Titre: Title:	Models for Integrated Demand Forecasting and Planning: Application to Large-scale Transportation Networks and Impact Assessment
Auteur: Author:	Greta Laage
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Laage, G. (2021). Models for Integrated Demand Forecasting and Planning: Application to Large-scale Transportation Networks and Impact Assessment [Ph.D. thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9122/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Models for Integrated Demand Forecasting and Planning: Application to Large-scale Transportation Networks and Impact Assessment

GRETA LAAGE
Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Mathématiques

Août 2021

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Models for Integrated Demand Forecasting and Planning: Application to Large-scale Transportation Networks and Impact Assessment

présentée par **Greta LAAGE**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Louis-Martin ROUSSEAU, président
Gilles SAVARD, membre et directeur de recherche
Emma FREJINGER, membre et codirectrice de recherche
Jonathan JALBERT, membre
Frederic SEMET, membre externe

DEDICATION

To my family

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the support of many that I would like to thank here.

I am indebted to my advisors Gilles Savard and Emma Frejinger, for their guidance and mentoring through the development of my research. Your wisdom and valuable advices allowed me to grow both scientifically and personally.

Special thanks to the co-authors of the third work presented in this thesis: Andrea Lodi and Guillaume Rabusseau. This work was integrated in a larger project at Ivado Labs with Air Canada, and I also wish to express my appreciation to all my dear colleagues over there, for their welcome and their support.

I gratefully acknowledge the close collaboration with personnel from different divisions of the Canadian National Railway Company. I am particularly grateful to Johanne Dandurand and Bill Mann, for always dedicating time to answer my numerous questions about the data and the operations.

I had the chance to be part of a fantastic learning group in Montréal. I would like to thank the GERAD and the CIRRELT staff for always being kind and assistive. A big thank you to all my friends from the lab. Antoine, Aurélien, Claudio, Giulia, Lucie, Luciano and Safae, we started as lab mates talking about optimization, and I feel very lucky now to have such great friends in my life. Special thanks to Julie and Stephanie, my dear friends, and office mates in non-covid times, for being constantly positive and motivating.

Many friends offered invaluable support throughout the difficulties of graduate life. Violette and Yohan, your dynamism and cheerfulness is a constant source of inspiration. Thank you also to Pierre-Emmanuel, Florian and Pauline, friends who are always near from afar.

Lastly, and most importantly, I thank my parents, Gilles and Sonia, and my sisters, Eudoxie, Marie and Louise for constantly questioning while being unconditionally encouraging, and an infinite source of advices. Eudoxie, thanks for the almost daily funny memes.

RÉSUMÉ

Pour une meilleure gestion de leurs opérations, les compagnies de transport ont mis en place des outils d'aide à la décision. Les décisions suggérées par ces outils reposent fortement sur les prévisions de demande, et peuvent affecter considérablement leur rentabilité. Les compagnies aériennes par exemple, utilisent des systèmes de gestion du revenu pour le transport de passagers. Les compagnies de transport de marchandises quant à elles, résolvent des problèmes de conception de réseaux de service, dont l'objectif est de définir les services à offrir sur leur réseau pour transporter la demande à coût minimal. Les prévisions de demande, la planification et la rentabilité des compagnies de transport sont donc très fortement reliées.

Dans cette thèse organisée en trois articles, nous réalisons des études interdisciplinaires portant sur l'intégration de la prévision de la demande et de la planification pour des réseaux de transport de grande taille, et sur l'analyse de l'impact sur des indicateurs de performance.

Tout d'abord, nous présentons pour la première fois le problème d'estimation de la demande périodique relatif aux décisions tactiques des compagnies de transport de marchandises. Ce problème incorpore à la fois le problème de prévision de la demande, et le problème de planification. Pour des raisons de complexité et de faisabilité, les modélisations cycliques et déterministes de ce dernier sont majoritairement utilisées en pratique pour les problèmes de grande taille. Celles-ci reposent sur l'hypothèse d'une demande fixe, connue et identique à chaque période de l'horizon de planification, appelée la *demande périodique*. Nous proposons une méthodologie en deux étapes pour son estimation, de façon à minimiser les coûts tactiques. L'objectif de la première étape est d'estimer la demande à venir pour toutes les commodités transportées, à chaque période. Nous développons et comparons des modèles de prévisions issus des statistiques et de l'apprentissage automatique utilisant d'importantes caractéristiques extraites des données. Puis, la deuxième étape définit la demande périodique comme la solution d'un problème multi-niveaux qui intègre le problème de planification et dont l'objectif est de transporter la demande prévue à coût minimal. C'est un problème difficile car les niveaux inférieurs sont de grande taille, non convexes, non différentiables et combinatoires. Nous le résolvons sur un petit ensemble discret de solutions réalisables. Des résultats numériques produits avec les données de la division intermodale de la Compagnie des Chemins de fer nationaux du Canada montrent que notre méthodologie permet de réduire considérablement les coûts par rapport au cas où la demande périodique est égale à la moyenne des prévisions, une pratique répandue chez les transporteurs.

Dans un deuxième temps, nous élargissons le problème d'estimation de la demande périodique

introduit précédemment, et définissons cette dernière comme étant une fonction linéaire de la moyenne des prévisions. Nous proposons donc une nouvelle formulation multi-niveaux du problème, dans laquelle les variables de décisions sont désormais les coefficients de la fonction linéaire pour toutes les commodités transportées. De par la complexité du problème, nous explorons deux pistes pour sa résolution: les méta-heuristiques et les algorithmes d'optimisation de boîte noire. Toutes deux exploitent la propriété de résolution séquentielle, affirmant que les niveaux inférieurs de la formulation peuvent être résolus séquentiellement lorsque les variables de premier niveau sont fixées. Nous proposons deux nouvelles méta-heuristiques de recherche locale, et comparons leurs résultats avec un logiciel de boîte noire dont les meilleures performances sont obtenues sur des problèmes avec peu de variables. Toutefois, les réseaux de transport de grande taille rencontrés en pratique transportent des centaines de commodités, ce qui correspond à des centaines de variables. Pour réduire le nombre de variables tout en conservant des bonnes solutions, nous présentons des approches heuristiques ayant pour but de créer des groupes de commodités ayant la même valeur de coefficient. Les groupes sont créés à partir de la structure du réseau ou de l'analyse de la distribution de la demande sur l'horizon de planification. Cette méthodologie surpasse les performances indiquées dans un premier temps sur le problème du réseau intermodal, et génère une amélioration des coûts. De plus, les heuristiques de groupement permettent non seulement d'obtenir les meilleurs résultats, mais elle rendent également l'utilisation des logiciels de boîte noire possible pour les applications de grande taille.

Dans un dernier temps, nous nous concentrons sur l'évaluation de l'impact d'une amélioration apportée à un outil d'aide à la décision. L'objectif est d'estimer, pour un transporteur, l'impact sur des indicateurs de performance tels que le revenu et le chiffre d'affaires. Nous considérons le cas particulier du système de gestion du revenu d'une compagnie aérienne, et proposons une approche qui évalue l'impact sur le revenu, indépendamment du type d'amélioration apportée au système. Le problème d'estimation de l'impact est intrinsèquement difficile car le revenu qui aurait été obtenu sans l'amélioration, en maintenant le système inchangé, n'est pas observable. L'approche que nous proposons repose sur des modèles dont l'objectif est d'estimer ce revenu, et qui nécessitent uniquement les données d'observation du revenu. L'impact est alors calculé comme étant la différence entre le revenu observé, résultant de l'amélioration, et le revenu estimé. Des résultats numériques issus des données d'Air Canada montrent que les modèles d'estimation du revenu non observable sont précis, et permettent d'estimer même un impact de petite ampleur.

ABSTRACT

Decision-making systems deployed in transportation companies are used for efficiency and better planning of their operations. Such systems rely on demand forecasts, and the decisions they suggest can substantially affect the profitability of a carrier. Airlines, for instance, use Revenue Management Systems for their passenger-based activities. Freight carriers solve service network design problems for their tactical planning, optimizing the services to offer to transport the demand at minimal cost. Hence the strong link between demand forecasting, planning and profitability.

In this research organized in three papers, we carry out interdisciplinary studies that focus on integrating demand forecasting and planning for large-scale transportation networks, and analyzing the impact on key performance indicators.

First, we introduce the periodic demand estimation problem that integrates demand forecasting with planning for tactical decisions of a freight carrier. Cyclic and deterministic formulations of tactical planning problems prevail in practice, due to the complexity and feasibility of large-scale applications. This is our focus. Those formulations assume that the demand is fixed, known and repeated at each period of the planning horizon. We refer to the latter as *periodic demand*. We propose a two-step methodology to estimate it. The first step consists in forecasting demand for each commodity transported in the network at each period of the planning horizon. We develop and compare models based on statistics and machine learning literature exploiting important features extracted from the data. The second step defines the periodic demand as a solution to a multilevel mathematical program which integrates the planning problem and aims at transporting the forecasted demand at minimal cost. The lower levels are non-convex, non-differentiable and combinatorial, hence the difficulty of solving the multilevel program. We apply the methodology on the intermodal network of the Canadian National Railway Company. Computational experiments on a limited feasible set of periodic demand show that this methodology allows substantial reduction of the tactical planning costs compared to the common practice which consists in taking the periodic demand as the average of the demand forecasts.

