 2005 for details).

Let Δ_k with $k = 1...52$ be the observed performance differences between the prescriptive approaches and SAA for $N_{test} = 52$ data samples and $\bar{\Delta}$ the mean performance improvement, and let $s^2_{test} = 1/(N_{test} - 1) \sum_{k=1}^{N_{test}} (\Delta_k - \bar{\Delta})^2$ be the sample variance.

Then, for a significance level of $p = 1 - \alpha$, we estimate the approximate confidence interval $[\bar{\Delta} - d_{Delta}, \bar{\Delta} + d_{Delta}]$ with $d_{Delta} = t \frac{s_{test}}{\sqrt{N_{test}}}$ and t the $1 - \alpha/2$-fractile of the t-distribution with $1 - N_{test}$ degrees of freedom for the true mean of the performance improvement (see Section 5.6 in Precht et al. 2005).



(a) Kernelized ERM. (b) Weighted SAA.

Figure A.9: Mean performance improvement of prescriptive approaches over SAA including 95% approximate confidence interval.

Following the service level variation setting, as described in Section 2.6.5, we estimate the $p = 95\%$ approximate confidence intervals for the true mean of the performance improvement, using $t = 2.0076$ for 51 degrees of freedom. As depicted in Figure A.9, the lower bound of this interval is above zero for all service levels across both approaches, with the exception for kERM with a very large service level of 90%, where the mean performance difference between kERM and SAA is close to zero. These results provide support for a statistically significant performance improvement of the prescriptive approaches over SAA for almost all service levels.

We use the same approach to estimate the $p = 95\%$ approximate confi-

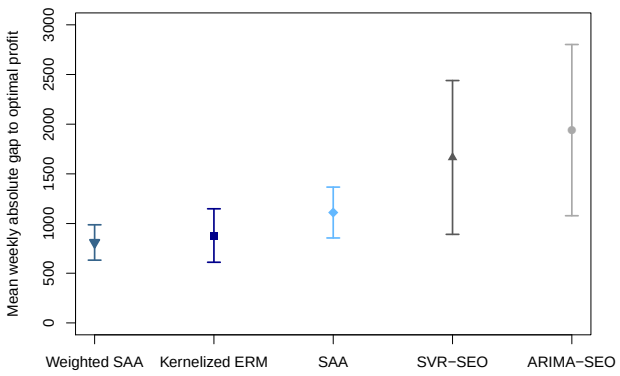


Figure A.10: Mean weekly absolute gap to optimal profit for real-world case including 95% approximate confidence interval.

dence intervals for the true mean of the weekly absolute gap to optimal profit for the real-world application presented in Section 2.6.4. The results, depicted in Figure A.10, show a significantly larger confidence interval for the two parametric approaches (ARIMA-SEO, SVR-SEO) compared to the prescriptive approaches and SAA.

B Appendix of Chapter 3

B.1 Proofs

Proof of Proposition 3.1

The cost function of the AMSSP consists of a linear expression that is convex in q_s ($\frac{1}{T}\sum_{s=1}^{S} c_q(t_{s+1} - t_s)q_s$) and the expectation of the queue length $\mathbb{E}[M_T]$ at the end of the final period. Because M_T is recursively defined by $M_t = (M_{t-1} + D_t - q_s)^+$, which is convex in q_s for each D_t because of the linearity in q_s and the convexity of the $(\cdot)^+ = \max(\cdot, 0)$ function, the expectation value $\mathbb{E}[M_T]$ is also convex in q_s. Consequently, the cost function is a sum of two convex functions and therefore jointly convex in $\vec{q}$.

Proof of Proposition 3.2

To derive the expression for $\mathbb{E}[M_T]$ stated in Proposition 3.2, we first provide, in Proposition B.1, a method to recursively calculate the distribution of M_T and then apply this result to $T = 2$ independent normally distributed demands.

Proposition B.1. *Assume demands D_t independently distributed for all t; then the distribution function $f_{M_T}(m)$ of M_T can be calculated by recursively solving:*

$$\begin{aligned} f_{M_0}(m) &= \delta(m - M_0) \\ f_{M_t}(m) &= \delta(m) \int_{\nu_1=-\infty}^{\nu_1=0} \int_{\nu_2=-\infty}^{\nu_2=\infty} f_{M_{t-1}}(\nu_2) f_{D_t}(\nu_1 + q_{s(t)} - \nu_2) d\nu_2 d\nu_1 \\ &\quad + \begin{cases} 0 \text{ for } m < 0 \\ \int_{-\infty}^{\infty} f_{M_{t-1}}(\nu) f_{D_t}(m + q_{s(t)} - \nu) d\nu \text{ for } m \geq 0, \end{cases} \end{aligned} \tag{B.1}$$

where $\delta(m)$ is the Dirac delta function and $s(t) = s : t \in [t_s, t_{s+1})$.

Proof of Proposition B.1: To prove the stated expression for the recursive calculation of the distribution of M_T, we start from the recursive expression for M_t stated in the AMSSP:

$$M_t = (M_{t-1} + D_t - q_s)^+. \tag{B.2}$$

The calculation of this recursion can be separated into three steps by introducing the variables $\tilde{M}_t$ and $\hat{M}_t$ as

$$\begin{aligned} \tilde{M}_t &= M_{t-1} + D_t, \\ \hat{M}_t &= \tilde{M}_t - q_s, \\ M_t &= (\hat{M}_t)^+. \end{aligned} \tag{B.3}$$

For the distribution functions of the variables $\tilde{M}_t$ and $\hat{M}_t$ we obtain:

$$f_{\tilde{M}_t}(\tilde{m}) = \int_{-\infty}^{\infty} f_{M_{t-1}}(\nu) f_{D_t}(\tilde{m} - \nu) d\nu, \tag{B.4}$$

$$f_{\hat{M}_t}(\hat{m}) = f_{\tilde{M}_t}(\hat{m} + q_s), \tag{B.5}$$

where we used Theorem 5.2.9 in Casella and Berger (2002) for Equation B.4 and assumed M_{t-1} and D_t to be independent continuous random variables. To estimate the distribution function of M_t, we first estimate the cumulative distribution function (cdf) $F_{M_t}(m)$ using Expression 2.1.4 provided in Casella and Berger (2002):

$$F_{M_t}(m) = \int_{\{\nu \in \mathbb{R}:\ (\nu)^+ \leq m\}} f_{\hat{M}_t}(\nu) d\nu = \begin{cases} 0 \text{ for } m < 0 \\ F_{\hat{M}_t}(m) \text{ for } m \geq 0 \end{cases}, \tag{B.6}$$

because

$$\{\nu \in \mathbb{R} :\ (\nu)^+ \leq m\} = \begin{cases} \emptyset \text{ for } m < 0 \\ \{\nu \in \mathbb{R} : -\infty < \nu \leq m\} \text{ for } m \geq 0. \end{cases} \tag{B.7}$$

This cdf can be expressed as follows, using the Heaviside function $\Theta(m)$:

$$F_{M_t}(m) = F_{\hat{M}_t}(0) \cdot \Theta(m) + \begin{cases} 0 \text{ for } m < 0 \\ F_{\hat{M}_t}(m) - F_{\hat{M}_t}(0) \text{ for } m \geq 0. \end{cases} \tag{B.8}$$

Then, using the Dirac Delta function $\delta(m) = d\Theta(m)/dm$, the distribution function $f_{M_t}(m)$ is obtained as:

$$\begin{aligned} f_{M_t}(m) = \frac{dF_{M_t}(m)}{dm} &= F_{\hat{M}_t}(0) \cdot \delta(m) + \begin{cases} 0 \text{ for } m < 0 \\ f_{\hat{M}_t}(m) \text{ for } m \geq 0 \end{cases} \\ &= \delta(m) \int_{-\infty}^{0} f_{\hat{M}_t}(\nu) d\nu + \begin{cases} 0 \text{ for } m < 0 \\ f_{\hat{M}_t}(m) \text{ for } m \geq 0, \end{cases} \end{aligned} \tag{B.9}$$

with $f_{M_0}(m) = \delta(m - M_0)$, from which we obtain the expression to be proven.

We apply this result (Equation B.1) to the case of $T = 2$ independent, normally distributed demands D_1, D_2 with $M_0 = 0$, such that $f_{M_0}(m) = \delta(m)$, and obtain for M_1:

$$\begin{aligned} f_{M_1}(m) &= \delta(m) \int_{\nu_1=-\infty}^{\nu_1=0} \int_{\nu_2=-\infty}^{\nu_2=\infty} \delta(\nu_2) f_{D_1}(\nu_1 + q_1 - \nu_2) d\nu_2 d\nu_1 \\ &\quad + \begin{cases} 0 \text{ for } m < 0 \\ \int_{-\infty}^{\infty} \delta(\nu) f_{D_1}(m + q_1 - \nu) d\nu \text{ for } m \geq 0 \end{cases} \\ &= \delta(m) \int_{-\infty}^{0} f_{D_1}(\nu_1 + q_1) d\nu_1 + \begin{cases} 0 \text{ for } m < 0 \\ f_{D_1}(m + q_1) \text{ for } m \geq 0 \end{cases} \\ &= \delta(m) F_{D_1}(q_1) + \begin{cases} 0 \text{ for } m < 0 \\ \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(m+q_1-\mu_1)^2}{2\sigma_1^2}\right] \text{ for } m \geq 0. \end{cases} \end{aligned} \tag{B.10}$$

Applying Equation B.1 again, we obtain the distribution function for M_2 as:

$$\begin{aligned} f_{M_2}(m) &= \delta(m) \int_{\nu_1=-\infty}^{\nu_1=0} \int_{\nu_2=-\infty}^{\nu_2=\infty} f_{M_1}(\nu_2) f_{D_2}(\nu_1 + q_2 - \nu_2) d\nu_2 d\nu_1 \\ &\quad + \begin{cases} 0 \text{ for } m < 0 \\ \int_{-\infty}^{\infty} f_{M_1}(\nu) f_{D_2}(m + q_2 - \nu) d\nu \text{ for } m \geq 0 \end{cases} \\ &= \delta(m) a_{M_2} + \begin{cases} 0 \text{ for } m < 0 \\ F_{D_1}(q_1) f_{D_2}(m + q_2) + \frac{\exp\left[-\frac{(m-\mu_1-\mu_2+q_1+q_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]}{2\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} \\ \cdot \left(1 + \text{Erf}\left[\frac{\sigma_1^2(m-\mu_2+q_2)+\sigma_2^2(\mu_1-q_1)}{\sqrt{2}\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}\right]\right) \text{ for } m \geq 0, \end{cases} \end{aligned} \tag{B.11}$$

where $a_{M_2} = \int_{\nu_1=-\infty}^{\nu_1=0} \int_{\nu_2=-\infty}^{\nu_2=\infty} f_{M_1}(\nu_2) f_{D_2}(\nu_1 + q_2 - \nu_2) d\nu_2 d\nu_1$ is irrelevant for the expectation value because of the factor $\delta(m)$. We calculate the expectation value of $M_T = M_2$ as:

$$\begin{aligned} \mathbb{E}[M_T] &= \int_{-\infty}^{\infty} m f_{M_2}(m) dm \\ &= \int_0^{\infty} m \Bigg[F_{D_1}(q_1) f_{D_2}(m + q_2) + \frac{\exp\left[-\frac{(m-\mu_1-\mu_2+q_1+q_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]}{2\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} \\ &\qquad \cdot \left(1 + \text{Erf}\left[\frac{\sigma_1^2(m-\mu_2+q_2)+\sigma_2^2(\mu_1-q_1)}{\sqrt{2}\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}\right]\right) \Bigg] dm, \end{aligned} \tag{B.12}$$

which concludes the proof.

Proof of Proposition 3.3

To prove the equivalency of the linearized version of the cost function, we first apply a linear version of the $(\cdot)^+$ function, and then solve the iterative approach by global minimization over the queue length in each period. The recursive equation of the queue length m_t

$$m_t = (m_{t-1} + d_t - q_s)^+ \tag{B.13}$$

can be expressed as optimization problem

$$\begin{aligned} m_t = \min_{u_t \in \mathbb{R}} \; & u_t \\ \text{s.t. } & u_t \geq m_{t-1} + d_t - q_s \\ & u_t \geq 0, \end{aligned} \tag{B.14}$$

where $s : t \in [t_s, t_{s+1})$ and with the final queue length $m_T = \min_{u_T \in \mathbb{R}} u_T$. While this expression requires iterative minimization, we simplify the calculation by showing equivalency to a single global minimization as:

$$\begin{aligned} v_T^* = \min_{\{v_t\} \in \mathbb{R}^T} \; & v_T \\ \text{s.t. } & v_t \geq v_{t-1} + d_t - q_s \; \forall t \in [t_s, t_{s+1}) \; \forall s = 1...S \\ & v_t \geq 0 \; \forall t \\ & v_0 = M_0. \end{aligned} \tag{B.15}$$

To prove equivalency, we need to show that $\{u_t\}$ that iteratively solves Problem B.14 is a feasible solution to Problem B.15, and that it is an optimal solution. The feasibility of $\{u_t\}$ is obvious due to identical constraints. Because $m_t = \min_{u_t} u_t =: u_t^*$ at each time step t, we know that for any other feasible solution $\{\tilde{v}_t\}$ to Problem B.15 we have $\tilde{v}_t \geq u_t^*$, leading to $\tilde{v}_{t+1} \geq u_{t+1}^*$ due to the first constraint. Therefore, $\tilde{v}_T \geq u_T^*$ by induction, from which we conclude that $\{u_t\}$ is an optimal solution to Problem B.15, and $m_T = v_T^* = u_T^*$, from which we derive the expression to be proven.

Proof of Proposition 3.4

The linearized cost function (Equation 3.11) consists of the convex expression $\frac{1}{T}\sum_{s=1}^{S} c_q(t_{s+1} - t_s)q_s$ (see Proposition 3.1) and $\min_{\{m_t\}} c_2 m_T$. The minimization $\min_{\{m_t\}} c_2 m_T$ is determined by solving a linear program with q_s determining the right-hand side, and therefore convex. Consequently, the linearized cost function $C(\vec{q}, \vec{d})$ is a sum of convex functions and, therefore, jointly convex in $\vec{q}$.

Proof of Proposition 3.5

To prove the stated expression for the kERM approach, we first provide an expression for a linear function space by deriving the primal Lagrangian, showing strong duality, and expressing the problem using the Lagrange dual function, and then kernelize the solution.

Assume a space of linear functions $\mathcal{F} = \{\vec{q}(\cdot) : \vec{q}(\vec{x}) = \boldsymbol{W}\vec{x} - \vec{b}\}$ with $\boldsymbol{W} = (\vec{w}_1, \vec{w}_2, ..., \vec{w}_S)$ and $\vec{b} = (b_1, b_2, ..., b_S)$. Then, the ERM approach can be stated as

$$\begin{aligned}
\min_{\boldsymbol{W},\vec{b},\{m_t^n\}\in\mathbb{R}^{T\times N}} & \sum_{s=1}^{S} \lambda_s ||\vec{w}_s||^2 + \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{T}\sum_{s=1}^{S} c_q(t_{s+1}-t_s)(\vec{w}_s\cdot\vec{x}^n - b_s) + c_2 m_T^n\right] \\
\text{s.t. } & m_t^n \geq m_{t-1}^n + d_t^n - (\vec{w}_s\cdot\vec{x}^n - b_s)\ \forall n\ \forall t\in[t_s, t_{s+1})\ \forall s = 1...S \\
& m_t^n \geq 0\ \forall n, t \\
& m_0^n = M_0\ \forall n \\
& \vec{w}_s\cdot\vec{x}^n - b_s \geq 0\ \forall n, s,
\end{aligned} \tag{B.16}$$

where we introduced a regularization term $\sum_{s=1}^{S}\lambda_s||\vec{w}_s||^2$ to prevent overfitting, similar as in Notz and Pibernik (2021).