Second, we extend the periodic demand estimation problem and define the periodic demand as a deviation from the average of the demand forecasts. We propose a new multilevel formulation for the periodic demand estimation problem, where the decision variables are now the deviation coefficients for all commodities transported in the network. Due to the complexity of the problem, we explore two avenues for its resolution, namely metaheuristics

and blackbox optimization algorithms. Both use the sequential property specific to our multilevel formulation, asserting that when the first-level variables are fixed, the lower levels can be solved sequentially. We propose two local search metaheuristics and compare their performances with an off-the-shelf blackbox solver known to perform best for problems with few variables. However, large-scale transportation networks met in practice carry hundreds of commodities, resulting in hundreds of deviation coefficient variables. To reduce the number of variables while keeping high-quality solutions, we propose heuristic approaches creating clusters of commodities having equal deviation coefficients. The clusters are formed by either exploiting the structure of the network or analyzing the distribution of demand over the planning horizon. The proposed methodology outperforms previous findings and yields decreases in costs on the Canadian National Railway Company application. Moreover, the clustering heuristics allow to reach the best tactical costs and leverage the use of off-the-shelf blackbox solvers even for large-scale applications.

Finally, we propose a counterfactual prediction approach for the impact assessment problem of an improvement to a decision-making system. The objective is to assess the impact on a key performance indicator of the carrier, such as the revenue or the income. We focus on the Revenue Management System of an airline, and the impact on the revenue. The approach is independent of the improvement itself and only needs the revenue data. Since the revenue that would have been without improving the system is not observable, we propose to estimate it with counterfactual prediction models. We can then estimate the impact as the difference between the observed impacted revenue and estimated non-impacted revenue. Computational experiments with data from Air Canada show that both linear and deep-learning models are highly accurate for aggregated counterfactual revenue predictions, which allows to estimate small impacts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

BBO	Blackbox Optimization
BP	Block Planning
CN	Canadian National Railway Company
DID	Differences-In-Differences
FC	Fully Connected
FFNN	Feed Forward Neural Network
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MCND	Multicommodity Capacitated Fixed-charge Network Design
MADS	Mesh Adaptive Direct Search
MCNNM	Matrix Completion with Nuclear Norm Minimization
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NN	Neural Network
NS	Neighborhood Search
NSDI	Neighborhood Search with Diversification and Intensification
OD	Origin-Destination
PDE	Periodic Demand Estimation
PoC	Proof of Concept
RMS	Revenue Management System
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RSC	Robust Synthetic control
SC	Synthetic control
SND	Service Network Design
tAPE	total Absolute Percentage Error
WAPE	Weighted Absolute Percentage Error

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

In the increasingly globalized world, the transportation sector has become essential for both international trade and tourism, two major components of countries' economy. In 2018, the value of world merchandise exports was 19.5 trillion US dollars in 2018[1]. At the same time, the total international tourism exports generated 1.7 trillion US dollars[2]. Efficient transportation services between continent and countries, and within countries are not only a major source of economic growth but also a necessity. In Canada, the transportation sector represented 4.5% of GDP in 2018[3]. Its trading partners are spread out across the world: besides to the U.S., Canada's top five trading partners includes China, Mexico, Japan and the United Kingdom.

The transportation sector allows people and goods to move worldwide. In the case of air transportation, both might share the same resources: airplanes with passengers sometimes also carry freight. For rail transportation, this is not the case in North America, and companies, also called *carriers*, share the same infrastructure but are dedicated to either freight or passenger transportation. This is because the equipment and the operations required to handle the movements of goods or people are deeply different.

Despite its importance, the transportation sector suffers from negative impacts. It is indeed the second CO_2 emitter, after the energy sector (Ritchie and Roser, 2020). Since 2005, the global greenhouse gas emissions from the transportation sector have increased in Canada despite the improvement of the energetic efficiency of air, rail and road carriers. Moreover, the transportation sector is affected by inefficiencies from either the under-utilization of capacity and congestion, that are due in part to the demand uncertainties. They are costly and carriers aim at planning their operations to avoid them as much as possible.

Transportation networks hence need to be wide, environmentally friendly, cost-efficient and flexible to support the global and domestic trade. With the development of mathematical modeling of operations and optimization algorithms, in parallel with the collection of large amount of data, both freight and passenger carriers have introduced highly sophisticated systems in their decision-making process.

The operations of a carrier and its whole planning process are remarkably complex. The multiple planning decisions, taken for a large-size network, often countrywide, are highly

[1] Annual report from the World Trade Organization
[2] Annual report from the World Tourism Organization
[3] Annual report from Transport Canada

dependent. Due to the size and the combinatorial aspect of the problems, having one model encompassing all operations would be impossible to solve. Therefore, carriers are typically decomposing their decision-making process into three levels, for three different planning horizons. Each level imposes constraints on the subsequent decision levels. At the strategic level, long-term decisions are taken, for instance the modification of the physical network or the acquisition of major resources such as locomotives for rail companies. The tactical level concerns medium-term decisions, namely the schedule and the allocation of resources to reach the strategic objectives. Finally, the operational level focuses on short-term decisions: the optimization of resources set out at higher levels to satisfy the actual demand.

Figure 1.1 illustrates the decision-making process of a carrier, where decisions flow from strategic to tactical and finally to operational while observed data, namely the transported demand and used resources, are used in the opposite order. We first observe the day-to-day transported demand, then aggregate the observations over longer periods. The three decision levels commonly rely at their respective time horizon on demand forecasts containing uncertainty. Strategic decisions usually require quarterly or yearly demand forecasts, while tactical decisions look at weekly or monthly forecasts and operational decisions at daily, or hourly forecasts. At the operational level, the transported demand results from the real demand, the available capacity and uncertain factors and constitutes the historical data used in the different forecasting modules.

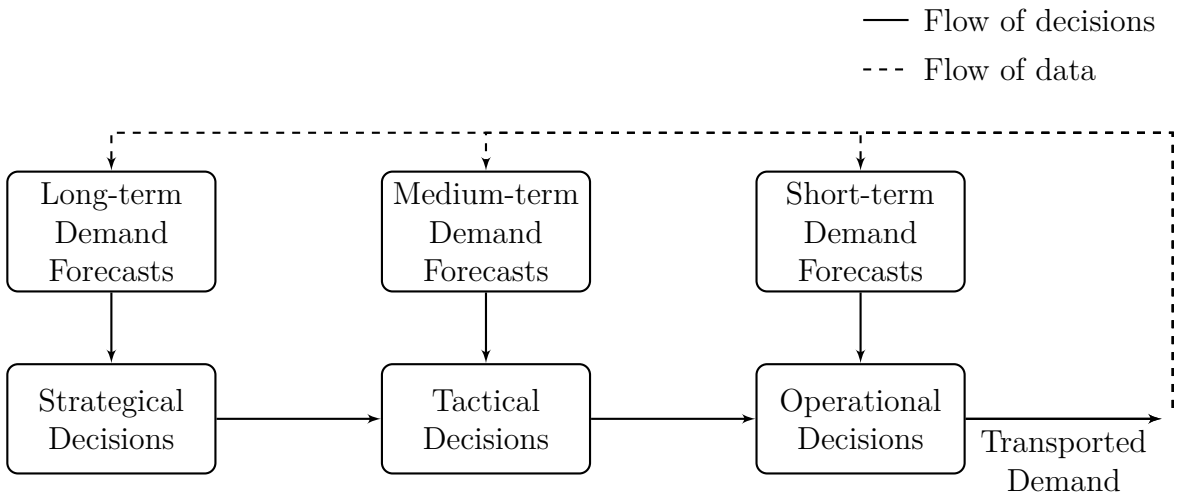


Figure 1.1 Decision making process of a carrier

Decisions at each level are the solutions to optimization problems with specific objectives, often included in analytics and decision-making systems. Tactical planning of freight carriers for instance aims at minimizing the costs of using the resources while satisfying demand and

is often modeled as a mathematical program solved with commercial softwares. Airlines make use of Revenue Management Systems for tactical and operational decisions. Because of the complex nature of both the forecasting and optimization problems, the prevailing practice is to first predict, then optimize separately: the forecasting models would then not consider the downstream optimization problem.

In this thesis, we focus on the link between demand and supply for transportation networks. We carry out interdisciplinary studies by combining methodologies from machine learning, statistics and optimization. We introduce new perspectives and concentrate on the integration of demand forecasting and planning, and the analysis of the impact on carriers' key performance indicators. We summarize the contributions in Figure 1.2. First, we link demand forecasts to the objective of the decision-making problems they should contribute to solve, and represent it with the red arrow. We also assess the impact of the demand forecasts on the planning costs. Then, we identify the problem of assessing the impact of improvements to decision-making systems and designate it by the blue arrows. This is a crucial problem for carriers for investment decisions: the value of the impact might lead to fundamental changes of their system. Even though each one of the forecasting, planning and impact assessment problems has been well studied in the literature in specific cases, we identify problems related to their respective integration to one another that have been overlooked in the literature while having high value in practice. Our work aims at addressing this gap.

The problems considered are strongly linked and affect one another: the value of the key performance indicator, e.g. the revenue, results from the planning decisions that result from the demand forecasts. Yet they draw from distinct literature which have their proper concepts and challenges. This dissertation, composed of self-contained chapters, proposes new formulations and experimental frameworks to tackle them. The following two sections are dedicated for each problem: integrating forecasting and planning, and assessing the impact. We introduce and explain the concepts essential to our research, and discuss our contributions and objectives.

1.1 Integrating Demand Forecasting and Tactical Planning

In this section, we focus on the tactical decision level for freight carriers. Their network is divided into terminals that, in turn, are linked by services, i.e. transportation modes such as trains or trucks. Goods transported on the network are referred to as *commodities*, defined by an origin terminal, a destination terminal and a type of good. The demand is then the quantity of each commodity that moves on the services available on the network. The tactical planning problems aim at answering questions regarding the schedule of the services

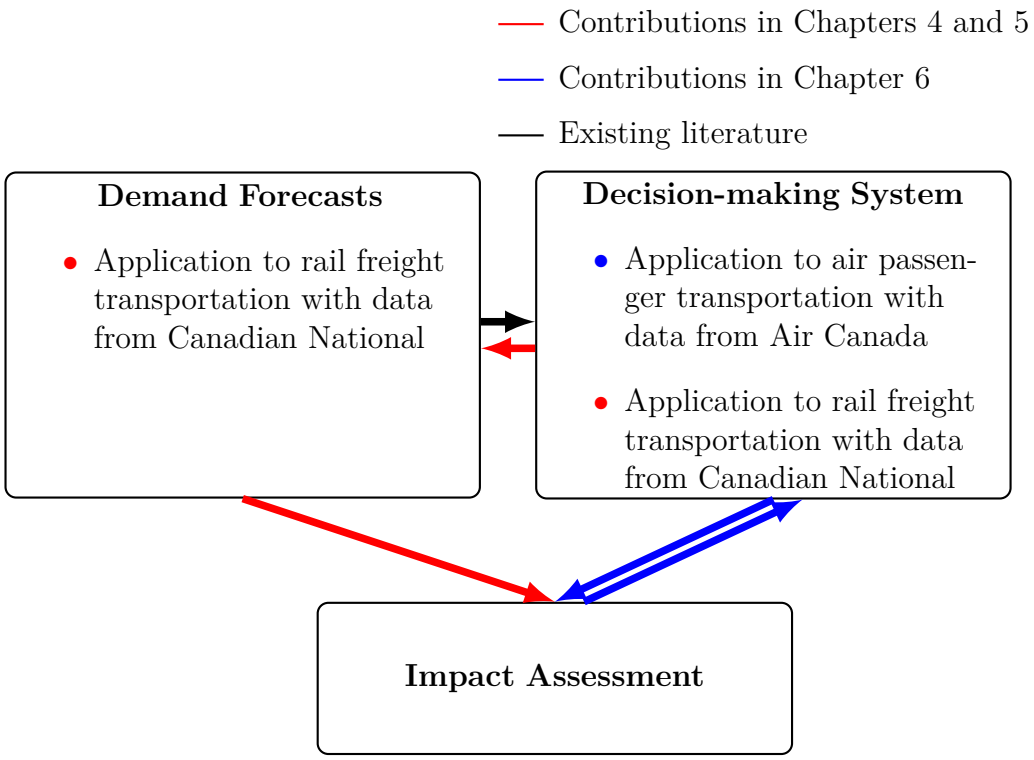


Figure 1.2 Overview of the research

to offer and their frequency, the type of vehicles to use and the routes of the commodities on the services. They are generally described as Service Network Design (SND) problems, a class of problems whose objective is to minimize the costs of routing the freight while satisfying the forecasted demand (Wieberneit, 2008). Those problems are typically modeled by Mixed Integer Programs that include time and space considerations. They integrate the costs and the constraints for all commodities transported in the network: which equipment they can be carried on, the capacity constraints, the possible routes, etc. The solution of the program gives a detailed plan of services at each period of the tactical horizon, their schedule, intermediate stops and the routes on the services for each commodity. In practice, once the design is set, the routes are adjusted operationally to support the actual demand realizations.

To increase the utilization of the different equipment and offer reliable customer services, large-scale transportation networks often use cyclic services: the tactical horizon is decomposed into periods of equal length and the same services are offered at each period. Moreover, due to the large number of commodities in real-life problems, SND problems are often modeled as deterministic, in which case the demand is assumed fixed and known.

The cyclic and deterministic SND formulations require one single estimate of the demand to come for each commodity that we call in this thesis the *periodic demand*. The latter has an important impact on the resulting tactical plan and the associated costs. However, time series forecasting models produce the demand forecasts for each period of the tactical horizon, hence the inconsistency between the available (multiple point estimates of demand) and the required (periodic demand) information.

Despite its importance, the question of estimating a good periodic demand has been overlooked in the literature, and the common practice consists in taking the average of the forecasts over the periods of the planning horizon. There is no study focusing on the Periodic Demand Estimation (PDE) problem linking time series forecasts to the tactical planning problem of interest. We aim to address this gap in this research, in Chapter 4 and Chapter 5. We now highlight their specific objectives as well as their achieved contributions.

Periodic Freight Demand Forecasting for Large-scale Tactical Planning Our first contribution consists of the formal introduction of the *periodic demand estimation problem*. We propose a two-step methodology, where the first step focuses on the time series demand forecasting problem, and the second step aims at estimating the best periodic demand from the forecasts. To do so, we propose a multilevel mathematical programming formulation whose solution is a periodic demand, defined as a mapping of the forecasts that minimized

the fixed design costs and the variable costs incurred from adapting the plan at each period. The proposed formulation links the demand forecasting to the tactical planning. We show the importance of the choice of the periodic demand estimate and the impact on the tactical costs for a large-scale application at the Canadian National Railway Company (CN). We start by forecasting the demand for the intermodal traffic. We develop and compare different forecasting models based on the statistics and deep learning literature on the first large-scale freight transportation network. Results show that statistical models perform best compared to neural networks on limited data. The latter reveal the importance of considering external features such as weather data to improve the forecast accuracy. Then, we solve the multilevel problem by enumerating the solutions on a restricted feasible set for the periodic demand. We show that this methodology lead to a substantial reduction of the tactical costs compared to the common practice using the average of the forecasts as periodic demand.

A Two-step Heuristic for the Periodic Demand Estimation Problem Our second contribution consists in extending the work from our first contribution by allowing a broad and continuous feasible set for the periodic demand and devising a solution approach. We propose a new formulation of the PDE problem, where we formalize the mappings from the demand forecasts to the periodic demand as a deviation from the vector of average forecasts. The first-level variables are then the deviation coefficient for each commodity. Due to the complexity of the problem and its lower levels, we explore two avenues for its resolution, namely metaheuristics and blackbox optimization algorithms. Both use the sequential property specific to our multilevel formulation, asserting that when the first-level variables are fixed, the lower levels can be solved sequentially. We propose two local search metaheuristics and compare their performances with an off-the-shelf blackbox solver known to perform best for problems with few variables. However, the number of variables corresponds to the number of commodities transported in the network, which can be up to hundreds for large-scale applications. To address this challenge, we propose heuristic approaches creating clusters of commodities that have an equal deviation coefficient, hence reducing the size of the feasible set of solutions and the number of variables. The clusters are formed by either exploiting the structure of the network or analyzing the distribution of demand over the planning horizon. We report results for the same large-scale application described in the first contribution. They show that the proposed methodology outperforms previous solutions allowing to further reduce the costs. Implementing the clustering heuristic before solving the problem allows to obtain the best solution. Moreover, it enables the use of the blackbox solver even for large-scale applications.

1.2 Impact Assessment

In this section, we switch focus to passenger transportation and consider Revenue Management Systems (RMSs) used by airlines to maximize revenue. Such decision-making systems handle demand bookings, cancellations and no-shows, as well as the optimization of seat allocations and overbooking levels. Improvements to existing systems are made by the airlines and solution providers in an iterative fashion, aligned with the advancement of the state-of-the-art where studies typically focus on one or a few components at a time. An improvement might be for instance a change of forecasting model, or a change in pricing policies. As RMSs are extremely intricate, the improvement affects the subsequent decisions. Thus, when it is implemented in practice, there are numerous intermediate steps between the improvement and the consequence on observed demand, and therefore on the revenue. It is then challenging to evaluate if the improvement to the RMS lead a significant improvement in revenue. This is even more challenging, as the value of interest, i.e. the difference between the revenue generated after the improvement and the revenue that would have been without it, is not observable.

While there is a wealth of studies aiming to improve RMSs, the literature focusing on assessing quantitatively the impact of such improvements is scarce. We aim to address this gap in Chapter 6, and highlight next its specific objective and achieved contributions.

Assessing the Impact: Does an Improvement to a Revenue Management System Lead to an Improved Revenue? In our third contribution, we propose a methodology for impact assessment that is independent of the improvement itself and only needs the data of the outcome of interest, that is the revenue in our case. We rely on counterfactual prediction models which estimate the reference revenue, that is the revenue that would have been if the RMS had not been changed. The impact is then estimated as the difference between the estimated reference revenue and the observed impacted revenue. We provide a comprehensive overview of counterfactual prediction models and the first extensive computational study that use those model in a large-scale airline application. It stands out from the usual macroeconomic synthetic control applications. We compare models from the literature with a tailored deep learning model for this task. We use data from Air Canada and present accurate counterfactual prediction models that allow to estimate fairly small impacts.

1.3 Thesis Outline

The remainder of this document is organized as follows. Chapter 2 contains a brief literature review of the main topics discussed in this thesis. Chapter 3 presents a synthesis of the work as a whole and the general organization of the manuscript. Chapters 4, 5 and 6 form the main body of this thesis, and contain the three aforementioned contributions. In Chapter 7, we further discuss our contributions regarding the three objectives of this thesis. Finally, in Chapter 8 we summarize the dissertation work, comment on its limitations, and identify future research directions.

CHAPTER 2 CRITICAL LITERATURE REVIEW

In this chapter, we present some important definitions and concepts related to our work. The forecasting, planning and impact assessment problems draw from distinct literature that include statistical models, machine learning, mathematical programming, metaheuristics and econometric models. We present in this chapter the key notions for each field and how they relate to our specific problems and applications. We describe in Figure 2.1 the specific topics for each problem considered in our research.