Because the objective function of this minimization problem is a sum of convex and linear functions, it is convex in $\boldsymbol{W}$, $\vec{b}$ and $\{m_t^n\}$. All constraints are affine in $\boldsymbol{W}$, $\vec{b}$ and $\{m_t^n\}$, and $\boldsymbol{W} = 0$, $\vec{b} = 0$, $m_t^n = M_0 + \sum_{\tau=0}^{t} d_\tau^n$ is a feasible solution and a relative interior point of the domain of definition $\mathbb{R}^\nu$ with $\nu = S \times \dim\mathcal{X} + S + T \times N$. Therefore, the Slater condition is fulfilled and strong duality holds (see Section 5.2.3 in Boyd and Vandenberghe 2004).

Let α_t^n, β_t^n, γ^n, δ_s^n be Lagrangian multipliers, then we derive the primal Lagrangian as

$$\begin{aligned}
L_{\text{primal}} = & \sum_{s=1}^{S}\lambda_s||\vec{w}_s||^2 + \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{T}\sum_{s=1}^{S} c_q\Delta t_s(\vec{w}_s\cdot\vec{x}^n - b_s) + c_2 m_T^n\right] \\
& + \sum_{n,t}\alpha_t^n\left(m_{t-1}^n + d_t^n - (\vec{w}_{s(t)}\cdot\vec{x}^n - b_{s(t)}) - m_t^n\right) - \sum_{n,t}\beta_t^n m_t^n \\
& + \sum_{n}\gamma^n\left(m_0^n - M_0\right) - \sum_{n,s}\delta_s^n\left(\vec{w}_s\cdot\vec{x}^n - b_s\right),
\end{aligned} \tag{B.17}$$

where $t = 1...T$, $n = 1...N$, $\Delta t_s := t_{s+1} - t_s$, and $s(t) = s : t \in [t_s, t_{s+1})$.

Because strong duality holds, we use the Karush-Kuhn-Tucker (KKT) conditions, which provide necessary and sufficient conditions for optimality, to simplify the expression for the primal Lagrangian (Boyd and Vandenberghe 2004, p. 244). The KKT conditions state:

$$\begin{aligned}
\frac{\partial L_{\text{primal}}}{\partial \vec{w}_s} &= 2\lambda_s \vec{w}_s + \sum_{n=1}^{N} \left(\frac{c_q \Delta t_s}{NT} - \sum_{t \in [t_s, t_{s+1})} \alpha_t^n - \delta_s^n \right) \vec{x}^n = 0 \\
\frac{\partial L_{\text{primal}}}{\partial b_s} &= \sum_{n=1}^{N} \left(-\frac{c_q \Delta t_s}{NT} + \sum_{t \in [t_s, t_{s+1})} \alpha_t^n + \delta_s^n \right) = 0 \\
\text{for } 1 \leq t < T\text{:} \ \frac{\partial L_{\text{primal}}}{\partial m_t^n} &= \alpha_{t+1}^n - \alpha_t^n - \beta_t^n = 0 \\
\text{for } t = 0\text{:} \ \frac{\partial L_{\text{primal}}}{\partial m_0^n} &= \alpha_1^n + \gamma^n = 0 \\
\text{for } t = T\text{:} \ \frac{\partial L_{\text{primal}}}{\partial m_T^n} &= -\alpha_T^n - \beta_T^n + \frac{c_2}{N} = 0.
\end{aligned} \tag{B.18}$$

We define

$$u_s^n := \frac{1}{2\lambda_s} \left(\sum_{t \in [t_s, t_{s+1})} \alpha_t^n + \delta_s^n - \frac{c_q \Delta t_s}{NT} \right) \tag{B.19}$$

and obtain

$$\begin{aligned}
\vec{w}_s &= \sum_{n=1}^{N} u_s^n \vec{x}^n, \\
0 &= \sum_{n=1}^{N} u_s^n \ \forall s,
\end{aligned} \tag{B.20}$$

and

$$\begin{aligned}
\text{for } 1 \leq t < T\text{:} \ \alpha_{t+1}^n &= \alpha_t^n + \beta_t^n \\
\text{for } t = 0\text{:} \ \gamma^n &= -\alpha_1^n \\
\text{for } t = T\text{:} \ \alpha_T^n + \beta_T^n &= \frac{c_2}{N}.
\end{aligned} \tag{B.21}$$

We can therefore state the primal Lagrangian as

$$L_{\text{primal}} = -\sum_{s=1}^{S} \lambda_s \sum_{p=1}^{N} \sum_{q=1}^{N} (u_s^p u_s^q (\vec{x}^p \cdot \vec{x}^q)) + \sum_{n,t} \alpha_t^n d_t^n + \sum_{n} \alpha_1^n M_0. \tag{B.22}$$

The primal Lagrangian L_{primal} is independent of the primal variables, therefore, we obtain for the dual Lagrangian:

$$L_{\text{dual}} = \inf_{\mathbf{w},\vec{b},\{m_t^n\}} L_{\text{primal}} = L_{\text{primal}}, \tag{B.23}$$

from which we derive for the dual problem:

$$\begin{aligned} &\max_{\{\alpha_t^n\},\{\delta_s^n\}} L_{\text{dual}} \\ &\quad \text{s.t. } \alpha_t^n,\ \delta_s^n \geq 0\ \forall t,n,s \\ &\qquad \alpha_{t+1}^n \geq \alpha_t^n\ \forall t,n \\ &\qquad \alpha_T^n \leq \frac{c_2}{N}\ \forall n. \end{aligned} \tag{B.24}$$

Because Problem B.24 (with the dual Lagrangian as in Equation B.22) depends on the feature vectors $\vec{x}^n$ only in the form of the scalar product between two feature vectors, we can employ the kernel trick and replace the scalar product by a kernel function:

$$(\vec{x}^p \cdot \vec{x}^q) \rightarrow K(\vec{x}^p, \vec{x}^q), \tag{B.25}$$

which allows us to solve the ERM approach over non-linear reproducing kernel Hilbert spaces.

The resulting expression is

$$\begin{aligned} &\max_{\{\alpha_t^n\},\{\delta_s^n\}} -\sum_{s=1}^{S} \lambda_s \sum_{p=1}^{N}\sum_{q=1}^{N} (u_s^p u_s^q K(\vec{x}^p, \vec{x}^q)) + \sum_{n,t} \alpha_t^n d_t^n + \sum_n \alpha_1^n M_0 \\ &\quad \text{s.t. } \alpha_t^n,\ \delta_s^n \geq 0\ \forall t,n,s \\ &\qquad \alpha_{t+1}^n \geq \alpha_t^n\ \forall t,n \\ &\qquad \alpha_T^n \leq \frac{c_2}{N}\ \forall n, \end{aligned} \tag{B.26}$$

with

$$\vec{q}^{\,\text{kERM}}(\vec{x}) = \boldsymbol{W}\vec{x} - \vec{b} = \sum_{n=1}^{N} \vec{u}^{\,n} K(\vec{x}^{\,n}, \vec{x}) - \vec{b}, \tag{B.27}$$

where $\vec{u}^{\,n} = (u_1^n, u_2^n, ..., u_S^n)$, which concludes the proof.

B.2 Detailed Description of Features and Importance Analysis

This section provides further details on the 142 features used in our numerical analyses presented in Section 3.5 and on their individual importance. As described in Section 3.5.2, the features are constructed based on the date, public holidays, lagged demands and related to the case company's processes. In the first group, we constructed date-based features describing

- the year (2014-2017),
- the half of the year (1-2),
- the quarter of the year (1-4),
- the month of the year (1-12),
- the day of the month (1-31),
- the day of the month integer-divided by 7 (0-4)[60]
- the day of the week (1-6),
- the day of the week as indicator features (0-1),
- the day of the quarter year (1-92),
- the day of the year (1-366),
- the week number (1-53) by US and ISO standards,
- the week number within the month (1-5),
- the week number modulo 2, 3, or 4 (0-1, 0-2, 0-3), and
- a time index as continuously increasing value (number of seconds after 1970, in the range 1454976000-1509753600).

[60]This represents a different (non-calendar-week) measure for the week of the month.

Most of these 22 date-based features were constructed using the *timetk* package in R.
For the second group, we constructed 12 features indicating the relation to public holidays:

- a public holiday (in Germany or US) is 1-3 days before the day in focus (0-1), and
- a public holiday (in Germany or US) 1-3 days after the day in focus (0-1).

In the third group, we constructed 36 features representing lagged demand as

- 1-7 days lagged demand as totals for first and second shift,
- 14, 21, and 28 days lagged demand as totals for first and second shift,
- 1-7 days lagged demand as daily totals,
- 14, 21, and 28 days lagged demand as daily totals,
- sum of 1-3 days lagged demand as totals for first and second shift,
- sum of 1-7 days lagged demand as totals for first and second shift,
- sum of 1-3 days lagged demand as daily totals, and
- sum of 1-7 days lagged demand as daily totals.

For the fourth group, 72 features have been created based on the company's processes, including

- demand for which a process ID was created 1-29 days before the day in focus,
- summed demand for which a process ID was created 1-29 days before the day in focus,
- expected demand based on contracted delivery date,

- expected demand based on updated (agreed) delivery date,
- weekday average of expected demand for contracted and updated delivery date,
- percentage difference of actuals to weekday average for contracted and updated delivery date,
- parts shipped 1-7 days before the day in focus by four country groups,
- sum of parts shipped between 8-29 days before the day in focus by four country groups, and
- sum of parts shipped between 1-29 days before the day in focus by four country groups.

Following this procedure, we created a total of 142 features.

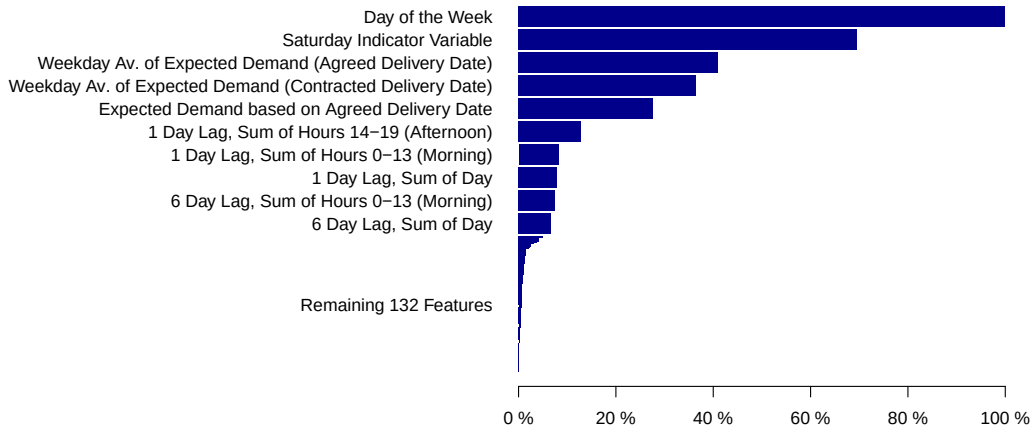


Figure B.1: Feature importance, measured as a decrease in node impurity in the random forest model.

The importance of several classes of features has been presented in Section 3.5.5. Figure B.1 shows the importance of individual features, measured as a decrease in node impurity based on the residual sum of squares in the random forest model used for wSAA or kERM. The feature that describes the day of the week is of largest importance, however, all four most important features are—at least to some extent—correlated with the weekday. It can,

therefore, be expected that the random forest model uses all of these features to distinguish between Saturdays and the rest of the week (which form the two groups between which demand differs most significantly).

B.3 Detailed Description of Approaches for Numerical Evaluation

This section describes in more detail the approaches used in our numerical analyses.

B.3.1 Weighted SAA

For the wSAA approach, we first train a random forest model using the *caret* R-package (Kuhn 2008) and cross validation parameter tuning with 10...142 variables to be selected for each tree. Based on this random forest model we implement (3.8) to calculate the weights $w_n^{\mathrm{RF}}(\vec{x}^j)$ between all training data samples ($n = 1...N$) and all test data samples ($j = 1...N_{test}$). Finally, we adapt the wSAA approach (Algorithm 3.1) to the specific capacity planning problem (3.17):

$$\begin{aligned}\vec{q}^{\,\mathrm{wSAA}}(\vec{x}) = \arg\min_{\vec{q}\in\mathcal{Q}} \min_{\{m_t^n\}\in\mathbb{R}^{T\times N}} \sum_{n=1}^{N} w_n(\vec{x}) \Big[c_q\Big((t_2-t_1)q_1 + (t_3-t_2)q_2\Big) \\ + c_2 m_T^n + c_3 m_{14}^n \Big] \\ \text{s.t. } m_t^n \geq \begin{cases} m_{t-1}^n + d_t^n \text{ for } 1 \leq t \leq 6 \ \forall n \\ m_{t-1}^n + d_t^n - q_1 \text{ for } 7 \leq t \leq 14 \ \forall n \\ m_{t-1}^n + d_t^n - q_2 \text{ for } 15 \leq t \leq 20 \ \forall n \end{cases} \\ m_t^n \geq 0 \ \forall t, n \\ m_0^n = 0 \ \forall n,\end{aligned} \tag{B.28}$$

and solve the resulting problem using *Gurobi Optimizer* for each day of the test period.

B.3.2 Kernelized ERM

Similar as in Proposition 3.3, we derive a linearized cost function for the specific capacity planning problem (3.17):

$$\begin{aligned} C(\vec{q},\vec{d}) =& c_q\Big((t_2-t_1)q_1+(t_3-t_2)q_2\Big) \\ &+ \min_{\{m_t\}\in\mathbb{R}^T} (c_2 m_T + c_3 m_{14}) \\ &\text{s.t. } m_t \geq \begin{cases} m_{t-1}+d_t \text{ for } 1\leq t\leq 6 \\ m_{t-1}+d_t-q_1 \text{ for } 7\leq t\leq 14 \\ m_{t-1}+d_t-q_2 \text{ for } 15\leq t\leq 20 \end{cases} \\ & m_t \geq 0 \ \forall t \\ & m_0 = 0. \end{aligned} \tag{B.29}$$

Proposition B.2. *The specific linearized cost function $C(\vec{q},\vec{d})$, as stated in (B.29), is jointly convex in $\vec{q}$.*

Proof of Proposition B.2: The specific linearized cost function (B.29) consists of a linear and therefore convex expression $(c_q\left((t_2-t_1)q_1+(t_3-t_2)q_2\right))$ and $\min_{\{m_t\}}(c_2 m_T + c_3 m_{14})$. The minimization $\min_{\{m_t\}}(c_2 m_T + c_3 m_{14})$ is determined by solving a linear program in which $\vec{q}$ determines the right-hand side, and therefore convex. Consequently, the specific linearized cost function $C(\vec{q},\vec{d})$ is a sum of convex functions and, therefore, jointly convex in $\vec{q}$.