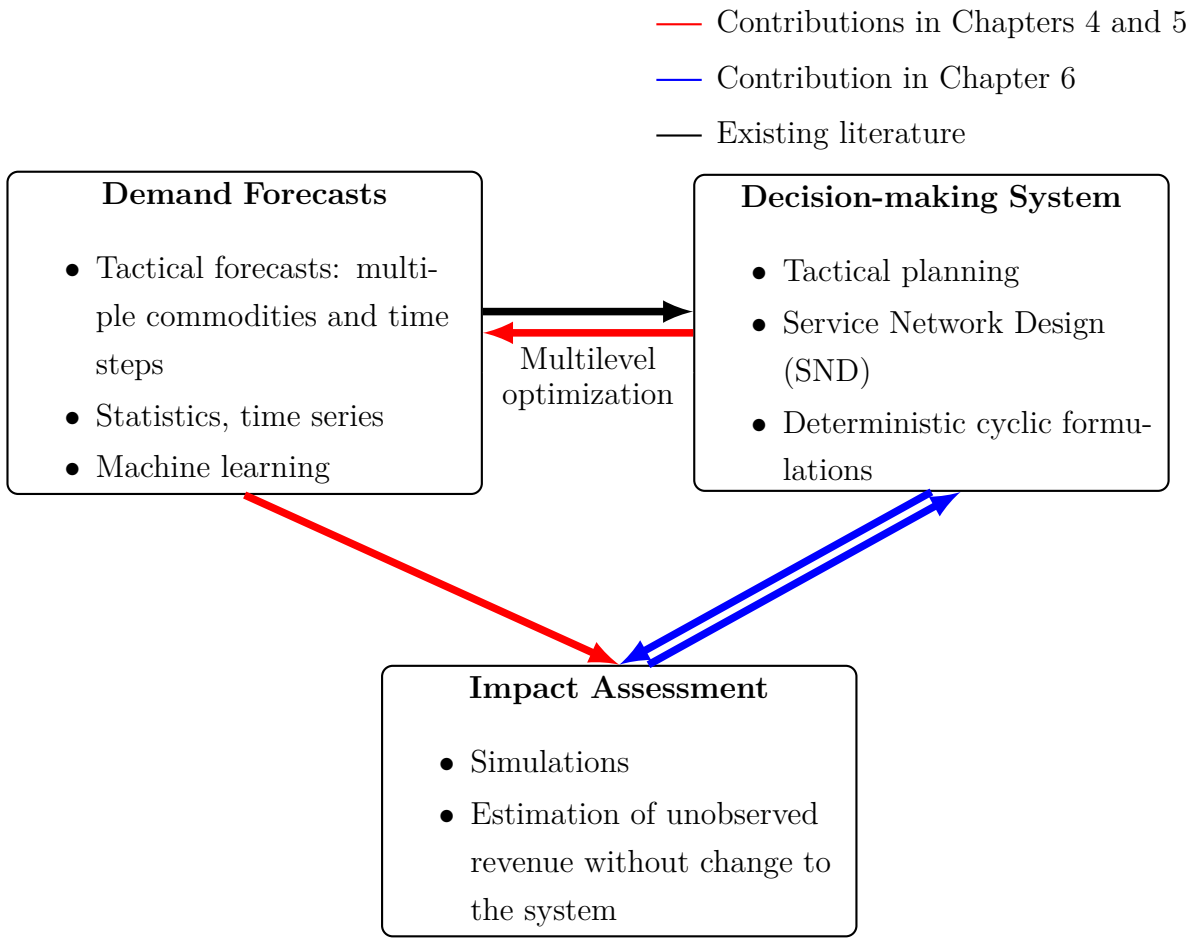


Figure 2.1 Topics related to our research

We first focus on the forecasting and planning problems for tactical decisions with emphasis on a freight transportation application. We review in Section 2.1 the literature related to forecasting, namely statistics and machine learning. In Section 2.2, we focus on the tactical

planning problems for freight transportation. They are modeled as SND problems, which consist in selecting and scheduling services, specifying terminal operations and routing of freight. For large-scale applications, the formulations are mainly deterministic and assume that the demand is fixed and known. In Section 2.3, we present the solution approaches, namely metaheuristics and blackbox optimization, that could be used to solve our multilevel formulation of integrated forecasting and planning problems. Its difficulty is explained by the large-scale, non-convex, not differentiable and combinatorial lower-levels problems. Finally, in Section 2.4, we focus on the question of how to assess the impact when improving the decision-making system, and consider an airline application. The more complex the system, the more difficult it is to estimate properly the impact of the change due to the multiple intermediate steps between the change and the consequences on the transported demand. The potential methods can be divided in two categories: simulations or field experiments. The latter consist in applying the change in practice and estimating the revenue that has not been subject to the change. It is not observed and thus called the *counterfactual revenue*.

2.1 Demand Forecasting

Demand forecasting for tactical planning consists in estimating the demand to come for each commodity traveling on the network at each period of the tactical horizon. This is known as a multivariate multistep demand forecasting problem: *multivariate* designates the multiple commodities and *multistep* designates the multiple periods.

Two types of forecasting methods emerge from the literature: statistics with mostly time series analysis, and machine learning (ML) with the recent development of neural networks (NN). As Breiman (2001) discusses, the former considers that data are generated from an underlying stochastic process and examines historical data to extract it and to predict future trends. Machine learning methods, on the other hand, assume that data have been generated by an unknown process. Karlaftis and Vlahogianni (2011) review the differences between statistical models and NN for forecasting problems in transportation research, and highlight that the former are often defined in terms of the mathematical model they use and its statistical properties, whereas the latter are defined by their architecture and learning algorithms. In a survey paper, Azadeh et al. (2014) review and propose a taxonomy of existing statistical and machine learning demand forecasting methods for revenue management.

One limitation of statistical models is multicollinearity, i.e., the correlation between two or more independent variables. In NN models, the assumption of independent variables being uncorrelated is not made. This is important in transportation networks, as demand for multiple commodities are often correlated. Moreover, NN models are flexible enough to

model complex non-linear relationships in an automated fashion. A multilayer feed-forward neural network for instance is able to approximate, as accurately as desired, a function from training examples (Hornik et al., 1989). The NN also allow to take custom lagged demand into account, to favor recent demand over earlier demand for instance, and external factors, by using them as inputs. This is useful in transportation applications, as the weather for instance might be responsible for sudden changes in the demand, that are difficult to predict with statistical models.

The comprehensive books, Makridakis et al. (2018) and Box et al. (2015), provide a complete description about statistical forecasting models. They include for instance autoregressive models, moving averages and exponential smoothing. An overview of ML models can be found in James et al. (2013), and of NN models in Goodfellow et al. (2016).

Both types of models have been used for demand forecasting in various applications. They each have their advantages and limitations on particular aspects. In terms of implementation, interpretation and data requirements, statistical models tend to outperform NN which have difficulties when data is limited, due to the problem of overfitting (Goodfellow et al., 2016). However, flexibility and external factors inclusion are their strength. Yet in terms of demand forecast accuracy, there is not a clear winner: some studies suggest that NN are more accurate, while others provide evidence for the opposite (Karlaftis and Vlahogianni, 2011).

2.2 Tactical Planning and Service Network Design

Freight transportation networks are composed of terminals, which are linked by infrastructure and services to provide high-quality and reliable customer services. A terminal designates e.g. a train station, an airport or a depot, and a service corresponds to a transportation mode between two terminals. Depending on the application, the service can either be flights (Yang, 2009), trains (Morganti et al., 2020), ships (Agarwal and Ergun, 2008), etc. Commodities moving on the network are defined by their origin and destination terminals, and the type of freight.

Crainic (2000) describes the tactical planning of a freight carrier as the design of the service network, with the objective of finding the optimal allocation and utilization of resources, to achieve the economic and customer service goals. For an efficient allocation of their resources, carriers consolidate the freight, and commodities that do not have the same origin and destination might be moved on the same services. Thus, detours and transshipments of commodities are possible, and are defined when solving the tactical planning problems. Their objective is to define the schedule of the services and their frequency, the type of vehicles

to use, the routes on the service for the different commodities to satisfy the demand with minimal costs. The latter are both the fixed costs of the selected services, and the variable costs of using the selected services.

Tactical planning problems are modeled by the class of SND problems, that are typically Mixed Integer Programs that include constraints specific to the problem addressed. The possible services are described using a time-space graph, where the vertices are the terminals, i.e., the origin and destination of the commodities transported in the network. Arcs joining the terminals represent the services and their cost, duration and capacity. The objectives of those models is to select the services to transport the demand at minimal cost. They integrate the constraints of the physical network, for instance the capacity in each terminal, the equipment required for each commodity, etc. The paper from Crainic (2000) and the one from Wieberneit (2008) propose a review of tactical planning and SND problems in freight transportation.

SND problems are either modeled as deterministic or stochastic problems. In the deterministic case, the demand is assumed fixed and known, and provided as an input to the model. Yet freight transportation networks are subject to various uncertainties: demand, travel time, equipment maintenance, etc. Lium et al. (2009) show that ignoring stochastic factors could result in poor quality of service and high operational costs. To tackle uncertain demand, researchers have been focusing on developing stochastic formulations. However, they have high computational costs for real large-scale instances (Bai et al., 2014), due to the large number of commodities, and the constraints at each terminal to take into account. Hence, deterministic formulations prevail in practice and most studies considering large-scale real-life instances have been focusing on deterministic approaches.

There are many variants of the SND problems, depending on the type of application one wishes to model. We detail below one variant relevant for our research: the Multicommodity Capacitated Fixed-charge Network Design (MCND) problem (Magnanti and Wong, 1984). It is a cyclic and deterministic formulation, where the network has a limited capacity and the objective is to design a tactical plan that allows to transport the demand for a set of commodities $\mathcal{K}$ at a minimum cost. The demand for a commodity $k \in \mathcal{K}$ is designated by y_k^{p} and is in fact the periodic demand, assumed to be repeated at each period of the tactical horizon. We present the path-based formulation **MCND**, where the set of potential paths $\mathcal{P}$ for the commodities is defined in advance. A path corresponds the route formed by one or multiple services that a commodity can take to go from its origin to its destination. The set $\mathcal{P}_k$ is the set of paths that can transport commodity k and $\mathcal{K}_p$ is the set of commodities that can be transported on path p. **MCND** has two categories of decision variables: Binary

design variables $z_p,\ \forall p \in \mathcal{P}$, equal to one if path p is used and zero otherwise, and flow variables $x_{pk} \geq 0$, representing the flow of each commodity $k \in \mathcal{K}$ on its possible paths $p \in \mathcal{P}_k$.

$$\textbf{MCND} \quad \min_{z,x} \sum_{p\in\mathcal{P}} C_p^{\text{design}} z_p + \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k} C_p^{\text{flow}} x_{pk} \tag{2.1}$$

$$\text{s.t.} \quad \sum_{p\in\mathcal{P}_k} x_{pk} = y_k^{\text{p}}, \qquad k \in \mathcal{K}, \tag{2.2}$$

$$\sum_{k\in\mathcal{K}_p} x_{pk} \leq u_p z_p, \qquad p \in \mathcal{P}, \tag{2.3}$$

$$x_{pk} \geq 0, \qquad k \in \mathcal{K}, p \in \mathcal{P}_k, \tag{2.4}$$

$$z_p \in \{0,1\}, \qquad p \in \mathcal{P}. \tag{2.5}$$

The objective function (2.1) includes a fixed design cost $C_p^{\text{design}} \geq 0$ for each path p built to transport the demand. The second cost is the variable flow cost $C_p^{\text{flow}} \geq 0$ which accounts for the cost of demand transported on each path. Constraints (2.2) ensure that the periodic demand is satisfied for each commodity. Constraints (2.3) enforce flows on selected paths only, and that the flow on a path p does not exceed the path capacity, u_p. **MCND** is solved to obtain a tactical plan, i.e, the solutions z^*, x^* based on a given periodic demand. In practice, demand varies from one period to another, and the plan is therefore adjusted at the operational planning level.

MCND is a generic program, and additional constraints might be considered for specific applications, such as the one in our research, i.e., intermodal transportation. Indeed, the latter corresponds to the movement of containers (Crainic and Kim, 2007) for which special services have been developed in the intermodal subdivisions of North American carriers. Long and double-stacked trains that move across country are one example. In **MCND**, this transposes by integer flow variables (the demand is formalized in numbers of containers), and specific capacity constraints to account for the double-staking of containers. Morganti et al. (2020) detail the tactical problem specific to rail intermodal transportation, called *the block planning problem*. Blocking aims to take advantage of economies of scale and reduce the cost of handling cars individually at intermediate yards. Cars are grouped within blocks to be moved together as a unique entity from the origin of the block to its destination. A block is moved by a sequence of trains, while a car can be moved by one or a sequence of blocks between its origin and destination. Finally, containers are stacked on moving cars.

2.3 Metaheuristics and Blackbox Optimization

Let us consider a generic multilevel optimization problem with no derivative information:

$$\min_{s \in \mathcal{S}} f(s, g^*(s)), \tag{2.6}$$

where s is the first-level decision variables, i.e, the periodic demand and S its feasible set, f the objective function and g^* the optimal solution of the lower level programs. We note that this generic formulation bears resemblance to a bilevel optimization program, except for the extreme max-min bilevel case (Colson et al., 2005). However, here $g^*(s)$ is a value and not a set of solutions, and (2.6) does not belong to the class of bilevel programs. We focus on cases where the program $g(s)$ for a fixed s is combinatorial, large-scale and with no derivative information.

The optimal solution to (2.6) could be found by enumerating the feasible solutions $s \in \mathcal{S}$ and storing the best one. This is computationally very expensive and cannot be used in practice, where it is necessary to find a good solution in reasonable times. *Heuristics* and *metaheuristics* have been developed to address this issue, as well as algorithms for black-box optimization problems.

Metaheuristics

Heuristics and metaheuristics have the same objective: to explore the set of feasible solutions and find good solutions. Often, there are no guarantees on the optimality of the solution. The foundational concepts of metaheuristics can be described in an abstract way, without referring to a specific problem. They sometimes include heuristics developed specifically for the problem being addressed, but nonetheless driven by a higher-level strategy (Hertz, 2016).

There are essentially two types of metaheuristics: the local search methods which iterate over one solution and the population methods working with a population of solutions (Talbi, 2009). The former are adapted to our context due to the difficulty of solving large-scale planning problems.

Local search methods improve an initial solution iteratively by building a trajectory in the set of feasible solutions. At each iteration, they improve the current solution by exploring its neighborhood. The latter corresponds to the set of solutions that are easy to reach from s and is defined by:

$$N : s \in S \mapsto N(s). \tag{2.7}$$

Metaheuristics and local search methods are important and mature research topics. Many different algorithms have been proposed in the literature: the Simulated Annealing (Kirkpatrick et al., 1983; Černỳ, 1985), the Variable Neighborhood Search (Mladenović and Hansen, 1997), the Greedy Randomized Adaptive Search Procedure (Feo and Resende, 1989, 1995), to name a few. They differ by the structure of the neighborhood N visited at each iteration, the use of memory during the search and the diversification and intensification procedures computed at each iteration to explore the set of solutions.

Black-box optimization

For black-box optimization (BBO) problems, either the analytic form of f or the one of the functions defining the feasible set S is not known and has no derivative information. It can be for instance the result of a computer code, referred to as the black box. BBO is used in various applications, such as aircraft takeoff trajectories (Torres et al., 2011) or snow water equivalent estimation (Alarie et al., 2013).

The book Audet and Hare (2017) provides a grasp of the foundational concepts in derivative-free and black-box optimization. We take a particular interest in NOMAD, an open source implementation of the Mesh Adaptive Direct Search (MADS), a recent method to solve BBO problems (Audet and Dennis Jr, 2006).

Le Digabel (2011) describes the solver's functionalities, its implementation and the underlying algorithm. MADS is an iterative method that belongs to the more general class of direct search methods, using only evaluations of the black-box functions to drive the exploration of the set of feasible solutions. The *mesh* refers to the spatial discretization of the latter and is defined by a set of directions and a mesh size parameter. Each iteration of MADS is decomposed into three steps: the poll, the search and the update. The poll and search steps generate trial points on the mesh. At the search step, the points lie anywhere on the mesh while the poll step generates points constructed from directions near the current solution. The poll directions grow dense, and the distance to the current solution is bounded by a poll size parameter. At the update step, the algorithm determines if the iteration is a success or not, according to the evaluations of the black-box functions on the trial points. The new solution is either the most promising evaluation or the current solution. Both the mesh size and the poll size parameters are updated such that the former is always smaller than the latter. When the current solution is not updated, the mesh size parameter is reduced.

The current implementation of NOMAD performs best for black-box problems with fewer than 50 variables.

2.4 Assessing the Impact

Potential methods to estimate the impact of improving a decision-making system can be divided in two categories: simulations or statistical field experiments. As a natural follow-up from the previous work, we assume here that the change is a modification in the demand forecasts, and illustrate the notions with this example. We consider applications to airlines whose decision-making systems are RMS that define seat allocation and pricing rules from the demand forecasts.

In the first category, the carrier uses the new forecasted demand and simulates every intermediate step of its planning systems. Assumptions to simplify the problem are usually made, and the results importantly rely on them. Simulations for airlines hence require to be able to reproduce the behavior of each customer facing new prices. This requires a large amount of data and is computationally expensive. The few works focusing on this topic consider simplified simulation settings (Fiig et al., 2019; Weatherford and Belobaba, 2002).

The second category, statistical field experiments, consists in implementing the changes in practice, often on a short period of time and on part of the network, and estimating the impact on a key performance indicator afterwards. The challenge lies in finding the reference to compare to since the observed indicator includes the impact. There is a body of the econometric literature focusing on this problem for macroeconomic applications. A typical example is the estimation of the economic impact of the German reunification on West Germany (Abadie et al., 2015). There are, however, few applications in revenue management and even less focus on the airline industry. Cohen et al. (2019) estimates the impact of differentiating lead-in fares, the fare of the lowest inventory class, on the revenue, yield (ratio of revenue to the number of tickets sold) and market share of the partner airline.

In the remainder, we describe key concepts for statistical field experiments (Abadie, 2021), and illustrate with the impact on the revenue. Data are decomposed in space, the *units*, and time. The improvement on the decision-making system is called a *treatment*. Some units are exposed to the treatment during a subset of the periods. The revenue that would have been without the treatment is not observed, and called the *counterfactual revenue*. The objective is to estimate the counterfactual revenue of the treated units. Then the impact is estimated as the difference between the observed revenue and the estimated counterfactual revenue. The set of units is partitioned into the set of treated units and the set of control units which neither receive, nor are affected by, the treatment.

Techniques to estimate the counterfactual revenue have known an increasing popularity after the development of synthetic control models (Abadie and Gardeazabal, 2003). The latter im-

putes the missing revenue for treated units using a weighted average of the revenue of control units. Synthetic controls have been described as "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens, 2017). Many works have been working on extending this method, defining the untreated revenue of the treated units as a mapping from the revenue of the control units. The paper of Doudchenko and Imbens (2016) and the one of Athey et al. (2021) present a review of the recent developments of the counterfactual prediction models.