To apply the kERM approach, we first adapt the solution presented in Proposition 3.5 to the specific capacity planning problem stated in (B.29). The resulting dual Lagrangian can be stated as

$$L_{\text{dual}} = -\sum_{s=1}^{2}\lambda_s\sum_{p=1}^{N}\sum_{q=1}^{N}(u_s^p u_s^q K(\vec{x}^p,\vec{x}^q)) + \sum_{n,t=7}^{t=20}\alpha_t^n d_t^n + \sum_{n,t=1}^{t=6}\alpha_7^n d_t^n, \tag{B.30}$$

where

$$\begin{aligned} u_1^n &= \frac{1}{2\lambda_1}\left(\sum_{t=7}^{14}\alpha_t^n + \delta_1^n - \frac{c_q(t_2-t_1)}{N}\right) \\ u_2^n &= \frac{1}{2\lambda_2}\left(\sum_{t=15}^{20}\alpha_t^n + \delta_2^n - \frac{c_q(t_3-t_2)}{N}\right). \end{aligned} \tag{B.31}$$

The dual problem can be stated as

$$\begin{aligned} \max_{\{\alpha_t^n\},\{\delta_s^n\}} \; & L_{\text{dual}} \\ \text{s.t. } & \alpha_t^n,\ \delta_s^n \geq 0 \ \forall s,t,n \\ & \alpha_{t+1}^n \geq \alpha_t^n \ \forall n, t \neq 14 \\ & \alpha_{t+1}^n \geq \alpha_t^n - \frac{c_3}{N} \ \forall n, t = 14 \\ & \alpha_{20}^n \leq \frac{c_2}{N} \ \forall n, \end{aligned} \tag{B.32}$$

where $n = 1...N$, $s = 1...2$ and $t = 7...20$.

We use the same random forest model as for the wSAA approach and calculate the random forest kernel $\boldsymbol{K}_{pq}^{\text{RF}} = K^{\text{RF}}(\vec{x}^p, \vec{x}^p)$ by solving (3.14) for $p = 1...N$, $q = 1...N$. Similarly we calculate the kernel matrix for the test data sample, with $q = 1...N_{test}$. The tuning parameter λ_s is estimated in the range of $5 \cdot 10^{-7}...5 \cdot 10^{-3}$ using simple cross validation with 3/4 of the training data being used for model training, and 1/4 of the training data for validation.

Based on the kernel matrices we solve Problem B.32 using *Gurobi Optimizer* to determine $\vec{u}^n$. We then use

$$\vec{q}^{\text{kERM}}(\vec{x}) = \sum_{n=1}^{N} \vec{u}^n K(\vec{x}^n, \vec{x}) - \vec{b} \tag{B.33}$$

and solve the primal problem for $\vec{b}$, which is a linear problem, using *Gurobi Optimizer*. The resulting function is used to prescribe capacity decisions for the test period.

B.3.3 Optimization Prediction Approach

The OP approach consists of two steps, first, the estimation of the ex-post optimal decisions for the training data set, and second, the training of random forest models to predict optimal decisions.

We use *Gurobi Optimizer* to solve Problem 3.17 for each day of the training data set. These ex-post optimal capacity decisions q_1, q_2 form, together with the features, the training data set for two random forest models. We train two

random forest models using the *caret* R-package (Kuhn 2008) and determine the optimal number of variables to be selected for each tree from the range of 20...140 by cross validation parameter tuning. The random forest models are used to prescribe capacity decisions for the test period.

B.3.4 SAA

We solve Problem B.28 with $w_n(\vec{x}) = 1/N$ using *Gurobi Optimizer* to determine a capacity prescription, which is constant for the test period.

B.3.5 PDE-T20 Approach

The PDE-T20 approach is based on estimating a large number of demand distributions, and then solving Problem 3.17 using Monte Carlo simulation, as described in Section 3.3.3.

We set $T = 20$ and partition the demand data into a total of 120 empirical demand distributions (for 20 time periods and 6 weekdays). For each of these distributions we estimate mean and standard deviation, thereby fit a normal distribution, and draw $N_{MC} = 300$ samples from each distribution. We then estimate the prescription $\vec{q}^{PDE-T20}$ for each weekday by solving

$$\begin{aligned}\vec{q}^{PDE-T20} = \arg\min_{\vec{q}\in\mathcal{Q}} \min_{\{m_t^n\}\in\mathbb{R}^{T\times N}} \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} \Big[& c_q\Big((t_2-t_1)q_1 + (t_3-t_2)q_2\Big) \\ & + c_2 m_T^n + c_3 m_{14}^n\Big] \\ \text{s.t. } m_t^n \geq & \begin{cases} m_{t-1}^n + d_t^n \text{ for } 1 \leq t \leq 6 \ \forall n \\ m_{t-1}^n + d_t^n - q_1 \text{ for } 7 \leq t \leq 14 \ \forall n \\ m_{t-1}^n + d_t^n - q_2 \text{ for } 15 \leq t \leq 20 \ \forall n \end{cases} \\ & m_t^n \geq 0 \ \forall t, n \\ & m_0^n = 0 \ \forall n, \end{aligned} \tag{B.34}$$

using *Gurobi Optimizer*.

B.3.6 PDE-T2 Approach

The PDE-T2 approach follows a similar structure as the PDE-T20 approach. However, we partition the demand data into a total of six two-dimensional empirical distributions (for 6 weekdays and 2 time periods) and estimate the vector-valued mean and the covariance matrix. We thereby fit a multivariate normal distribution and draw $N_{MC} = 300$ samples from each of the six distributions. We then estimate the prescription $\vec{q}^{PDE-T2}$ for each weekday by solving

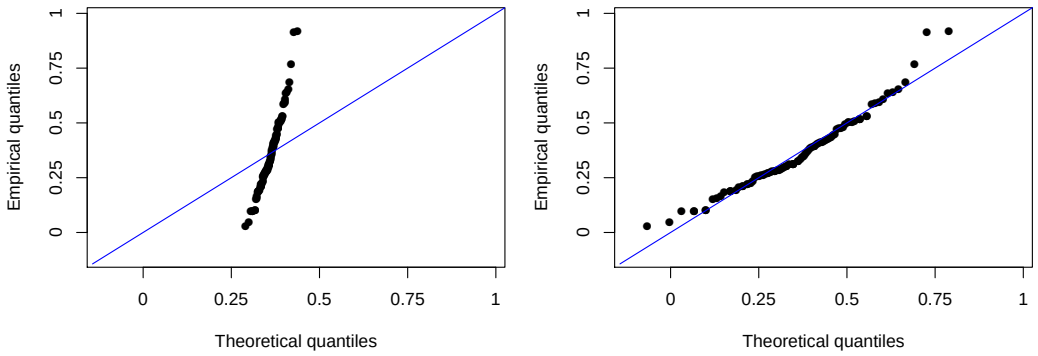
$$\begin{aligned} \vec{q}^{PDE-T2} = \arg\min_{\vec{q}\in\mathcal{Q}} \min_{\{m_t^n\}\in\mathbb{R}^{2\times N}} \frac{1}{N_{MC}} \sum_{n=1}^{N_{MC}} \Big[& c_q\Big((t_2-t_1)q_1 + (t_3-t_2)q_2\Big) \\ & + c_2 m_2^n + c_3 m_1^n \Big] \\ \text{s.t. } & m_1^n \geq d_1^n - q_1 \ \forall n \\ & m_2^n \geq n_1^n + d_2^n - q_2 \ \forall n \\ & m_t^n \geq 0 \ \forall t, n, \end{aligned} \tag{B.35}$$

using *Gurobi Optimizer*.

B.4 Analysis of Demand Arrival Distributions

In our analysis of the historical demand data in Section 3.1, we observed an empirical coefficient of variation $CV_{\bar{\lambda}}$ that was much larger than the CV_{Poisson} that would have been assumed by a Poisson distribution for the demand arriving between 9 a.m. and 10 a.m. for each weekday, suggesting a doubly stochastic demand process. This section first illustrates the mismatch of assuming a simple Poisson distribution for the demand arrivals in this time period on Mondays and then extends this observation to all time periods on all weekdays.

Figure B.2 illustrates the mismatch of assuming a Poisson distribution by showing quantile-quantile plots (Q-Q plots) that compare the quantiles of the observed empirical distribution with a theoretically fitted distribution. Figure B.2a shows that the Q-Q values for the Poisson distribution (black dots)



(a) Poisson distribution. (b) Normal distribution.

Figure B.2: Quantile-Quantile plots between empirical and fitted (theoretical) demand distributions (black) for Monday, 9 a.m. - 10 a.m., with optimal fit indicated as blue line (normalized).

deviate strongly from the optimal fit (blue line), indicating a low goodness-of-fit. Moreover, the plot shows that the variance of the fitted Poisson distribution is much smaller than the empirical variance that can be measured from the data set. In contrast, the Q-Q plot of a normal distribution (Figure B.2b) demonstrates a much better fit to the empirical demand data.

In order to extend our observations for the particular time period from 9 a.m. to 10 a.m. to all time periods, we present in Table B.1 the empirical and Poisson CVs for all weekdays and time periods. This table shows that $CV_{\text{emp}} \gg CV_{\text{Pois}}$ for all time periods, except for some early morning, late evening or Saturday afternoon time periods, in which the demand is almost zero. The assumption of an uncertainty-dominated regime is therefore valid.

B.5 Solution to the AMSSP without Further Constraints

This section shows that, without any further constraints, assigning the full required capacity to the last shift $s = S$ is an optimal solution to the AMSSP. To simplify the argument, we assume a service level larger than 50 percent and the number of shifts S to equal the number of time periods T. Then, it

Table B.1: Empirical CV_{emp} and Poisson CV_{Pois} for all weekdays and time periods in percent.

Hour	Monday		Tuesday		Wednesday		Thursday		Friday		Saturday	
	CV_{emp}	CV_{Pois}	CV_{emp}	CV_{Pois}	CV_{emp}	CV_{Pois}	CV_{emp}	CV_{Pois}	CV_{emp}	CV_{Pois}	CV_{emp}	CV_{Pois}
0	312.9	298.3	367.7	174.1	242.1	161.1	436.4	159.4	272.4	268.9	979.8	979.8
1	260.1	222.4	229.3	197.9	192.1	185.2	245.3	183.2	227.7	168.8	689.2	692.8
2	148.9	111.2	150.0	120.3	153.4	109.5	137.5	110.5	144.2	124.1	482.1	489.9
3	151.3	137.6	150.7	133.2	176.9	137.2	172.6	144.5	190.9	114.3	365.2	271.7
4	276.2	89.1	214.2	147.9	249.1	147.7	198.4	141.4	202.7	125.2	385.3	224.8
5	190.8	52.1	301.2	47.3	182.3	48.8	198.1	46.4	159.4	53.7	308.3	112.4
6	72.8	12.0	78.7	14.0	90.4	16.0	64.7	17.4	70.0	17.6	122.5	23.0
7	66.7	9.6	103.4	10.5	56.3	10.5	63.3	9.9	48.3	10.4	81.2	14.9
8	59.1	8.1	68.8	10.8	69.5	10.9	72.2	10.2	62.7	10.1	105.6	14.5
9	44.2	7.4	48.0	8.3	54.6	7.8	48.7	7.3	48.5	7.3	114.1	18.2
10	39.5	7.0	42.0	7.9	41.7	7.4	47.9	7.3	41.0	7.3	90.0	14.6
11	36.3	7.0	43.8	9.3	45.2	8.3	43.3	8.3	42.1	8.0	77.4	12.3
12	41.1	7.7	44.3	10.2	46.1	9.0	46.1	9.0	44.8	8.4	81.0	13.8
13	48.4	8.1	58.8	10.0	61.3	10.9	55.3	9.5	63.8	8.9	136.2	20.1
14	70.4	10.6	57.6	12.7	73.2	12.4	83.3	11.9	76.4	12.3	491.3	163.3
15	62.1	10.9	60.7	12.9	69.7	12.9	74.8	12.1	87.4	12.9	344.4	261.9
16	87.8	14.3	76.3	15.6	83.4	14.1	78.0	14.6	82.8	16.4	339.5	213.8
17	100.4	14.6	105.6	16.4	84.1	15.7	92.6	15.9	84.5	17.2	283.8	237.6
18	68.4	15.2	91.0	16.1	82.2	16.6	78.0	16.2	81.7	16.3	559.7	565.7
19	133.1	22.6	302.6	22.0	156.7	23.6	126.6	22.3	155.5	27.6	428.9	438.2
20	105.9	19.8	116.9	21.7	148.3	22.4	175.6	22.6	126.1	25.8	535.6	370.3
21	174.4	25.2	148.9	33.9	198.9	35.0	141.4	42.4	146.5	40.2	482.1	489.9
22	182.9	41.4	179.9	53.2	161.4	71.1	186.1	71.7	233.7	92.4	689.2	489.9
23	233.1	70.7	306.7	101.1	417.5	208.9	499.0	292.3	681.9	685.6	979.8	979.8

is optimal to only employ capacity in the last time period T:

$$\begin{aligned} q_t &= 0 \ \forall t < T \\ q_T &= \arg\min_q \left(c_q q + c_2 \mathbb{E}\left[\left(M_0 + \sum_{t=1}^{T} D_t - q \right)^+ \right] \right), \end{aligned} \tag{B.36}$$

where q_T is determined by a single-period newsvendor-type problem.

The reason for this being an optimal solution is that the standard deviation of the sum of arriving demands $D = \sum_{t=1}^{T} D_t$, which equals the standard deviation of M_{T-1} if $q_t = 0 \ \forall t < T$, is always smaller or equal than the sum

of the standard deviations of each arriving demand:

$$\begin{aligned}\sigma_D &= \sqrt{\sum_{t=1}^{T} \sigma_{D_t}^2 + 2 \sum_{t_1<t_2} \mathrm{Cov}(D_{t_1}, D_{t_2})} \\ &\leq \sqrt{\sum_{t=1}^{T} \sigma_{D_t}^2 + 2 \sum_{t_1<t_2} \sigma_{D_{t_1}} \sigma_{D_{t_2}}} \qquad = \sum_{t=1}^{T} \sigma_{D_t},\end{aligned} \tag{B.37}$$

where equality holds only for perfect correlation. Therefore, in all cases, except for perfect correlation, it is more profitable to only use capacity q_T, because planning for the total demand D requires less safety buffer due to the reduced standard deviation.

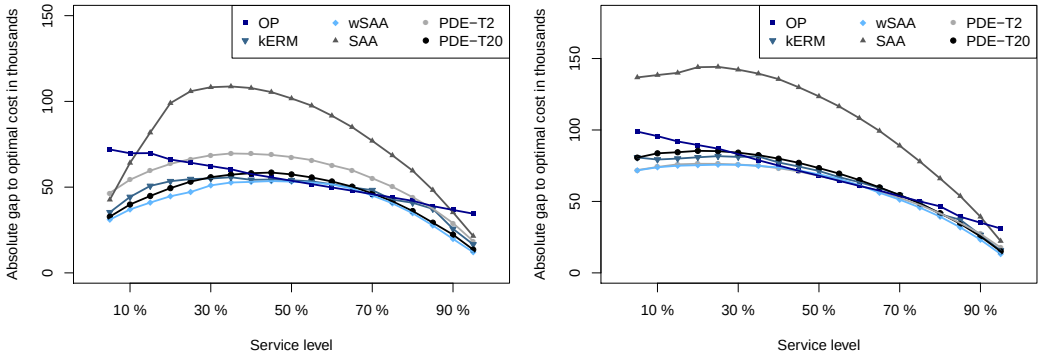
B.6 Robustness Analysis

This section analyses the robustness of all approaches by comparing their performance for various cost parameter settings (Table B.2) that induce service levels between 5 and 95 percent and load-balancing factors between 0.05 and 0.95.