While those methods have been developed mainly for social sciences, they apply very well to transportation networks. In this context, units correspond to Origin-Destination pairs (ODs) or commodities, and carriers, in particular airlines, can apply a treatment on a few ODs.

CHAPTER 3 SYNTHESIS OF THE WORK

This thesis encompasses both the integration of demand forecasting and planning, and the impact assessment. While those subjects are large, we focus on two main applications of large-scale transportation networks: the rail intermodal division of the Canadian National Railway Company (CN), one of the largest freight carriers in North America, and the passenger activities of Air Canada, the largest Canadian airline with a worldwide network.

We introduced our research and framework in Chapter 1. We examine the tactical planning problem of freight carriers, modeled as a Service Network Design problem. The formulations for large-scale networks are often cyclic and deterministic due to their complexity, with a fixed and known demand input. Demand forecasting models, on the other hand, yield point estimates of the demand, that vary during the planning horizon. This contradiction of available and required demand inputs for planning optimization models is the base of our focus. We aim at integrating the tactical forecasting and planning problems. Then, we focus on assessing the impact of improving the decision-making system, not only for freight carriers but for network-based transportation applications.

In Chapter 2, we reviewed the literature related to our work and objectives, which include several subjects: statistical and machine learning forecasting models, service network design problems, metaheuristics and counterfactual prediction models.

Periodic Freight Demand Forecasting for Large-scale Tactical Planning In Chapter 4, we introduce the Periodic Demand Estimation (PDE) problem which aims at integrating demand forecasting and planning. We propose a two-step methodology for large-scale cyclic and deterministic formulations of service network design problems. The first step consists in obtaining time series forecasts of the demand for each commodity transported at each time period of the tactical horizon. The second step defines the periodic demand as a solution to a multilevel mathematical program that explicitly connects the estimation problem to the tactical planning problem, and minimizes the costs incurred by adapting the tactical plan at an operational level. We consider in this chapter a limited feasible set of periodic demand and solve the problem by enumerating the solutions. We report results on CN's intermodal network and compare the periodic demand estimate resulting from the proposed methodology to the approach commonly used in practice which simply consists in using the mean of the time series forecasts. Even with the restrictions, results show the importance of the periodic demand estimate. Moreover, despite the uncertainty contained in demand

forecasts, a good periodic demand estimate can lead to substantial cost reductions.

A Two-step Heuristic for the Periodic Demand Estimation Problem In Chapter 5, we extend the work from Chapter 4 and no longer restrict the periodic demand to be taken from a small discrete set. It is rather defined as a deviation from the average of the demand forecasts. We present a new multilevel formulation for the PDE problem, where the variables are the deviation coefficients. We first propose two local search metaheuristics to solve the new formulation and compare them with NOMAD, an off-the-shelf black-box optimization solver. Both might be challenged by the number of variables, which can be up to hundreds for the large-scale applications met in practice. We then propose clustering heuristic approaches which aim at reducing the size of the feasible set of the PDE problem while keeping high-quality solutions. It consists in creating clusters of commodities that have an equal deviation coefficient, hence reducing the number of variables. The results on CN's intermodal network show that considering only the second step of the heuristic, i.e. defining the periodic demand as a deviation from the average of the forecasts lead to substantial cost reductions. The clustering heuristics allow to leverage the solution algorithms even for large-scale applications, as best tactical costs are obtained when they are computed before solving the problem.

Assessing the Impact: Does an Improvement to a Revenue Management System Lead to an Improved Revenue? In Chapter 6, we focus on assessing the impact of a change in a sophisticated decision-making system. This is a challenging problem as the impact corresponds to the difference between the generated value and the value that would have been generated keeping the system as usual, and the latter is not observable. We consider the specific case of Revenue Management Systems used for the traffic of passengers by airlines, and the impact on the revenue. We propose to model the problem as a counterfactual prediction problem which objective is to estimate the unobserved revenue. The revenue impact is therefore the difference between the observed revenue subject to the improvement and the estimated revenue. We compare counterfactual prediction models developed for macroeconomic contexts with deep learning models. The counterfactual prediction models achieve between 1% and 1.1% of error allowing to estimate a small impact quite accurately.

In Chapter 7, we provide a general discussion on the three works developed in this thesis. Finally, in Chapter 8, we summarize our work and discuss some of the limitations associated with the proposed methods, and indicate future research avenues.

CHAPTER 4 ARTICLE 1: PERIODIC FREIGHT DEMAND FORECASTING FOR LARGE-SCALE TACTICAL PLANNING

The text of this chapter is the one of the research paper *Periodic Freight Demand Forecasting for Large-scale Tactical Planning* submitted to the journal *Transportation Research Part B: Methodological*.

Authors Greta Laage, Emma Frejinger, Gilles Savard

Abstract Crucial to freight carriers is the tactical planning of the service network. The aim is to obtain a cyclic plan over a given tactical planning horizon that satisfies predicted demand at a minimum cost. A central input to the planning process is the periodic demand, that is, the demand expected to repeat in every period in the planning horizon. We focus on large-scale tactical planning problems that require deterministic models for computational tractability. The problem of estimating periodic demand in this setting broadly present in practice has hitherto been overlooked in the literature. We address this gap by formally introducing the periodic demand estimation problem and propose a two-step methodology: Based on time series forecasts obtained in the first step, we propose, in the second step, to solve a multilevel mathematical programming formulation whose solution is a periodic demand estimate that minimizes fixed costs, and variable costs incurred by adapting the tactical plan at an operational level.

We report results in an extensive empirical study of a real large-scale application from the Canadian National Railway Company. We compare our periodic demand estimates to the approach commonly used in practice which simply consists in using the mean of the time series forecasts. The results clearly show the importance of the periodic demand estimation problem. Indeed, the planning costs exhibit an important variation over different periodic demand estimates and using an estimate different from the mean forecast can lead to substantial cost reductions. For example, the costs associated with the period demand estimates based on forecasts were comparable to, or even better than those obtained using the mean of *actual* demand.

Key words Freight transportation, tactical planning, large-scale, periodic demand, forecasting demand.

4.1 Introduction

Freight transportation is essential to society and its economic development. In order to satisfy demand in a cost effective way, freight carriers are faced with a multitude of planning problems. In this context, Service Network Design (SND) is an important class of problems. Consider, for example, the Multicommodity Capacitated Fixed-charge Network Design (MCND) problem (Magnanti and Wong, 1984). The objective is to design a capacitated network – a tactical plan – that allows to transport demand for a set of commodities between different origin-destination pairs at a minimum cost. The latter is given by the sum of fixed and variable costs. The tactical plan is defined over a given period (e.g., a week) and is repeated over a planning horizon (e.g., a few months). Given this cyclic nature of the tactical plan, it relies on an accurate representation of *periodic demand*.

In any realistic setting, demand for commodities is subject to uncertainty. This has naturally led to stochastic SND formulations (e.g., Crainic et al., 2020). As even deterministic SND problems are NP-hard, stochastic formulations are limited to fairly small size problems and cannot yet be used in most real large-scale applications. Hence, a wealth of practical applications rely on deterministic formulations and point estimates of periodic demand. In turn, periodic demand has an important impact on the resulting tactical plan and the associated costs. Despite its importance, there is no study in the literature focused on the periodic demand estimation problem linking time series forecasts to the tactical planning problem of interest. Our work addresses this gap and we use a MCND formulation for illustration purposes.

The impact of demand forecast errors on revenue has been studied in the context of airline revenue management. Through simulation analysis in a simplified setting, Weatherford and Belobaba (2002) show that reducing demand forecast errors by 25% increase revenue by a minimum of 1-2% which is a significant number. Fiig et al. (2019) confirm those findings in a more complex airline revenue management setting and show that reduced forecast errors lead to increased revenue. As opposed to passenger transportation, freight carriers typically have flexibility regarding the routing of demand as long as it respects certain constraints, such as delivery time. Moreover, the freight demand origin-destination matrices are often unbalanced, meaning that there can be excess supply in certain directions. It implies that the cost associated with demand forecast errors can vary over commodities. This further motivates the importance of linking the periodic demand estimation and the corresponding SND problem.

Our proposed methodology proceeds in two steps: First we forecast demand for a given set

of commodities for *each period* of the planning horizon. This corresponds to a multivariate multistep time series forecasting problem. Then, we define mappings from the time series forecasts to periodic demand. The different mappings lead to different periodic demand estimates. We solve a mathematical program explicitly linking these mappings and the MCND formulation. It selects the periodic demand that minimizes the fixed and variable costs over the planning horizon.

Brief Background on Time Series Forecasting. The periodic demand estimates are based on time series forecasts. There is an extensive related body of literature in statistics and machine learning. Our problem is particularly challenging for a number of reasons. First, the historical data on which the forecasting models rely, are results of operational decisions that are constrained by available capacity. Second, there are a large number of commodities and relatively long forecasting horizon which can lead to spatiotemporal correlations. Third, the demand varies over time and long-term dependencies and seasonality can be specific to each commodity. This highlights the potential need for modeling both commodity specific behavior and correlation between commodities while classic time series models and exponential smoothing methods assume independence across time series.

Freight demand forecasting works mainly focus on small networks of port terminals (Milenković et al., 2019) with either statistical models (Schulze and Prinz, 2009) or neural networks (Tsai and Huang, 2017). With the recent development in intelligent transport systems and availability of large sources of data, forecasting methods have been shifting from model-based statistical models to data-driven machine learning approaches and more specifically deep learning models (Karlaftis and Vlahogianni, 2011). Neural networks challenge the statistical models such as AR processes and Holt Winters method (Holt, 2004; Winters, 1960) with their augmented capacity to model non-linearities (Hornik et al., 1989). The capacity of neural networks to model complex data to forecast traffic flows is increasingly exploited (Nguyen et al., 2018). The Long Short-Term Memory (LSTM) recurrent neural network (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014) is a successful architecture to model both short-term and long-term dependencies (Längkvist et al., 2014). Nevertheless, empirical evidence shows that it is still hard to achieve a level of accuracy comparable to that of classic time series models (e.g., Makridakis et al., 2018).