Table B.2: Variations in Cost Parameters for Robustness Analysis.

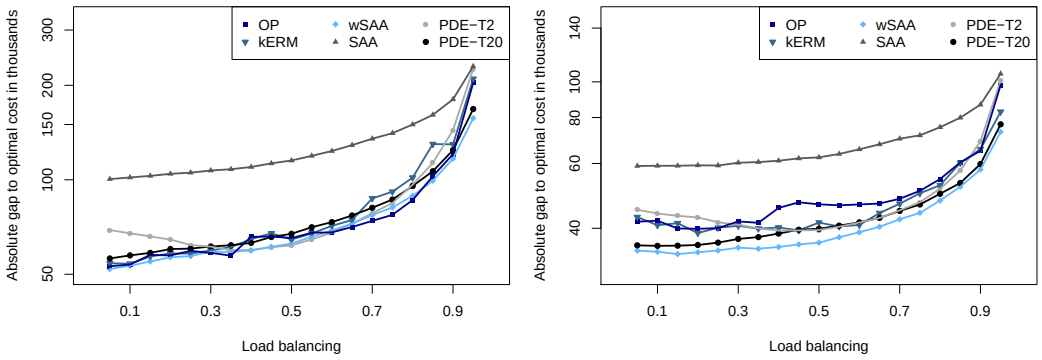
Figure	c_q	c_2	c_3	**Service Level**	**Load Balancing**
B.3a	5...0.3	5.25	0.6...0.06	5%...95%	0.1
B.3b	5...0.3	5.25	7.5...0.4	5%...95%	0.6
B.4a	2.63	5.25	0.14...50	50%	0.05...0.95
B.4b	1.05	5.25	0.06...20	80%	0.05...0.95

Figure B.3 shows the performance of all approaches in terms of the gap to optimal cost for load-balancing factors $LB = 0.1$ (Figure B.3a) and $LB = 0.6$ (Figure B.3b). In addition to the discussion in Section 3.5.6, we observe that, when comparing Figures B.3a and B.3b, the PDE-T2 approach leads to a significantly lower performance than PDE-T20 for $LB = 0.1$. This difference in performance can be explained by the time-structure effect (Section 3.5.4).



(a) Load balancing $LB = 0.1$. (b) Load balancing $LB = 0.6$.

Figure B.3: Absolute gap to optimal cost across service levels (SL) for all approaches and different levels of load-balancing (LB).



(a) Service level $SL = 50\%$. (b) Service level $SL = 80\%$.

Figure B.4: Absolute gap to optimal cost across levels of load-balancing (LB) for different service levels (SL).

Figure B.4 shows the performance of all approaches in terms of the gap to optimal cost in logarithmic scale for service levels of $SL = 50\%$ (Figure B.4a) and $SL = 80\%$ (Figure B.4b). The prescriptive approaches' and PDE-T20's performance vary similarly with the load-balancing factor LB, which is in line with all of these approaches' incorporation of the full time structure ($T = 20$). In contrast, the performance of the PDE-T2 approach is significantly lower than that of PDE-T20 for very small or very large levels of load-balancing, which, again, can be explained by the time-structure effect. When LB is comparably large, it is most cost-effective to increase the capacity of shift 1,

such that (almost) no demand is backlogged between the shifts, while accepting idle capacity at the beginning of shift 1. Similarly, it is most cost-effective to decrease the capacity of shift 1 when LB is very small. PDE-T2, because it neglects the time structure of demand within the shifts, does not adapt the shift 1 capacity adequately, and thus leads to lower performance for a very small or very large LB.

B.7 Analytical Results for Prescriptive Analytics Approaches

In this section we provide a bound on the cost function $C(\vec{q}, \vec{d})$ and show its Lipschitz-continuity, which allows us to establish analytical results including asymptotic optimality for wSAA and out-of-sample performance guarantees for kERM.

We begin by defining demand and decision sets:

Definition B.1. *Let $\mathcal{Q}$ and $\mathcal{D}$ be defined as*

$$\begin{aligned} \mathcal{Q} &= \{\vec{q} = \{q_s\} \in \mathbb{R}^S : 0 \leq q_s \leq \bar{q}\ \forall s\} \\ \mathcal{D} &= \{\vec{d} = \{d_t\} \in \mathbb{R}^T : 0 \leq d_t \leq \bar{d}\ \forall t\}. \end{aligned} \tag{B.38}$$

Clearly, demand and capacity can only take on positive values, which justifies the lower bound. The upper bounds can be large, however demand is typically limited by the market, and capacity may be limited by the overall available staff. Based on this definition, we can derive a bound and the Lipschitz property for the cost function (3.11).

Proposition B.3. *The linearized cost function, stated in (3.11), is positive-valued:*

$$0 \leq C(\vec{q}, \vec{d})\ \forall \vec{q} \in \mathcal{Q}, \vec{d} \in \mathcal{D} \tag{B.39}$$

and bounded over $\mathcal{Q}$, $\mathcal{D}$ as

$$\sup_{\vec{q} \in \mathcal{Q}, \vec{d} \in \mathcal{D}} C(\vec{q}, \vec{d}) \leq \bar{l}. \tag{B.40}$$

Proof of Proposition B.3: Because the cost factors are positive ($c_q, c_2 \geq 0$) and $q_s, m_t \geq 0\ \forall s, t$, the cost function is defined as the sum of positive values and, consequently, positive-valued.

To prove the upper bound of the cost function, we first derive a bound on the overtime cost, which, for given $\vec{q}$ and $\vec{d}$, can be defined as

$$\begin{aligned} C_{\text{overtime}}(\vec{q}, \vec{d}) = \min_{\{m_t\} \in \mathbb{R}^T} & c_2 m_T \\ \text{s.t. } & m_t \geq m_{t-1} + d_t - q_s \ \forall t \in [t_s, t_{s+1}) \ \forall s = 1...S \\ & m_t \geq 0 \ \forall t \\ & m_0 = M_0. \end{aligned} \tag{B.41}$$

We bound this overtime cost as

$$C_{\text{overtime}}(\vec{q}, \vec{d}) \leq C_{\text{overtime}}(0, \vec{d}) \ \forall \vec{d} \in \mathcal{D}, \tag{B.42}$$

because the solution $\{m_t\}$ of $C_{\text{overtime}}(0, \vec{d})$ is also feasible for $C_{\text{overtime}}(\vec{q}, \vec{d})$ (the first constraint in Problem B.41 is weaker when setting $\vec{q} \geq 0$ compared to $\vec{q} = 0$), and, therefore, when $\vec{q} \geq 0$, the minimum of $c_2 m_T$ can only be smaller or equal compared to a setting in which $\vec{q} = 0$.

Similarly, we obtain

$$C_{\text{overtime}}(\vec{q}, \vec{d}) \leq C_{\text{overtime}}(\vec{q}, \{\bar{d}\}) \ \forall \vec{q} \in \mathcal{Q} \tag{B.43}$$

by the respective argument, because $d_t \leq \bar{d}\ \forall t, \vec{d} \in \mathcal{D}$.

Therefore, we obtain a bound on the overtime cost as

$$C_{\text{overtime}}(\vec{q}, \vec{d}) \leq C_{\text{overtime}}(0, \{\bar{d}\}) = c_2 \left(M_0 + T\bar{d}\right), \tag{B.44}$$

which allows us to bound the cost function:

$$\begin{aligned} C(\vec{q}, \vec{d}) &= \frac{1}{T} \sum_{s=1}^{S} c_q (t_{s+1} - t_s) q_s + C_{\text{overtime}}(\vec{q}, \vec{d}) \\ &\leq c_q \bar{q} + c_2 \left(M_0 + T\bar{d}\right) =: \bar{l}, \end{aligned} \tag{B.45}$$

which concludes the proof.

Proposition B.4. *The linearized cost function, stated in (3.11), is equi-Lipschitz in $\vec{q}$ over $\mathcal{Q}$ and $\mathcal{D}$, and there is some $M_{Lip} < \infty$, such that*

$$\sup_{\vec{q},\vec{q}' \in \mathcal{Q}, \vec{q} \neq \vec{q}', \vec{d} \in \mathcal{D}} \frac{|C(\vec{q},\vec{d}) - C(\vec{q}',\vec{d})|}{||\vec{q} - \vec{q}'||_\infty} \leq M_{Lip}. \tag{B.46}$$

Proof of Proposition B.4: To prove the equi-Lipschitz property of the linearized cost function, we need to show that $C_{\text{overtime}}(\vec{q},\vec{d})$, as defined in (B.41), is equi-Lipschitz, because the linear part of the cost function is equi-Lipschitz:

$$\begin{aligned}
|C(\vec{q},\vec{d}) - C(\vec{q}',\vec{d})| &= \left| \frac{1}{T} \sum_{s=1}^{S} c_q (t_{s+1} - t_s)(q_s - q'_s) \right. \\
&\qquad \left. + C_{\text{overtime}}(\vec{q},\vec{d}) - C_{\text{overtime}}(\vec{q}',\vec{d}) \right| \\
&\leq \left| \frac{1}{T} \sum_{s=1}^{S} c_q (t_{s+1} - t_s)(q_s - q'_s) \right| \\
&\quad + \left| C_{\text{overtime}}(\vec{q},\vec{d}) - C_{\text{overtime}}(\vec{q}',\vec{d}) \right| \\
&\leq c_q ||\vec{q} - \vec{q}'||_\infty + \left| C_{\text{overtime}}(\vec{q},\vec{d}) - C_{\text{overtime}}(\vec{q}',\vec{d}) \right|.
\end{aligned} \tag{B.47}$$

To show the equi-Lipschitz property of the overtime cost, we define a sequence of capacity configurations $\vec{q}_\tau$ with $\tau = 0...S$ such that $\vec{q}_0 = \vec{q}$, $\vec{q}_S = \vec{q}'$ and $\vec{q}_\tau, \vec{q}_{\tau+1}$ differing only in dimension $\tau + 1$. Then, we obtain

$$\begin{aligned}
\left| C_{\text{overtime}}(\vec{q},\vec{d}) - C_{\text{overtime}}(\vec{q}',\vec{d}) \right| &= \left| \sum_{\tau=0}^{S-1} \left[C_{\text{overtime}}(\vec{q}_\tau,\vec{d}) - C_{\text{overtime}}(\vec{q}_{\tau+1},\vec{d}) \right] \right| \\
&\leq \sum_{\tau=0}^{S-1} \left| C_{\text{overtime}}(\vec{q}_\tau,\vec{d}) - C_{\text{overtime}}(\vec{q}_{\tau+1},\vec{d}) \right|.
\end{aligned} \tag{B.48}$$

Assuming $(\vec{q}_\tau)_{\tau+1} \geq (\vec{q}_{\tau+1})_{\tau+1}$ (both differ only along dimension $\tau+1$), without loss of generality, we know for the queue length $\{m_t^*\}$ (defined as solution to

Problem B.41) that

$$m_t^*(\vec{q}_\tau, \vec{d}) = m_t^*(\vec{q}_{\tau+1}, \vec{d}) \text{ for } t < t_{\tau+1}, \tag{B.49}$$

because shift $\tau + 1$ begins at $t_{\tau+1}$, and

$$m_{t_{\tau+1}}^*(\vec{q}_\tau, \vec{d}) \leq m_{t_{\tau+1}}^*(\vec{q}_{\tau+1}, \vec{d}) \tag{B.50}$$

with the difference being bounded as

$$m_{t_{\tau+1}}^*(\vec{q}_{\tau+1}, \vec{d}) - m_{t_{\tau+1}}^*(\vec{q}_\tau, \vec{d}) \leq (\vec{q}_\tau - \vec{q}_{\tau+1})_{\tau+1}. \tag{B.51}$$

Then, by iteratively applying the first condition of (3.11), we obtain

$$m_T^*(\vec{q}_{\tau+1}, \vec{d}) - m_T^*(\vec{q}_\tau, \vec{d}) \leq (t_{\tau+2} - t_{\tau+1})(\vec{q}_\tau - \vec{q}_{\tau+1})_{\tau+1}, \tag{B.52}$$

because the remaining q_s and d_t are identical for both $\vec{q}_\tau$ and $\vec{q}_{\tau+1}$. Consequently,

$$\begin{aligned} C_{\text{overtime}}(\vec{q}_{\tau+1}, \vec{d}) - C_{\text{overtime}}(\vec{q}_\tau, \vec{d}) &\leq c_2(t_{\tau+2} - t_{\tau+1})(\vec{q}_\tau - \vec{q}_{\tau+1})_{\tau+1} \\ &\leq c_2(t_{\tau+2} - t_{\tau+1})||\vec{q} - \vec{q}'||_\infty, \end{aligned} \tag{B.53}$$

from which we obtain

$$\begin{aligned} |C(\vec{q}, \vec{d}) - C(\vec{q}', \vec{d})| &\leq c_q||\vec{q} - \vec{q}'||_\infty + \left|C_{\text{overtime}}(\vec{q}, \vec{d}) - C_{\text{overtime}}(\vec{q}', \vec{d})\right| \\ &\leq c_q||\vec{q} - \vec{q}'||_\infty + \sum_{\tau=0}^{S-1} \left|C_{\text{overtime}}(\vec{q}_\tau, \vec{d}) - C_{\text{overtime}}(\vec{q}_{\tau+1}, \vec{d})\right| \\ &\leq (c_q + Tc_2)\,||\vec{q} - \vec{q}'||_\infty =: M_{Lip}||\vec{q} - \vec{q}'||_\infty, \end{aligned} \tag{B.54}$$

which concludes the proof.

In the following we show that the results of Bertsimas and Kallus (2020) on asymptotic optimality of wSAA also apply to our wSAA approach for our particular cost function.

Proposition B.5. *(Following Bertsimas and Kallus 2020) Assume a closed,*

bounded, non-empty decision space $\tilde{\mathcal{Q}} \subset \mathbb{R}_+^T$*, a bounded, non-empty demand space* $\tilde{\mathcal{D}} \subset \mathbb{R}_+^T$*, and a data set* S_N^T *generated by iid sampling from a joint distribution of* $\vec{X} \times \vec{D}$*. Then, the wSAA approach that solves the linearized AMSSP is asymptotically optimal when using any of the following weight functions:*

a) Based on k-nearest-neighbors (kNN):

$$w_n^{kNN}(\vec{x}) = \frac{1}{k} \mathbb{1}[\vec{x}^n \text{ is a kNN of } \vec{x}], \tag{B.55}$$

with $k = \min(\lceil CN^\delta \rceil, N-1)$ *for some* $C > 0$ *and* $0 < \delta < 1$*.*

b) Based on kernel methods:

$$w_n^K(\vec{x}) = \frac{K((\vec{x}^n - \vec{x})/h_N)}{\sum_{k=1}^N K((\vec{x}^k - \vec{x})/h_N)}, \tag{B.56}$$

with $h_N = CN^{-\delta}$ *for some* $C > 0$*,* $0 < \delta < 1/p$ *with* $\vec{x} \in \mathbb{R}^p$*, and* K *being one of the following kernels: naïve* $K(\vec{x}) = \mathbb{1}[||\vec{x}|| \leq 1]$*, Epanechnikov* $K(\vec{x}) = (1 - ||\vec{x}||^2)\mathbb{1}[||\vec{x}|| \leq 1]$*, Tri-cubic* $K(\vec{x}) = (1 - ||\vec{x}||^3)^3 \mathbb{1}[||\vec{x}|| \leq 1]$*, or Gaussian* $K(\vec{x}) = \exp(-||\vec{x}||^2/2)$*.*

c) Based on recursive kernel methods:

$$w_n^{rK}(\vec{x}) = \frac{K((\vec{x}^n - \vec{x})/h_n)}{\sum_{k=1}^N K((\vec{x}^n - \vec{x})/h_n)}, \tag{B.57}$$

with $h_n = Cn^{-\delta}$ *for some* $C > 0$*,* $0 < \delta < 1/(2p)$ *with* $\vec{x} \in \mathbb{R}^p$*, and* K *being the naïve kernel* $K(\vec{x}) = \mathbb{1}[||\vec{x}|| \leq 1]$*.*

d) Based on local linear methods:

$$w_n^{LL}(\vec{x}) = \frac{\tilde{w}_n(\vec{x})}{\sum_{k=1}^N \tilde{w}_k(\vec{x})}, \tag{B.58}$$

with

$$\begin{aligned}\tilde{w}_n(\vec{x}) &= k_n(\vec{x})\Big(1 - \sum_{l=1}^{N} k_l(\vec{x})(\vec{x}^l - \vec{x})^T \Xi(\vec{x})^{-1}(\vec{x}^n - \vec{x})\Big),\\ \Xi(\vec{x}) &= \sum_{n=1}^{N} k_n(\vec{x})(\vec{x}^n - \vec{x})(\vec{x}^n - \vec{x})^T,\\ k_n(\vec{x}) &= K((\vec{x}^n - \vec{x})/h_N),\end{aligned} \tag{B.59}$$

and $h_N = CN^{-\delta}$ *for some* $C > 0$, $0 < \delta < 1/p$ *with* $\vec{x} \in \mathbb{R}^p$, *and* K *being one of the following kernels: naïve* $K(\vec{x}) = \mathbb{1}[||\vec{x}|| \leq 1]$, *Epanechnikov* $K(\vec{x}) = (1-||\vec{x}||^2)\mathbb{1}[||\vec{x}|| \leq 1]$, *Tri-cubic* $K(\vec{x}) = (1 - ||\vec{x}||^3)^3\mathbb{1}[||\vec{x}|| \leq 1]$, *or Gaussian* $K(\vec{x}) = \exp(-||\vec{x}||^2/2)$, *if the distribution of feature vectors* $\vec{x}$ *is absolutely continuous and the probability density* $f(\vec{x})$ *is bounded away from zero and* ∞ *over* $\mathcal{X}$.

Proof of Proposition B.5: To prove asymptotic optimality of the wSAA approach, we show that all assumptions of Theorems 2-5 in Bertsimas and Kallus (2020) are fulfilled, from which the statement to be proven follows. For any $\tilde{\mathcal{Q}}$ and $\tilde{\mathcal{D}}$ we find $\mathcal{Q}$ and $\mathcal{D}$ that follow Definition B.1, such that $\tilde{\mathcal{Q}}$ and $\tilde{\mathcal{D}}$ are subsets of $\mathcal{Q}$ and $\mathcal{D}$. Therefore, the following assumptions are fulfilled:

- Existence (Assumption 1 in Bertsimas and Kallus 2020): Proposition B.3 establishes the bound on the cost function, therefore we obtain that $\mathbb{E}(|C(\vec{q}, \vec{D})|) \leq \bar{l} < \infty$, and $\tilde{\mathcal{Q}}$ is non-empty by definition.
- Continuity (Assumption 2 in Bertsimas and Kallus 2020): Proposition B.4 establishes the equi-Lipschitz property of the cost function.
- Regularity (Assumption 3 in Bertsimas and Kallus 2020): $\tilde{\mathcal{Q}}$ is closed, bounded and non-empty by definition.

In addition, using the bound on the cost function provided in Proposition B.3, we obtain

$$\mathbb{E}[|C(\vec{q}, \vec{D})| \max(\log|C(\vec{q}, \vec{D})|, 0)] \leq \bar{l} \log \bar{l} < \infty.$$

Therefore, the results of Theorems 2-5 in Bertsimas and Kallus (2020) apply, which concludes the proof.

For kERM, we can, based on the bound and the Lipschitz-property of the cost function, obtain (following the results presented in Notz and Pibernik 2021) an out-of-sample performance guarantee when using the random forest kernel and the universal approximation property when using a universal kernel.

An out-of-sample performance guarantee is a bound on the true risk, which is defined as the expected cost over the (unknown) joint distribution of $\vec{X} \times \vec{D}$: $R(\vec{q}(\cdot)) := \mathbb{E}_{\vec{X}\times\vec{D}}\left[C(\vec{q}(\vec{X}), \vec{D})\right]$, based on the (measurable) empirical risk $R_N^T(\vec{q}(\cdot)) := \sum_{n=1}^{N}\left[C(\vec{q}(\vec{x}^n), \vec{d}^n)\right]$ over the data sample S_N^T. Because the random forest kernel is data-dependent, we need to use a sample-splitting approach to derive an out-of-sample performance guarantee for kERM when using the random forest kernel. In particular, we use N_{RF} data samples to compute the random forest kernel function, and the remaining $N - N_{RF}$ data samples to train the kERM approach.

Proposition B.6. *(Following Notz and Pibernik 2021) Assume a data set $S_N^T = S_{N_{RF}}^T \uplus S_{N-N_{RF}}^T$, generated by iid sampling from a joint distribution of $\vec{X} \times \vec{D}$, $C(\vec{q}, \vec{d})$, as defined in (3.11), $||\vec{b}||_\infty \leq B_C$, $||q_{U,s}||_K \leq B_U \ \forall s$, $K(\vec{x}_1, \vec{x}_2) = K^{RF}(\vec{x}_1, \vec{x}_2)$, as defined in (3.14) and computed using $S_{N_{RF}}^T$; and let $\delta > 0$. Then, with probability of at least $1 - \delta$ for any function $\vec{q}(\cdot) \in \mathcal{F}^{RF}$, the true risk is bounded as*

$$\begin{aligned} R(\vec{q}(\cdot)) \leq & R_{N-N_{RF}}^T(\vec{q}(\cdot)) + 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2(N-N_{RF})}} \\ & + M_{Lip}\left(\frac{2\sqrt{2}IB_C e}{\sqrt{\pi}\sqrt{N-N_{RF}}} + \frac{2IB_U}{\sqrt{N-N_{RF}}}\right), \end{aligned} \tag{B.60}$$

where $\bar{l}$ is the bound and M_{Lip} is the Lipschitz constant of $C(\vec{q}, \vec{d})$.

Proof of Proposition B.6: The proof follows a similar structure as the proofs of Theorems 2 and 3 in Notz and Pibernik (2021). The cost function $C(\vec{q}, \vec{d})$ is bounded and equi-Lipschitz (Propositions B.3 and B.4), therefore the results of Theorem 8 in Bertsimas and Kallus (2020) apply. Combining these results with the results of Lemmas 2, 3, and 4 in Notz and Pibernik (2021) and the bound on the random forest kernel function $K^{\mathrm{RF}}(\vec{x}_1, \vec{x}_2) \leq 1$

shown in the proof of Theorem 3 in Notz and Pibernik (2021), we obtain the expression to be proven.

Limiting the function space as in Proposition B.6 leads to a rate of convergence of $1/\sqrt{N - N_{RF}}$, however, the prescription function $\vec{q}^{\,\mathrm{kERM}}(\vec{x})$ only converges to the best function within the function space. In contrast, using universal kernels, such as the Gaussian RBF kernel, allows for convergence to the best of all continuous functions without any limitations on the function space, as shown in Proposition B.7. However, the same limitation on the speed of convergence as for wSAA applies (see Notz and Pibernik 2021 for details).

Proposition B.7. *(Following Notz and Pibernik 2021) Assume a data set S_N^T containing N iid samples of a joint distribution $\vec{X} \times \vec{D}$, $C(\vec{q}, \vec{d})$, as defined in (3.11), $||\vec{b}||_\infty \leq B_{C,N}$, $||q_{U,s}||_K \leq B_{U,N}$ $\forall s$, and $K(\vec{x}_1, \vec{x}_2) = K^{RBFG}(\vec{x}_1, \vec{x}_2)$ the RBF Gauss kernel. Assume also regularization sequences $B_{C,N}, B_{U,N}$ such that $\lim_{N\to\infty} B_{C,N}, B_{U,N} = \infty$ and $\lim_{N\to\infty} B_{C,N}/\sqrt{N}, B_{U,N}/\sqrt{N} = 0$. Then the kERM approach fulfills the universal approximation property and the associated true risk converges in probability for $N \to \infty$ toward the minimum risk*

$$R^* = \min_{\vec{q}(\cdot)\in C(\mathcal{X},\mathbb{R}^S)} R(\vec{q}(\cdot)), \tag{B.61}$$

where $C(\mathcal{X}, \mathbb{R}^S)$ is the space of continuous functions that map from $\mathcal{X}$ to $\mathbb{R}^S$.

Proof of Proposition B.7: The proof of Proposition B.7 follows a similar structure as the proof of Proposition 8 in Notz and Pibernik (2021). Combining the results of Theorem 8 in Bertsimas and Kallus (2020) (because $C(\vec{q}, \vec{d})$ is bounded and equi-Lipschitz, Propositions B.3 and B.4) with the results of Lemmas 2, 3, and 4 in Notz and Pibernik (2021) and the bound on the RBF Gauss kernel function $K^{\mathrm{RBFG}}(\vec{x}_1, \vec{x}_2) \leq 1$, we obtain an out-of-sample performance guarantee as provided in Equation 61 in Notz and Pibernik (2021). Based on this Equation 61, Notz and Pibernik (2021) show that the difference $\Delta_N(\vec{q}(\cdot))$ between the risk associated with the kERM prescription function and the minimum risk R^* converges to zero for $N \to \infty$, while the function space over which kERM optimizes converges to $C(\mathcal{X}, \mathbb{R}^S)$, which proves the universal approximation property and, therefore, concludes the proof.

C Appendix of Chapter 4

C.1 Proofs

Proof of Proposition 4.1

To prove that $\partial L(\vec{q}_0)$ is a non-empty, bounded set, we first observe that $L(\vec{q})$ is a proper convex function, because $L(\vec{q}) < \infty$ for at least one $\vec{q}$ and $L(\vec{q}) > -\infty$ for all $\vec{q} \in \mathcal{Q}$ (see definition of proper convex functions in Chapter 4 in Rockafellar 1970). Because $L(\vec{q})$ is proper convex, the results of Theorem 23.4 in Rockafellar (1970) apply, stating that the subdifferential $\partial L(\vec{q}_0)$ is a non-empty, bounded set for all $\vec{q}_0 \in \operatorname{int} \mathcal{Q}$.

Proof of Proposition 4.2

The unique subgradient of a convex differentiable function ($L(q)$ for $q \neq q_0$) is given as the gradient of the function: $L'(q) = \frac{dL(q)}{dq}$ (Theorem 25.1 in Rockafellar 1970). Furthermore, s_0 is a subgradient of $L(q)$ at $q = q_0$ (Definition 4.1), therefore, the subgradient of $L(q)$ can be expressed as the piecewise-defined function specified in Proposition 4.2.

Proof of Proposition 4.3

To prove the stated expression for the subdifferential $\partial L_{\boldsymbol{d}}(\vec{q})$, we show that both elements of the loss function $f_0(\vec{q}, \boldsymbol{d})$ and $f^*_{1,\boldsymbol{d}}(\vec{q})$ are proper convex, and then use the results of Theorem 23.8 in Rockafellar (1970).

Because by assumption $f_{0,\boldsymbol{d}}(\vec{q}), f^*_{1,\boldsymbol{d}}(\vec{q}) < \infty$ for at least one $\vec{q} \in \mathcal{Q}$ and $f_{0,\boldsymbol{d}}(\vec{q}), f^*_{1,\boldsymbol{d}}(\vec{q}) > -\infty$ for all $\vec{q} \in \mathcal{Q}$, both $f_{0,\boldsymbol{d}}(\vec{q})$ and $f^*_{1,\boldsymbol{d}}(\vec{q})$ are proper convex functions (Chapter 4 in Rockafellar 1970). Furthermore, because the relative interior set relint $\mathcal{Q}$ is non-empty and both $f_{0,\boldsymbol{d}}(\vec{q})$ and $f^*_{1,\boldsymbol{d}}(\vec{q})$ are defined on $\mathcal{Q}$,

the results of Theorem 23.8 in Rockafellar (1970) state that

$$\partial L_{\boldsymbol{d}}(\vec{q}) = \partial f_{0,\boldsymbol{d}}(\vec{q}) + \partial f^*_{1,\boldsymbol{d}}(\vec{q}) \;\forall \vec{q} \in \mathcal{Q}. \tag{C.1}$$

Because $f_{0,\boldsymbol{d}}(\vec{q})$ is proper, finite ($|f_{0,\boldsymbol{d}}(\vec{q})| < \infty$) and differentiable for all $\vec{q} \in \mathcal{Q}$, the gradient $\nabla f_{0,\boldsymbol{d}}(\vec{q})$ is the unique subgradient of $f_{0,\boldsymbol{d}}(\vec{q})$ for all $\vec{q} \in \mathcal{Q}$ (Theorem 25.1 in Rockafellar 1970), and consequently

$$(\partial f_{0,\boldsymbol{d}}(\vec{q}))_j = \frac{\partial f_{0,\boldsymbol{d}}(\vec{q})}{\partial q_j}, \tag{C.2}$$

which concludes the proof.

Proof of Theorem 4.1

Because all assumptions of Proposition 4.3 are fulfilled, a subgradient of $L_{\boldsymbol{d}}(\vec{q})$ is given as

$$(\vec{s}_{L,\boldsymbol{d},\vec{q}})_j = \frac{\partial f_{0,\boldsymbol{d}}(\vec{q})}{\partial q_j} + (\vec{s}_{f^*_1,\boldsymbol{d},\vec{q}})_j, \tag{C.3}$$

where $\vec{s}_{f^*_1,\boldsymbol{d},\vec{q}}$ is a subgradient of $f^*_{1,\boldsymbol{d}}(\vec{q})$.