Contributions. The paper offers both methodological and empirical contributions. First, we formally introduce the periodic demand estimation problem and propose a two-step methodology. Based on time series forecasts obtained in the first step, we propose a multilevel mathematical programming formulation whose solution is a periodic demand estimate that

minimizes fixed and variable costs. The formulation hence explicitly links the periodic demand estimates to the tactical planning problem of interest. It is computationally tractable as it can be solved sequentially to optimality. Second, we describe a real large-scale application at the Canadian National Railway Company (CN). We present an extensive empirical study that clearly shows the importance of the periodic demand estimation problem. In this context, we compare different forecasting models from the statistics and deep learning literature. In turn we analyze the impact of the definition of periodic demand distinguishing between time series forecast errors and the errors introduced by different periodic demand estimates. Moreover, we benchmark against an approach used in practice that consists in averaging time series forecasts.

Paper Organization. The remainder paper is structured as follows. Next we formally introduce the periodic demand estimation problem. In Section 4.3 we describe the proposed two-step methodology. We then focus on empirical results, first introducing our application in Section 4.4, followed by the results in Section 4.5. Finally, Section 4.6 concludes and outlines some directions for future research.

4.2 Problem Description

We start by briefly summarizing the planning process we consider: We take the point of view of a freight carrier that wishes to define a tactical plan. First, the carrier estimates the periodic demand for each commodity over the tactical planning horizon. Second, the periodic demand estimates are used as an input to solve the tactical planning problem of interest. The latter involves design decisions that are fixed over the tactical planning horizon, and flow decisions. At the operational level, the tactical plan is adjusted – i.e., the flow decisions – according to actual demand realizations. In addition to these adjustments, other decisions could be taken to cope with demand fluctuations, such as outsourcing. There are hence two sources of costs to consider in the tactical planning process: the fixed cost of the tactical plan (design decisions) and the variable cost (flow and outsourcing decisions) resulting from the adjustments in each period.

We attend to large-scale problems that require a deterministic formulation to be tractable. Therefore, at the tactical planning level, the demand is treated as fixed and known while, in reality, it varies in each period. In this section we describe in detail the problem of estimating the periodic demand so as to minimize fixed and variable costs. We first introduce notation related to tactical and operational planning time horizons. We then define the various concepts of demand we encounter, followed by a description of an MCND formulation.

Finally, we describe the link between observed demand, periodic demand and the MCND formulation, which formally introduces our problem.

For the demand forecasting problem, it is important to distinguish the tactical and operational planning horizons. We therefore introduce two different notations related to time. First, the tactical planning horizon $\mathcal{T}$ can be divided into periods of equal length $t = 1, \ldots, T$. Second, each period t can be further divided into D time periods, $d = 1, \ldots, D$ that we here refer to as the operational horizon. Figure 4.1 provides an illustration where the tactical horizon is composed of $T = 4$ weeks, and a week t is composed of $D = 7$ days.

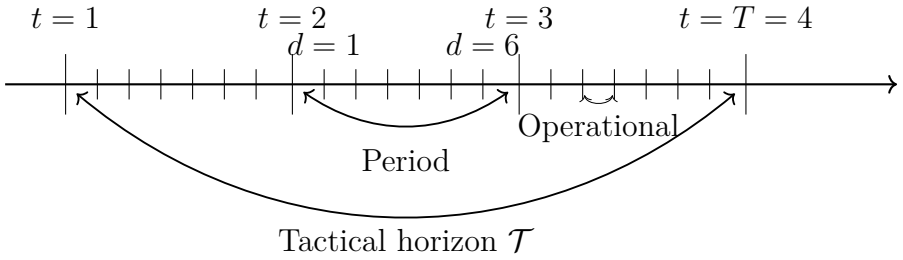


Figure 4.1 Time scales for planning of a freight carrier

Let $\mathbf{y}_t$ be the demand vector of period t, $\mathbf{y}_t = (y_{t1}, \ldots, y_{tK})^\top$ where y_{tk} is the quantity of commodity k to be transported during period t. In this context, a commodity k is characterized by its origin o_k, destination d_k and type γ_k. We denote the set of commodities $\mathcal{K}$ and its cardinality K. Let $\mathbf{y}_d^t$ be the demand vector for each operational time $d = 1, \ldots, D$ within period t, $\mathbf{y}_d^t = (y_{d1}^t, \ldots, y_{dK}^t)^\top$ where y_{dk}^t is the demand for commodity k to be carried at time d in period t. The demand for a period t is hence

$$y_{tk} = \sum_{d=1}^{D} y_{dk}^t, \quad k \in \mathcal{K}. \tag{4.1}$$

We introduce the demand matrix for horizon $\mathcal{T}$, $\mathbf{Y}^\mathcal{T} \in \mathbb{R}_+^{T \times K}$ with $[\mathbf{Y}^\mathcal{T}]_{tk} = y_{tk}$. For a given tactical planning horizon $\mathcal{T}$, the plan is repeated at each $t = 1, \ldots, T$. Let $\mathbf{y}^{\mathrm{p}\mathcal{T}}$ be the periodic demand vector for tactical horizon $\mathcal{T}$, $\mathbf{y}^{\mathrm{p}\mathcal{T}} = (y_1^{\mathrm{p}\mathcal{T}}, \ldots, y_K^{\mathrm{p}\mathcal{T}})^\top$ where $y_k^{\mathrm{p}\mathcal{T}}$ is the periodic demand for commodity k. To simplify the notation, we henceforth remove the superscript $\mathcal{T}$ but recall that the periodic demand and the demand matrix always refer to a given horizon.

We focus on estimating the periodic demand $\mathbf{y}^{\mathrm{p}}$. In this context it is important to note that time series forecasting models produce demand forecasts for each commodity in *each period* t. That is, at period t_0, the forecasting models output an estimate of $\mathbf{Y}$ denoted $\hat{\mathbf{Y}}$ which consists in T point estimates $\hat{\mathbf{y}}_{t_0+1}, \ldots, \hat{\mathbf{y}}_{t_0+T}$. The periodic demands y_k^{p} are then estimated

from these forecasts $\hat{\mathbf{Y}}$, or, for validation or analysis, from $\mathbf{Y}$. Let h denote the mapping of $\mathbf{Y}$ to a periodic demand vector $\mathbf{y}^{\mathrm{p}}$:

$$\begin{aligned} h\colon \quad & \mathbb{R}_+^{T\times K} \to \mathbb{R}_+^K \\ & \mathbf{Y} \mapsto \mathbf{y}^{\mathrm{p}} = h(\mathbf{Y}). \end{aligned} \tag{4.2}$$

When the periodic demand is a mapping of the forecasts, we use the notation $\hat{\mathbf{y}}^{\mathrm{p}} = h(\hat{\mathbf{Y}})$. Our objective is to define the mapping h minimizing fixed and variable planning costs.

We provide an illustrative example in Figure 4.2. The two graphs show two demand distributions (shown with one dot per period) with identical mean (depicted with a solid line). The mean represents one particular mapping $\mathbf{y}^{\mathrm{p}} = h(\mathbf{Y})$. In the example, we illustrate three other possible mappings: the maximum, median and third quartile. As opposed to the mean, these other mappings do not result in the same periodic demand estimates in the right and left-hand graphs.

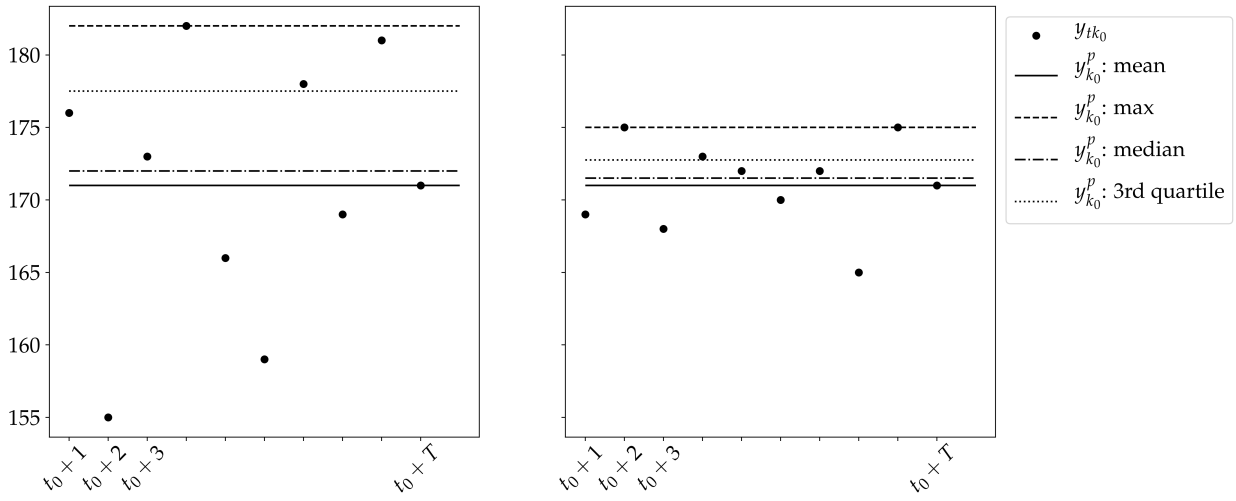


Figure 4.2 Illustration of a periodic demand from point estimates for T periods

We now introduce a path-based MCND formulation (Crainic, 2000) that we use for illustrating our methodology. An arc-based formulation can be found, e.g., in Chouman et al. (2017). Let $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ denote a space-time graph where $\mathcal{N}$ is the set of nodes and $\mathcal{A}$ is the set of arcs. Commodity k uses a path p, i.e., a sequence of arcs in $\mathcal{G}$. The source node of the first arc is o_k, and the sink node of the last arc is d_k. Let $\mathcal{P}$ denote the set of paths. In the case of insufficient capacity, demand is outsourced and we denote $\mathcal{P}^{\mathrm{out}}$ the paths corresponding to outsourcing options, such that $\mathcal{P}^{\mathrm{out}} \subset \mathcal{P}$. Furthermore, let $\mathcal{P}_k$ denote the set of paths for

commodity k, $\mathcal{P}_k^{\text{out}}$ the outsourcing paths for commodity k such that $\mathcal{P}_k^{\text{out}} \subset \mathcal{P}_k$ and $\mathcal{K}_p$ the set of commodities that can use p. Note that here we refer to outsourcing in a broad sense. It could mean outsourcing to a third party, or making use of additional capacity from the same carrier that was not originally part of the plan. For example, in our intermodal rail transporation application (Section 4.4), outsourcing means using capacity from non-intermodal trains.