We derive such a subgradient $\vec{s}_{f^*_1,\boldsymbol{d},\vec{q}}$ by using the Lagrange formalism and applying perturbation theory as presented in Chapter 5.6.2 in Boyd and Vandenberghe (2004) and in Boyd et al. (2018). Let

$$\begin{aligned} \tilde{g}_j(\boldsymbol{z}, \boldsymbol{d}) &:= g_j(\boldsymbol{z}, \boldsymbol{d}) - (\vec{q}_0)_j \\ \Delta\vec{q} &:= \vec{q} - \vec{q}_0 \\ \tilde{h}_{1,l}(\boldsymbol{z}, \boldsymbol{d}) &:= h_{1,l}(\boldsymbol{z}, \boldsymbol{d}) - u_l \\ \tilde{h}_{2,m}(\boldsymbol{z}, \boldsymbol{d}) &:= h_{2,m}(\boldsymbol{z}, \boldsymbol{d}) - v_m \end{aligned} \tag{C.4}$$

and let $\tilde{f}^*_{1,\boldsymbol{d},\vec{q}_0}(\Delta\vec{q}) := f^*_{1,\boldsymbol{d}}(\Delta\vec{q} + \vec{q}_0) = f^*_{1,\boldsymbol{d}}(\vec{q})$, then we can express $f^*_{1,\boldsymbol{d}}(\vec{q})$

defined in (4.17) as

$$\begin{aligned} \tilde{f}^*_{1,\boldsymbol{d},\vec{q}_0}(\Delta\vec{q}) = \min_{\boldsymbol{z}\in\mathcal{Z}} \; & f_1(\boldsymbol{z},\boldsymbol{d}) \\ \text{s.t. } & \tilde{g}_j(\boldsymbol{z},\boldsymbol{d}) \leq (\Delta\vec{q})_j \;\forall j \\ & \tilde{h}_{1,l}(\boldsymbol{z},\boldsymbol{d}) \leq 0 \;\forall l \\ & \tilde{h}_{2,m}(\boldsymbol{z},\boldsymbol{d}) = 0 \;\forall m. \end{aligned} \tag{C.5}$$

The primal Lagrangian corresponding to Problem C.5 can be expressed as

$$\begin{aligned} L_{Primal} = f_1(\boldsymbol{z},\boldsymbol{d}) + \sum_j \alpha_j \left[\tilde{g}_j(\boldsymbol{z},\boldsymbol{d}) - (\Delta\vec{q})_j\right] \\ + \sum_l \beta_{1,l}\tilde{h}_{1,l}(\boldsymbol{z},\boldsymbol{d}) + \sum_m \beta_{2,m}\tilde{h}_{2,m}(\boldsymbol{z},\boldsymbol{d}), \end{aligned} \tag{C.6}$$

where $\alpha_j, \beta_{1,l} \geq 0$ and $\beta_{2,m}$ are the Lagrange variables. Because, by assumption, strong duality holds and the dual optimum at $\Delta\vec{q} = 0$ (i.e., $\vec{q} = \vec{q}_0$) is attained at $(\vec{\alpha}^*_{\vec{q}_0}, \vec{\beta}^*_{1,\vec{q}_0}, \vec{\beta}^*_{2,\vec{q}_0})$, all assumptions required for Inequality 5.57 in Boyd and Vandenberghe (2004) to hold are fulfilled, and we obtain

$$\tilde{f}^*_{1,\boldsymbol{d},\vec{q}_0}(\Delta\vec{q}) \geq \tilde{f}^*_{1,\boldsymbol{d},\vec{q}_0}(0) - \langle\vec{\alpha}^*_{\vec{q}_0}, \Delta\vec{q}\rangle, \tag{C.7}$$

which is equivalent to

$$f^*_{1,\boldsymbol{d}}(\vec{q}) \geq f^*_{1,\boldsymbol{d}}(\vec{q}_0) - \langle\vec{\alpha}^*_{\vec{q}_0}, \vec{q} - \vec{q}_0\rangle. \tag{C.8}$$

Because Inequality C.8 holds for any $\vec{q}, \vec{q}_0 \in \mathcal{Q}$ and $f^*_{1,\boldsymbol{d}}(\vec{q})$ is jointly convex in $\vec{q}$ by assumption, $-\vec{\alpha}^*_{\vec{q}_0}$ is a subgradient of $f^*_{1,\boldsymbol{d}}(\vec{q})$ at $\vec{q}_0$ by Definition 4.1. Therefore, setting $\vec{s}_{f^*_1,\boldsymbol{d},\vec{q}} = -\vec{\alpha}^*_{\vec{q}}$ in Equation C.3, we obtain the expression to be proven.

Proof of Proposition 4.4

Convexity of the loss function defined in (4.21) has been shown in Proposition 2 in Notz and Pibernik (2021).

Proof of Proposition 4.5

To prove the existence of at least one subgradient, we show that all assumptions of Proposition 4.1 are fulfilled, which guarantees that $\partial L_{\boldsymbol{d}}(\vec{q})$ is non-empty.

By definition the loss function is positive: $L(\vec{q}, \boldsymbol{d}) \geq 0 > -\infty$. Furthermore, for any $\vec{q}, \boldsymbol{d}$ exist $\tilde{\mathcal{Q}}, \tilde{\mathcal{D}}$ bounded as in Definition 2 in Notz and Pibernik (2021), such that $\vec{q} \in \tilde{\mathcal{Q}}, \boldsymbol{d} \in \tilde{\mathcal{D}}$. Therefore, following Lemma 1 in Notz and Pibernik (2021), the loss function is bounded: $\forall \vec{q} \in \tilde{\mathcal{Q}}, \boldsymbol{d} \in \tilde{\mathcal{D}} : L(\vec{q}, \boldsymbol{d}) < \infty$.

With these bounds, and because the loss function is convex (Proposition 4.4), all assumptions of Proposition 4.1 are fulfilled and $\forall \vec{q} \in int\mathcal{Q}$ the subdifferential is non-empty. Because $\mathcal{Q}$ is open, all elements are interior points such that $\vec{q} \in \mathcal{Q} \Rightarrow \vec{q} \in int\mathcal{Q}$, which concludes the proof.

Proof of Proposition 4.6

To obtain the stated expression for a subgradient of the loss function, we first derive the primal Lagrange function of Problem 4.22 and prove strong duality, and then use the results of Theorem 4.1.

Let α_i, β_j, γ_{ij} and δ_{ij} be Lagrange multipliers, then we obtain for the primal Lagrangian of (4.22):

$$\begin{aligned} L_{\text{primal}} = & \sum_i c_i d_i - \sum_{i,j} a_{ij} y_{ij} \\ & + \sum_i \alpha_i \left(\sum_j y_{ij} - d_i \right) + \sum_j \beta_j \left(\sum_i y_{ij} - q_j \right) \\ & - \sum_{i,j} \gamma_{ij} y_{ij} + \sum_{i<j} \delta_{ij} y_{ij}, \end{aligned} \tag{C.9}$$

where $i, j = 1...I$. Because the objective function is linear in $\{y_{ij}\}$ and therefore convex, all constraints of Problem 4.22 are affine in the primal variables $\{y_{ij}\}$, and $\{y_{ij}\} = 0$ is a feasible solution with 0 being a relative interior point of the domain of definition $\mathbb{R}^{I \times I}$, the Slater condition is fulfilled and strong duality holds (see Section 5.2.3 in Boyd and Vandenberghe 2004). Therefore, by the Karush-Kuhn-Tucker (KKT) conditions, the partial deriva-

tives of L_{primal} with respect to the primal variables y_{ij} vanish, which is a necessary and sufficient condition for optimality (Boyd and Vandenberghe 2004, p. 244):

$$\begin{aligned} \text{for } i < j\text{: } & \frac{\partial L_{\text{primal}}}{\partial y_{ij}} = -a_{ij} + \alpha_i + \beta_j - \gamma_{ij} + \delta_{ij} = 0 \\ \text{for } i \geq j\text{: } & \frac{\partial L_{\text{primal}}}{\partial y_{ij}} = -a_{ij} + \alpha_i + \beta_j - \gamma_{ij} \qquad = 0. \end{aligned} \tag{C.10}$$

Therefore, we can express the primal Lagrangian as

$$L_{\text{primal}} = \sum_i (c_i - \alpha_i)\, d_i - \sum_j \beta_j q_j. \tag{C.11}$$

Because L_{primal} is independent of the primal variables, we obtain for the dual Lagrangian:

$$L_{\text{dual}} = \inf_{\{y_{ij}\}} L_{\text{primal}} = L_{\text{primal}}, \tag{C.12}$$

from which we derive for the dual problem:

$$\begin{aligned} \max_{\{\alpha_i\},\{\beta_j\}} & \sum_i (c_i - \alpha_i)\, d_i - \sum_j \beta_j q_j \\ \text{s.t. } & \alpha_i,\ \beta_j \geq 0\ \forall i, j \\ & \alpha_i + \beta_j \geq a_{ij}\ \forall i \geq j, \end{aligned} \tag{C.13}$$

with the solution $\beta^*_{j,\vec{d},\vec{q}}$.

We further observe that the loss function (4.21) follows the form of (4.17) with $f_{0,\boldsymbol{d}}(\vec{q}) = \Pi^*(\boldsymbol{d}) + \sum_j f_j q_j$ convex in $\vec{q}$ and differentiable, and $|f_{0,\boldsymbol{d}}(\vec{q})| < \infty$ for all finite $\boldsymbol{d}$ and $\vec{q}$; and $f^*_{1,\boldsymbol{d}}(\vec{q}) = \sum_t -\pi(\vec{d^t}, \vec{q})$ jointly convex in $\vec{q}$ (Proposition 4.4), with $|f^*_{1,\boldsymbol{d}}(\vec{q})| < \infty$ for all $\boldsymbol{d}$ and $\vec{q}$.

Then, because

$$\frac{\partial f_{0,\boldsymbol{d}}(\vec{q})}{\partial q_j} = f_j, \tag{C.14}$$

and because a subgradient of $f^*_{1,\boldsymbol{d}}(\vec{q})$ is given as $(\vec{s}_{f^*_1,\boldsymbol{d},\vec{q}})_j = -\sum_t \beta^*_{j,\vec{d^t},\vec{q}}$ (Theorem 4.1), we obtain the expression to be proven.

Proof of Proposition 4.7

Convexity of the loss function defined in (4.26) has been shown in Proposition 7 in Notz et al. (2020).

Proof of Proposition 4.8

Similar as in the proof of Proposition 4.5, we show that all assumptions of Proposition 4.1 are fulfilled, which guarantees that at least one subgradient exists.

For any $\vec{q}$ with $||\vec{q}||_\infty < \infty$ and any $\vec{d} \geq 0$ with $||\vec{d}||_\infty < \infty$, we can bound the loss function by bounding the backlogging cost. A feasible solution to Problem 4.27 is given as

$$m_t = \sum_{\tau=1}^{t} d_\tau, \tag{C.15}$$

which allows us to bound the backlogging cost as

$$C_{\text{backlog}}(\vec{q}, \vec{d}) \leq (c_2 + c_3) \sum_t d_t. \tag{C.16}$$

Therefore, the loss function is bounded as

$$|L(\vec{q}, \vec{d})| \leq c_q(\tau_a + \tau_b)||\vec{q}||_\infty + (c_2 + c_3) \sum_t d_t < \infty. \tag{C.17}$$

Based on this bound, and because the loss function is convex (Proposition 4.7), all assumptions of Proposition 4.1 are fulfilled and $\forall \vec{q} \in int\mathcal{Q}$ the subdifferential $\partial L_{\vec{d}}(\vec{q})$ is non-empty. Because $\mathcal{Q}$ is open, all elements are interior points $\vec{q} \in \mathcal{Q} \Rightarrow \vec{q} \in int\mathcal{Q}$, which concludes the proof.

Proof of Proposition 4.9

Similar as in the proof of Proposition 4.6, we first derive the primal Lagrange function of Problem 4.27 and prove strong duality, and then use the results of Theorem 4.1.

Let β_t, γ_t and δ_0 be Lagrange multipliers, then we obtain for the primal

Lagrangian:

$$\begin{aligned} L_{\text{primal}} =& c_2 m_{20} + c_3 m_{14} \\ &+ \sum_{t=1}^{6} \beta_t \left(m_{t-1} + d_t - m_t \right) + \sum_{t=7}^{14} \beta_t \left(m_{t-1} + d_t - m_t - q_a \right) \\ &+ \sum_{t=15}^{20} \beta_t \left(m_{t-1} + d_t - m_t - q_b \right) - \sum_{t=1}^{20} \gamma_t m_t + \delta_0 m_0. \end{aligned} \tag{C.18}$$

Because the objective function is linear in $\{m_t\}$ and therefore convex, all constraints of Problem 4.27 are affine in the primal variables $\{m_t\}$, and $m_t = \sum_{\tau=1}^{t} d_\tau$—which equals a continuous queue build-up with all arriving demand being added to the queue—is a feasible solution and a relative interior point of the domain of definition $\mathbb{R}^T$, the Slater condition is fulfilled and strong duality holds (see Section 5.2.3 in Boyd and Vandenberghe 2004). Therefore, the KKT conditions, which state that the partial derivatives of L_{primal} with respect to the primal variables $\{m_t\}$ equal 0, provide necessary and sufficient conditions for optimality (Boyd and Vandenberghe 2004, p. 244):

$$\begin{aligned} \text{for } t = 0\text{:} \quad & \frac{\partial L_{\text{primal}}}{\partial m_t} = \beta_1 + \delta_0 && = 0 \\ \text{for } t = 14\text{:} \quad & \frac{\partial L_{\text{primal}}}{\partial m_t} = \beta_{15} - \beta_{14} - \gamma_{14} + c_3 && = 0 \\ \text{for } t = 20\text{:} \quad & \frac{\partial L_{\text{primal}}}{\partial m_t} = c_2 - \beta_{20} - \gamma_{20} && = 0 \\ \text{for } t \neq 0, 14, 20\text{:} \quad & \frac{\partial L_{\text{primal}}}{\partial m_t} = \beta_{t+1} - \beta_t - \gamma_t && = 0. \end{aligned} \tag{C.19}$$

Using these results, the primal Lagrangian can be expressed as

$$L_{\text{primal}} = \sum_{t=1}^{20} \beta_t d_t - \sum_{t=7}^{14} \beta_t q_a - \sum_{t=15}^{20} \beta_t q_b. \tag{C.20}$$

Because L_{primal} is independent of the primal variables, we obtain for the dual Lagrangian:

$$L_{\text{dual}} = \inf_{\{m_t\}} L_{\text{primal}} = L_{\text{primal}}, \tag{C.21}$$

from which we derive for the dual problem:

$$\begin{aligned} \max_{\{\beta_t\}} & \sum_{t=1}^{20} \beta_t d_t - \sum_{t=7}^{14} \beta_t q_a - \sum_{t=15}^{20} \beta_t q_b \\ \text{s.t. } & \beta_{t+1} \geq \beta_t \ \forall t \neq 14, 20 \\ & \beta_{15} \geq \beta_{14} - c_3 \\ & \beta_{20} \leq c_2, \end{aligned} \tag{C.22}$$

with the solution $\{\beta^*_{t,\vec{d},\vec{q}}\}$.

We observe that the loss function (4.26) follows the form of (4.17) with $f_{0,\vec{d}}(\vec{q}) = c_q(\tau_a q_a + \tau_b q_b)$ convex in $\vec{q}$ and differentiable, $|f_{0,\vec{d}}(\vec{q})| < \infty$ for all finite $\vec{d}$ and $\vec{q}$. Also, $f^*_{1,\vec{d}}(\vec{q}) = C_{\text{backlog}}(\vec{q}, \vec{d})$ with $f^*_{1,\vec{d}}(\vec{q})$ jointly convex in $\vec{q}$ (Proposition 4.7) and $|f^*_{1,\vec{d}}(\vec{q})| < \infty$ for all $\vec{d}$ and $\vec{q}$. We further obtain

$$\frac{\partial f_{0,\vec{d}}(\vec{q})}{\partial q_i} = c_q \tau_i, \tag{C.23}$$

and a subgradient of $f^*_{1,\vec{d}}(\vec{q})$ as

$$(\vec{s}_{f^*_1,\vec{d},\vec{q}})_j = \begin{cases} -\sum_{t=7}^{14} \beta^*_{t,\vec{d},\vec{q}} & \text{for } j = 1 \\ -\sum_{t=15}^{20} \beta^*_{t,\vec{d},\vec{q}} & \text{for } j = 2, \end{cases} \tag{C.24}$$

using the results of Theorem 4.1, which concludes the proof.

C.2 Detailed Description of Aggregate Features

A detailed description of the 162 (mail sorting case) and 142 (aviation maintenance case) features used in our numerical analyses can be found in Notz and Pibernik (2021) and Notz et al. (2020). In the following, we provide details on the feature aggregation used in Section 4.5.5 to obtain explanations of the prescribed capacities based on SHAP values.
For the mail sorting case we define the following aggregate features:

- 1-3 weeks lag, by service line and total of all service lines: aggregate of

daily lagged demands and total week lagged demand,

- 1 month lag, by service line and total of all service lines: aggregate of daily lagged demands and total week lagged demand,
- 1 year lag, by service line and total of all service lines: aggregate of daily lagged demands and total week lagged demand,
- difference between 1 and 2 weeks lagged demand, by service line: aggregate of daily differences.

This yields a total of 23 aggregate features. Combined with the remaining 11 date-based features, the 11 indicator features representing the relation to public holidays, and 5 features describing sums of lagged demands (1-2 weeks and 1-3 weeks) and summed differences between weeks (see Appendix C in Notz and Pibernik 2021 for details), we obtain a total of 50 features used to derive SHAP value-based explanations for the mail sorting case.
For the aviation maintenance case we define the following aggregate features:

- 1-7 days lagged demand: aggregate of both shifts and totals,
- 14, 21, and 28 days lagged demand: aggregate of both shifts and totals,
- sums of 1-3 or 1-7 days lagged demand: aggregate of both shifts and totals,
- indicator if a public holiday is 1-3 days before the day of interest: aggregate of Germany and US,
- indicator if a public holiday is 1-3 days after the day of interest: aggregate of Germany and US,
- weekday average of expected demand: aggregate of contracted and updated delivery date,
- parts shipped before day of interest by four country groups: aggregate across 1-29 days before day of interest,

- demand for which a process ID was created: aggregate across 1-29 days before day of interest.

This yields a total of 24 aggregate features. Combined with the remaining 22 date-based features and 4 features describing expected demand based on contracted or updated delivery date and the percentage difference to the weekday average (see Appendix B in Notz et al. 2020 for details), we obtain a total of 50 features used to derive explanations based on SHAP values for the aviation maintenance capacity case.

C.3 Detailed Description of Approaches for Numerical Evaluation

In this section we describe the implementation of each of the approaches presented in Section 4.5.3 in more detail.

C.3.1 Subgradient Tree Boosting

We implement Algorithm 4.1 presented in Section 4.3.1 to obtain the prescription function $\vec{q}^{STB}(\cdot)$. In particular, we implement the loss functions (4.21) and (4.26) and use *Gurobi Optimizer* for the sample average approximation in Steps 1 and 7 of Algorithm 4.1. The subgradients of the loss functions $\vec{s}_L$ are calculated as described in (4.23) and (4.28), by solving Problems 4.24 and 4.29 using *Gurobi Optimizer*.[61] The decision tree learning (Step 5) is done using the *DecisionTreeRegressor* class of the *scikit-learn* package for python, which allows us to learn bi- and trivariate regression trees with a maximum of five leaf nodes.[62] For the mail sorting case, the hyper parameters $\xi = 0.3...1.0$ (fraction of data samples used for tree learning in each iteration), $\nu = 0.001...0.5$

[61] In case of $\vec{q}_k^{STB}(\vec{x}^n) = 0$ in one dimension j, we approximate the subgradient by setting $(\vec{q}_k^{STB}(\vec{x}^n))_j = 0.001$, because the subgradient is not defined for vanishing $\vec{q}_k^{STB}(\vec{x}^n)$. However, assuming $\nu < 1$, this can only occur when the initialization (SAA) assigns zero capacity to a service line (mail sorting case) or shift (aviation maintenance case), which is almost never the case in realistic settings.

[62] Hastie et al. (2009) describe a number of 4 to 8 leaf nodes for a base learner as common range (see Hastie et al. 2009, p. 363).

(shrinkage factor or learning rate) and $K = 1...4000$ (number of iterations and trees in final model) are tuned using simple cross-validation with 2/3 of the training data used to train the model, and the remaining 1/3 of the training data to evaluate the achieved profit. The hyper parameters are tuned similarly for the aviation maintenance case, where $\xi = 0.3...1.0$, $\nu = 0.001...0.5$ and $K = 1...2000$ with 3/4 of the training data used to train the model, and the remaining 1/4 of the training data used to evaluate the incurred cost of the prescriptions.

We use the estimated prescription function $\vec{q}^{STB}(\cdot)$ to determine the capacity decisions for each week (mail sorting case) or each day (aviation maintenance case) of the test period.

C.3.2 Weighted SAA

The implementation of the wSAA approach consists of two steps: we first calculate the random forest weights, and then solve the wSAA optimization problem. To calculate the weights, we train random forest models and use the weight function

$$w_n^{\mathrm{RF}}(\vec{x}) = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathbb{1}[\mathcal{R}^k(\vec{x}) = \mathcal{R}^k(\vec{x}^n)]}{\sum_{j=1}^{N} \mathbb{1}[\mathcal{R}^k(\vec{x}) = \mathcal{R}^k(\vec{x}^j)]}, \tag{C.25}$$

as described in Notz and Pibernik (2021) and Notz et al. (2020).

We then solve the wSAA optimization problem

$$\vec{q}^{\,\mathrm{wSAA}}(\vec{x}) = \arg\min_{\vec{q} \in \mathcal{Q}} \sum_{n=1}^{N} w_n^{\mathrm{RF}}(\vec{x}) L(\vec{q}, \boldsymbol{d}^n) \tag{C.26}$$

with the loss functions $L(\vec{q}, \boldsymbol{d})$ defined in (4.21) and (4.26) for each instance of the test period using *Gurobi Optimizer*.

C.3.3 Kernelized ERM

Similar to the implementation of wSAA, we first calculate the kernel function defined in (4.8) using the same random forest models as for wSAA, and then

solve the kERM approach for the loss functions $L(\vec{q}, \boldsymbol{d})$ defined in (4.21) and (4.26) using *Gurobi Optimizer*. The kERM prescription function is given as:

$$\vec{q}^{\,\mathrm{kERM}}(\vec{x}) = \sum_{n=1}^{N} \vec{u}^{n} K\left(\vec{x}^{n}, \vec{x}\right) - \vec{b}, \tag{C.27}$$

where, for the mail sorting case, the components of the coefficients $\vec{u}^{n}$ are defined as $u_j^n = \frac{1}{2\lambda_j}\left(\sum_{t=1}^{T}(\beta_j^{tn}) + \epsilon_j^n - f_j\right)$ and β_j^{tn}, ϵ_j^n are the solution to

$$\begin{aligned}
\max_{\{\alpha_i^{tn}\},\{\beta_j^{tn}\},\{\epsilon_j^n\}} L_{\mathrm{dual}} := & -\sum_{j=1}^{I}\lambda_j \sum_{p,q=1}^{N}\left(u_j^p u_j^q K(\vec{x}^p, \vec{x}^q)\right) + \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{t=1}^{T}(c_i - \alpha_i^{tn}) d_i^{tn} \\
\text{s.t. } & \alpha_i^{tn},\ \beta_j^{tn},\ \epsilon_j^n \geq 0\ \forall i, n, t \\
& \alpha_i^{tn} + \beta_j^{tn} \geq a_{ij}\ \forall i \geq j,\ \forall n, t \\
& \sum_{n=1}^{N} u_j^n = 0\ \forall j,
\end{aligned} \tag{C.28}$$

as described in Notz and Pibernik (2021). Similarly, for the aviation maintenance case, we define the components of $\vec{u}^{n}$ as

$$\begin{aligned}
u_1^n &= \frac{1}{2\lambda_1}\left(\sum_{t=7}^{14}\alpha_t^n + \delta_1^n - \frac{c_q \tau_a}{N}\right) \\
u_2^n &= \frac{1}{2\lambda_2}\left(\sum_{t=15}^{20}\alpha_t^n + \delta_2^n - \frac{c_q \tau_b}{N}\right),
\end{aligned} \tag{C.29}$$

where α_t^n and δ_j^n are the solution to

$$\begin{aligned}
\max_{\{\alpha_t^n\},\{\delta_j^n\}} L_{\mathrm{dual}} := & -\sum_{j=1}^{2}\lambda_j \sum_{p=1}^{N}\sum_{q=1}^{N}\left(u_j^p u_j^q K(\vec{x}^p, \vec{x}^q)\right) + \sum_{n,t=7}^{t=20}\alpha_t^n d_t^n + \sum_{n,t=1}^{t=6}\alpha_7^n d_t^n \\
\text{s.t. } & \alpha_t^n,\ \delta_j^n \geq 0\ \forall j, t, n \\
& \alpha_{t+1}^n \geq \alpha_t^n\ \forall n, t \neq 14 \\
& \alpha_{t+1}^n \geq \alpha_t^n - \frac{c_3}{N}\ \forall n, t = 14 \\
& \alpha_{20}^n \leq \frac{c_2}{N}\ \forall n,
\end{aligned} \tag{C.30}$$

where $n = 1...N$, $j = 1...2$ and $t = 7...20$, as described in Notz et al. (2020). We then apply the prescription function $\vec{q}^{\,\mathrm{kERM}}(\cdot)$ to each instance of the test period.

C.3.4 SAA

To estimate the SAA prescriptions we solve

$$\vec{q}^{\,\mathrm{SAA}}(\vec{x}) = \underset{\vec{q}\in\mathcal{Q}}{\arg\min} \frac{1}{N} \sum_{n=1}^{N} L(\vec{q}, \boldsymbol{d}^n) \tag{C.31}$$

with the loss functions $L(\vec{q}, \boldsymbol{d})$ defined in (4.21) and (4.26) using *Gurobi Optimizer*. This optimization problem is equivalent to that of wSAA when setting $w_n(\vec{x}) = 1/N$.

C.4 Impact of the Number of STB Iterations on Performance

In this section we study the dependence of the prescription performance of STB on the number of iterations K. First, we analyze the dependence theoretically, based on out-of-sample performance guarantees, and second, we conduct numerical experiments, in which we vary K and determine the prescription performance.

The theoretical analysis is based on the concept of Rademacher complexities, which was extended to multivariate OM settings in Bertsimas and Kallus (2020). We show in the following that the empirical Rademacher complexity of the function space used by STB increases with K, which emphasizes the requirement to limit K in practical applications.

Proposition C.1. *Let $\mathcal{F}_0$ be the function space of the STB base learners (e.g., decision trees) with empirical Rademacher complexity $\mathrm{Rad}_N(\mathcal{F}_0, S_N)$ over S_N, let K be the number of STB iterations, $\nu > 0$ the constant shrinkage factor, and let $\mathcal{F} = \sum_{k=1}^{K} \nu\mathcal{F}_0$ be the prescription function space of the STB approach.*

Then, the empirical Rademacher complexity of $\mathcal{F}$ is bounded as

$$\mathrm{Rad}_N(\mathcal{F}, S_N) \leq \nu K \,\mathrm{Rad}_N(\mathcal{F}_0, S_N). \tag{C.32}$$

Proof of Proposition C.1: Following Theorem 12 Part 7 in Bartlett and Mendelson (2002), the empirical Rademacher complexity of a sum of function spaces $\mathcal{F}_k$ is bounded as

$$\mathrm{Rad}_N\left(\sum_{k=1}^{K} \mathcal{F}_k, S_N\right) \leq \sum_{k=1}^{K} \mathrm{Rad}_N(\mathcal{F}_k, S_N). \tag{C.33}$$

Furthermore, following Theorem 12 Part 3 in Bartlett and Mendelson (2002), the empirical Rademacher complexity of scalar multiplication of a function space $\mathcal{F}$ is given as

$$\mathrm{Rad}_N(\nu\mathcal{F}, S_N) = |\nu| \,\mathrm{Rad}_N(\mathcal{F}, S_N). \tag{C.34}$$

Combining both, we derive for the Rademacher complexity of $\mathcal{F}$:

$$\begin{aligned} \mathrm{Rad}_N(\mathcal{F}, S_N) &= \mathrm{Rad}_N\left(\sum_{k=1}^{K} \nu\mathcal{F}_0, S_N\right) \\ &\leq \nu \sum_{k=1}^{K} \mathrm{Rad}_N(\mathcal{F}_0, S_N) \\ &= \nu K \,\mathrm{Rad}_N(\mathcal{F}_0, S_N), \end{aligned} \tag{C.35}$$

which concludes the proof.

We observe that this bound on the empirical Rademacher complexity $\mathrm{Rad}_N(\mathcal{F}, S_N)$ increases linearly in the number of STB iterations K, therefore, the function space complexity increases with K. Proposition C.2 shows how this leads to a bound on the out-of-sample performance that increases with K.

Proposition C.2. *(Following Bertsimas and Kallus 2020) Assume S_N, generated by iid sampling from a joint distribution of $X \times D$, $L(\vec{q}, \boldsymbol{d})$ the loss function of the mail sorting case (4.21) or the loss function of the aviation*

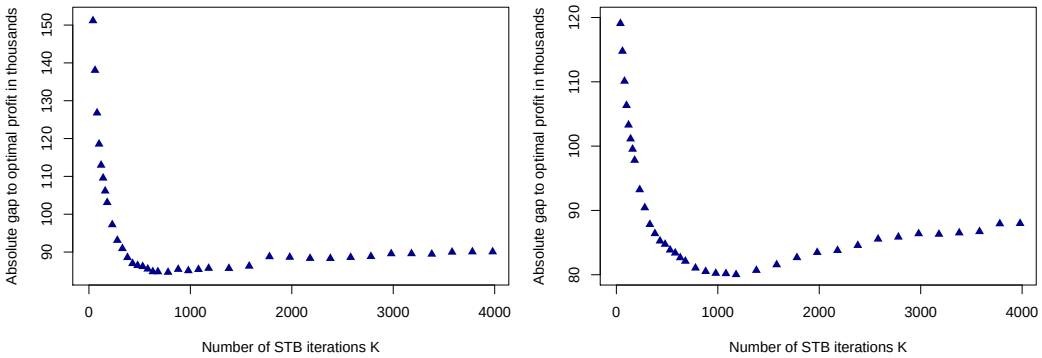
maintenance case (4.26), and let $\delta > 0$*. Then, with probability of at least* $1-\delta$ *for any function* $\vec{q}(\cdot) \in \mathcal{F} = \sum_{k=1}^{K} \nu\mathcal{F}_0$*, the true risk is bounded as*

$$R(\vec{q}(\cdot)) \leq R_N(\vec{q}(\cdot)) + 3\bar{l}\sqrt{\frac{\log(2/\delta)}{2N}} + M_{Lip}\nu K \,\mathrm{Rad}_N(\mathcal{F}_0, S_N), \tag{C.36}$$

where $\bar{l}$ *is the bound and* M_{Lip} *is the Lipschitz constant of* $L(\vec{q}, \boldsymbol{d})$*.*

Proof of Proposition C.2: When combining the results of Proposition C.1 with Theorem 4 in Notz and Pibernik (2021) for the mail sorting case or Theorem 8 in Bertsimas and Kallus (2020) and Propositions 8 (cost function is bounded) and 9 (cost function is equi-Lipschitz) in Notz et al. (2020) for the aviation maintenance case, the stated bound on the true risk directly follows.

The RHS of the result of Proposition C.2 demonstrates the common bias-variance trade-off between i) choosing a low complexity of the function space $\mathcal{F}$ (small Rademacher complexity, high bias), which leads to a large empirical (in-sample) risk $R_N(\vec{q}(\cdot))$, and ii) choosing a high complexity of the function space $\mathcal{F}$ (large Rademacher complexity, high variance), which leads to a small $R_N(\vec{q}(\cdot))$. This trade-off can be tuned by varying K, which determines the complexity of the function space $\mathcal{F}$ (Proposition C.1). The result of Proposition C.2 therefore underlines the need for *early stopping*, which is a strategy to prevent overfitting by choosing a finite number of STB iterations $K < \infty$, because the bound on the out-of-sample performance grows with K.



(a) Service Level=50%, $\nu = 0.01$, $\xi = 1.0$.(b) Service Level=80%, $\nu = 0.01$, $\xi = 0.3$.

Figure C.1: Absolute gap to optimal profit for the mail sorting case.

Our numerical experiments on the dependence of the STB prescription performance on the number of iterations K lead to consistent results. Figure C.1 shows the absolute gap to optimal profit for the mail sorting case for cost parameters that induce optimal service levels of 50 percent (Figure C.1a) and 80 percent (Figure C.1b). For both cost parameter settings the out-of-sample performance of STB first increases with the number of iterations K (most strongly for up to 500 iterations), which indicates the high bias regime, then reaches a minimum gap to optimal profit, and finally decreases with K, which indicates the high variance regime and suggests that the STB approach overfits the data for large K. The optimal value of K depends on the planning problem, the cost parameters and the data set, and is therefore typically determined through hyper parameter tuning.

List of Abbreviations

AI Artificial Intelligence

AMSSP Approximated Multi-Shift Staffing Problem

BK Bertsimas and Kallus (2020)

BR Ban and Rudin (2019)

CC Coefficient of Correlation

CV Coefficient of Variation

DRL Deep Reinforcement Learning

ERM Empirical Risk Minimization

GAN Generative Adversarial Network

GDP Gross Domestic Product

iid Independent and Identically Distributed

INFORMS Institute for Operations Research and the Management Sciences

kERM Kernelized Empirical Risk Minimization

KKT Karush-Kuhn-Tucker

kNN k-nearest-neighbors

LB Load Balancing

MC Monte Carlo

MRO Maintenance, Repair, and Overhaul

MSSP Multi-Shift Staffing Problem

NV Newsvendor

OM Operations Management

OP Optimization Prediction

OR/MS Operations Research and Management Science

PDE Partitioned Distribution Estimation

RBF Radial Basis Function

RF Random Forest

RHS Right-hand side

RKHS Reproducing Kernel Hilbert Space

SAA Sample Average Approximation

SEO Sequential Estimation and Optimization

SHAP Shapley additive explanation

SIPP Stationary Independent Period-by-Period

SL Service Level

STB Subgradient Tree Boosting

SVM Support Vector Machine

SVR Support Vector Regression

TBATS Trigonometric, Box-Cox transf., ARMA errors, Trend, Seasonality

wSAA Weighted Sample Average Approximation

List of Figures

List of Tables

Bibliography

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–18, New York, NY. ACM Press.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press, Boston, MA.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akcay, A., Biller, B., and Tayur, S. (2011). Improved inventory targets in the presence of limited historical demand data. *Manufacturing & Service Operations Management*, 13(3):297–309.

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Ban, G.-Y. (2020). Confidence intervals for data-driven inventory policies with demand censoring. *Operations Research*, 68(2):309–326.

Ban, G.-Y. and Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bassamboo, A., Randhawa, R. S., and Zeevi, A. (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686.

Bassok, Y., Anupindi, R., and Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research*, 47(4):632–642.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (1993). *Nonlinear programming: Theory and algorithms*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, New York, NY, second edition.

Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J.-F., Vincent, P., and Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219.

Bertsekas, D. P. (1995). *Nonlinear programming*. Athena Scientific, Belmont, MA.

Bertsimas, D. (2017). Vision for the new informs journal on optimization. *OR/MS Today*, 44(6):14–15.

Bertsimas, D., Delarue, A., Jaillet, P., and Martin, S. (2019). The price of interpretability. *arXiv preprint arXiv:1907.03419*.

Bertsimas, D., Gupta, V., and Kallus, N. (2018). Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Biau, G. and Cadre, B. (2017). Optimization by gradient boosting. *arXiv preprint arXiv:1707.05023*.

Bickel, P. J., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732.

Boyd, S., Duchi, J., and Vandenberghe, L. (2018). Subgradients: Notes for EE364b, Stanford University, Spring 2014–2015.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY.

Breiman, L. (2000). *Some infinity theory for predictor ensembles*. Technical Report 577, Statistics Department, University of California, Berkeley, CA.

Brynjolfsson, E. and McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. Norton & Company, New York, NY.

Bughin, J., Seong, J., Manyika, J., Chui, M., and Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy*. McKinsey Global Institute.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505.

Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury advanced series. Duxbury Thomson Learning, Pacific Grove, CA, second edition.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 785–794, New York, NY. ACM Press.

Chen, Y.-W. and Lin, C.-J. (2006). Combining SVMs with various feature selection strategies. In Guyon, I., Nikravesh, M., Gunn, S., and Zadeh, L. A., editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 315–324. Springer, Berlin, Heidelberg.

Cherkassky, V. and Ma, Y. (2002). Selection of meta-parameters for support vector regression. In Dorronsoro, J. R., editor, *Artificial Neural Networks — ICANN 2002*, volume 2415 of *Lecture Notes in Computer Science*, pages 687–693. Springer, Berlin, Heidelberg.

Cheung, W. C. and Simchi-Levi, D. (2019). Sampling-based approximation schemes for capacitated stochastic inventory control models. *Mathematics of Operations Research*, 44(2):668–692.

Choi, T.-M., Wallace, S. W., and Wang, Y. (2018). Big data analytics in operations management. *Production and Operations Management*, 27(10):1868–1883.

Davies, A. and Ghahramani, Z. (2014). The random forest kernel and other kernels for big data from random partitions. *arXiv preprint arXiv:1402.4293*.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527.

Defraeye, M. and Van Nieuwenhuyse, I. (2016). Staffing and scheduling under nonstationary demand for service: A literature review. *Omega*, 58:4–25.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Giesecke, K., Liberali, G., Nazerzadeh, H., Shanthikumar, J. G., and Teo, C. P. (2018). Call for papers—Management Science—special issue on data-driven prescriptive analytics. *Management Science*, 64(6):2972.

Gijsbrechts, J., Boute, R. N., Van Mieghem, J. A., and Zhang, D. (2019). Can deep reinforcement learning improve inventory management? Performance on dual sourcing, lost sales and multi-echelon problems. *SSRN Electronic Journal*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA. MIT Press.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., and Rueckert, D. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, 65:167–175.

Grubb, A. and Bagnell, J. A. (2011). Generalized boosting algorithms for convex optimization. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 1209–1216, Madison, WI. Omnipress.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, NY.

Harrison, J. M. and Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1):20–36.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, NY, second edition.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hogg, R. V., Tanis, E. A., and Zimmerman, D. L. (2015). *Probability and statistical inference*. Pearson, Upper Sadles River, NJ, ninth edition.

Horn, R. A. and Johnson, C. R. (2013). *Matrix analysis*. Cambridge University Press, New York, NY, second edition.

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

Kim, S.-H. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480.

Klabjan, D., Simchi-Levi, D., and Song, M. (2013). Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5).

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72.

Levi, R., Perakis, G., and Uichanco, J. (2015). The data-driven newsvendor problem: New bounds and insights. *Operations Research*, 63(6):1294–1306.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.

Liyanage, L. H. and Shanthikumar, J. G. (2005). A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4):341–348.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY. Curran Associates Inc.

Marsden, J. R., Coussement, K., and Benoit, D. (2020). Call for papers: Special issue on interpretable data science for decision making. *Decision Support Systems*.

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. (1999). Boosting algorithms as gradient descent. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 512–518, Cambridge, MA. MIT Press.

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. (2000). Functional gradient techniques for combining hypotheses. In Smola, A. J., Bartlett, P. L., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large-Margin Classifiers*, pages 221–246. MIT Press, Cambridge, MA.

McAfee, A. and Brynjolfsson, E. (2017). *Machine, Platform, Crowd: Harnessing our digital future*. Norton & Company, New York, NY.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., de Fauw, J., and Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94.

Meir, R. and Rätsch, G. (2003). An introduction to boosting and leveraging. In Mendelson, S. and Smola, A. J., editors, *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia, February 11–22, 2002 Revised Lectures*, pages 118–183. Springer, Berlin, Heidelberg.

Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

MIT Technology Review Custom (2016). *The Rise of Data Capital*. MIT Technology Review Custom in partnership with Oracle.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Netessine, S., Dobson, G., and Shumsky, R. A. (2002). Flexible service capacity: Optimal investment and the impact of demand correlation. *Operations Research*, 50(2):375–388.

Newey, W. K. and McFadden, D. (1994). Chapter 36: Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

Notz, P. M. and Pibernik, R. (2021). Prescriptive analytics for flexible capacity management. *Management Science*.

Notz, P. M., Wolf, P. K., and Pibernik, R. (2020). Prescriptive analytics for a multi-shift staffing problem. *SSRN Electronic Journal*.

Otto, B., Auer, S., Cirullies, J., Jürjens, J., Menz, N., Schon, J., and Wenzel, S. (2016). *Industrial Data Space: Digital Sovereignity over Data*. Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.; Industrial Data Space e.V., München, Berlin.

Precht, M., Kraft, R., and Bachmaier, M. (2005). *Angewandte Statistik 1*. Oldenbourg Wissenschaftsverlag, München, seventh edition.

Ratliff, N., Bagnell, J. A., and Srinivasa, S. S. (2007). Imitation learning for locomotion and manipulation. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*, pages 392–397. IEEE.

Ratliff, N., Bradley, D., Bagnell, J. A., and Chestnutt, J. (2006). Boosting structured prediction for imitation learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 1153–1160, Cambridge, MA. MIT Press.

Ratliff, N. D., Silver, D., and Bagnell, J. A. (2009). Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 27(1):25–53.

Robbins, H. (1955). A remark on stirling's formula. *The American Mathematical Monthly*, 62(1):26–29.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series, volume 28. Princeton University Press, Princeton, NJ.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sandulescu, V. and Chiru, M. (2016). Predicting the future relevance of research institutions - the winning solution of the KDD cup 2016. *arXiv preprint arXiv:1609.02728*.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Denison, D. D., Hansen, M. H., Holmes, C. C., Mallick, B., and Yu, B., editors, *Nonlinear Estimation and Classification*, volume 171 of *Lecture Notes in Statistics*, pages 149–171. Springer, New York, NY.

Schapire, R. E. and Freund, Y. (2012). *Boosting: Foundations and algorithms*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

Schölkopf, B., Luo, Z., and Vovk, V. (2013). *Empirical Inference*. Springer, Berlin, Heidelberg.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press, New York, NY.

Shapiro, A. (2003). Monte carlo sampling approach to stochastic programming. *ESAIM: Proceedings*, 13:65–73.

Shapiro, A. and Homem-de-Mello, T. (1998). A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming*, 81(3):301–325.

Shapiro, A. and Kleywegt, A. (2002). Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542.

Shi, C., Chen, W., and Duenyas, I. (2016). Technical note—nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research*, 64(2):362–370.

Shumsky, R. A. and Zhang, F. (2009). Dynamic capacity management with substitution. *Operations Research*, 57(3):671–684.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D.

(2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Information science and statistics. Springer, New York, NY.

Stolletz, R. (2008). Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research*, 190(2):478–493.

Studniarski, M. (1989). An algorithm for calculating one subgradient of a convex function of two variables. *Numerische Mathematik*, 55(6):685–693.

Taigel, F. and Meller, J. (2020). Prescriptive call center staffing. *SSRN Electronic Journal*.

Taigel, F., Meller, J., and Rothkopf, A. (2019). Data-driven capacity management with machine learning: A novel approach and a case-study for a public service office. In Yang, H. and Qiu, R., editors, *Advances in Service Science*, INFORMS-CSS 2018, Springer Proceedings in Business and Economics, pages 105–115, Cham. Springer.

Vaidya, P. M. (1989). Speeding-up linear programming using fast matrix multiplication. In *30th Annual Symposium on Foundations of Computer Science*, pages 332–337. IEEE.

Van Mieghem, J. A. (2003). Commissioned paper: Capacity management, investment, and hedging: Review and recent developments. *Manufacturing & Service Operations Management*, 5(4):269–302.

Vapnik, V. N. (1991). Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, pages 831–838, Denver, CO. Morgan Kaufmann.

Vapnik, V. N. (1998). *Statistical learning theory*. A Wiley-Interscience publication. Wiley, New York, NY.

Vapnik, V. N. and Chervonenkis, A. Y. (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Proceedings of the USSR Academy of Sciences*, 181(4):781–783. English translation: *Soviet Mathematics Doklady* 9(4):915–918 (1968); English reprint: Chapter 2 in Schölkopf, B., Luo, Z., and Vovk, V. (2013). *Empirical Inference*. Springer, Berlin, Heidelberg.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280.

Vens, C. and Costa, F. (2011). Random forest based feature induction. In *2011 IEEE 11th International Conference on Data Mining*, pages 744–753. IEEE.

Vert, J.-P., Tsuda, K., and Schölkopf, B. (2004). A primer on kernel methods. In Schölkopf, B., Tsuda, K., and Vert, J.-P., editors, *Kernel Methods in Computational Biology*, pages 35–70. MIT Press, Cambridge, MA.

Volkovs, M., Yu, G. W., and Poutanen, T. (2017). Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017*, RecSys Challenge '17, pages 1–6, New York, NY. ACM Press.

Westerman, G., Bonnet, D., and McAfee, A. (2014). *Leading digital: Turning technology into business transformation*. Harvard Business Review Press, Boston, MA.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. N. (2000). Feature selection for SVMs. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 647–653, Cambridge, MA. MIT Press.

Yu, Y., Chen, X., and Zhang, F. (2015). Dynamic capacity management with general upgrading. *Operations Research*, 63(6):1372–1389.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579.

Eidesstattliche Erklärung (Statement of Academic Integrity)

Hiermit erkläre ich gemäß § 7 Abs. 2 Nr. 2 der Promotionsordnung der wirtschaftswissenschaftlichen Fakultät der Universität Würzburg, dass ich diese Dissertation eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters angefertigt habe. Ausgenommen davon sind jene Abschnitte, bei deren Erstellung ein Koautor mitgewirkt hat. Diese Abschnitte sind entsprechend gekennzeichnet und die Namen der Koautoren sind vollständig und wahrheitsgemäß aufgeführt. Bei der Erstellung der Abschnitte, bei denen ein Koautor mitgewirkt hat, habe ich einen signifikanten Beitrag geleistet, der meine eigene Koautorschaft rechtfertigt.

Außerdem erkläre ich, dass ich außer den im Schrifttumsverzeichnis angegebenen Hilfsmitteln keine weiteren benutzt habe und alle Stellen, die aus dem Schrifttum ganz oder annähernd entnommen sind, als solche kenntlich gemacht und einzeln nach ihrer Herkunft nachgewiesen habe.

Würzburg, den 18. August 2020

Pascal Markus Notz