The MCND problem consists in satisfying demand at minimum cost. It has two categories of decision variables: Binary design variables $z_p,\ \forall p \in \mathcal{P}$, equal to one if path p is used and zero otherwise, and flow variables $x_{pk} \geq 0,\ \forall k \in \mathcal{K}, p \in \mathcal{P}_k$. Depending on the type of freight (bulk versus containers, for instance), x_{pk} is either continuous or integer. The path-based mixed integer linear programming formulation is:

$$\textbf{MCND} \quad \min_{z,x} \sum_{p\in\mathcal{P}} C_p^{\text{design}} z_p + \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k\setminus\mathcal{P}_k^{\text{out}}} C_p^{\text{flow}} x_{pk} + \sum_{k\in\mathcal{K}} \sum_{p\in\mathcal{P}_k^{\text{out}}} C_p^{\text{out}} x_{pk} \tag{4.3}$$

$$\text{s.t.} \sum_{p\in\mathcal{P}_k} x_{pk} = y_k^{\text{p}}, \qquad k \in \mathcal{K}, \tag{4.4}$$

$$\sum_{k\in\mathcal{K}_p} x_{pk} \leq u_p z_p, \qquad p \in \mathcal{P}, \tag{4.5}$$

$$x_{pk} \geq 0, \qquad k \in \mathcal{K}, p \in \mathcal{P}_k, \tag{4.6}$$

$$z_p \in \{0,1\}, \qquad p \in \mathcal{P}. \tag{4.7}$$

The objective function (4.3) includes a fixed design cost $C_p^{\text{design}} \geq 0$ for the paths built to transport demand. The second cost is the variable flow cost $C_p^{\text{flow}} \geq 0$ which accounts for satisfied demand and the third cost $C_p^{\text{out}} \geq 0$ is the flow cost of outsourced demand. Constraints (4.4) ensure that the periodic demand is satisfied for each commodity. Constraints (4.5) enforce flows on selected paths only, and that the flow does not exceed the path capacity, u_p.

MCND is solved to obtain a tactical plan based on a given periodic demand. However, in practice, demand varies from one period to another. The tactical plan is therefore adjusted at the operational planning level. That is, in each period the commodity flows can be adjusted to satisfy the actual demand value of this period also taking into account other uncertain aspects, such as schedule delays. The observed data $\mathbf{y}_d^t$, typically used for training forecasting models, result from this operational planning process. Consequently, data at this level of detail can be constrained by the available services and the observed demand may therefore not correspond to the true demand. This is known as censored data in the literature (e.g., Park et al., 2007).

In summary, we focus on estimating periodic demand $\mathbf{y}^{\text{p}}$ at a given time t_0 for a tactical

planning horizon $\mathcal{T}$. The demand forecasts $\hat{\mathbf{y}}_{t_0+1}, \ldots, \hat{\mathbf{y}}_{t_0+T}$ are obtained using historical data of demand $\{\mathbf{y}_d^s, s = t_0 - 1, t_0 - 2, \ldots, t_0 - H, d = 1, \ldots, D_s\}$ where H is the number of periods in the historical data and D_s is the number of operational time intervals in each period s. The periodic demand should be defined such that it minimizes fixed costs, as well as variable costs associated with adapting the plan over the tactical planning horizon. It is hence necessary to link the mapping h to the tactical planning problem of interest. In the following section, we propose a formulation for this purpose using **MCND** as an example of tactical planning problem formulation.

4.3 Periodic Demand Estimation

Each time a tactical plan is to be defined, our approach proceeds in two steps. First, we use a time series forecasting model to predict demand for each period in the planning horizon. Second, we solve a multilevel formulation for the joint periodic demand estimation and tactical planning problem. In the following subsection we describe assumptions and their implications on the time series forecasting problem. In Section 4.3.2, we delineate the mathematical programming formulation.

4.3.1 Time Series Forecasting

We use time series forecasting methods to predict, at period t, the T point estimates $\hat{\mathbf{y}}_{t+1}, \ldots, \hat{\mathbf{y}}_{t+T}$. As we highlight in the previous section, historical data captures operational flows which can be constrained by the supply and, therefore, may not correspond to actual demand. Time series forecasting with censored data is challenging and difficult to validate. In this section we introduce two weak assumptions that allow us to work with historical *uncensored* data with which we use time series forecasting methods.

Recall from the problem description that demand which cannot be satisfied by the planned capacity is outsourced. This is typically the case for carriers as unsatisfied demand would otherwise accumulate over time periods. Taking into account the outsourcing, demand is hence assumed satisfied in each time period t. This leads us to the following assumption on the historical data.

Assumption 1 *Historical data is uncensored when aggregated over time periods. That is,* $\mathbf{y}_s = \sum_{d=1}^{D_s} \mathbf{y}_d^s, s = t - 1, t - 2, \ldots, t - H$ *are uncensored.*

While this assumption simplifies the forecasting problem we note that the aggregation results in fewer data points to learn from.

The tactical planning problem formulation is based on a space-time graph. The departure and arrival times of a commodity k are hence implicitly given by o_k and d_k. We assume that the arrival and departure times are endogenous decisions, stated in other words in the following assumption.

Assumption 2 *Predicted demand per time period for each commodity,* $\hat{\mathbf{y}}_t,\ t = 1, \ldots, T$*, are sufficiently precise for tactical planning.*

This is a weak assumption considering that, if exogenous predictions of $\hat{\mathbf{y}}_d^t,\ d = 1, \ldots, D_t$ are required for each $t = 1, \ldots, T$, it is possible to define a model (different from the time series one) that projects $\hat{\mathbf{y}}_t$ down to that level.

4.3.2 A Multilevel Formulation

We define the feasible set of periodic demand vectors

$$\mathcal{Y} = \{\hat{\mathbf{y}}^{\mathrm{p}} = h_i(\hat{\mathbf{y}}_1, \ldots \hat{\mathbf{y}}_T), i = 1, \ldots, I\} \tag{4.8}$$

by a finite set of mappings $h_i,\ \ i = 1, \ldots, I$ (4.2). We propose the following multilevel formulation **PDE** for the periodic demand estimation problem.

$$\textbf{PDE} \quad \min_{\hat{\mathbf{y}}^{\mathrm{p}}} \quad C^{\mathrm{PDE}} = \sum_{t=1}^{T} \left[\sum_{p \in \mathcal{P}} C_p^{\mathrm{design}} z_p + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \setminus \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{flow}} x_{tpk} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{out}} x_{tpk} \right] \tag{4.9}$$

$$\text{s.t.} \quad \hat{\mathbf{y}}^{\mathrm{p}} \in \mathcal{Y} \tag{4.10}$$

$$\textbf{MCND} \quad \min_{z,x} \sum_{p \in \mathcal{P}} C_p^{\mathrm{design}} z_p + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \setminus \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{flow}} x_{pk} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{out}} x_{pk} \tag{4.11}$$

$$\text{s.t.} \sum_{p \in \mathcal{P}_k} x_{pk} = \hat{y}_k^{\mathrm{p}}, \qquad k \in \mathcal{K}, \tag{4.12}$$

$$\sum_{k \in \mathcal{K}_p} x_{pk} \leq u_p z_p, \qquad p \in \mathcal{P}, \tag{4.13}$$

$$x_{pk} \geq 0, \qquad k \in \mathcal{K}, p \in \mathcal{P}_k, \tag{4.14}$$

$$z_p \in \{0,1\}, \qquad p \in \mathcal{P}, \tag{4.15}$$

$$\textbf{wMCND} \quad \min_{x_1,\ldots,x_T} \sum_{t=1}^{T} \left[\sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \setminus \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{flow}} x_{tpk} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{out}} x_{tpk} \right] \tag{4.16}$$

$$\text{s.t.} \sum_{p \in \mathcal{P}_k} x_{tpk} = \hat{y}_{tk}, \qquad t = 1, \ldots, T, k \in \mathcal{K}, \tag{4.17}$$

$$\sum_{k \in \mathcal{K}_p} x_{tpk} \leq u_p z_p, \qquad t = 1, \ldots, T, p \in \mathcal{P}, \tag{4.18}$$

$$x_{tpk} \geq 0, \qquad t = 1, \ldots, T, k \in \mathcal{K}, p \in \mathcal{P}_k. \tag{4.19}$$

The upper level selects $\hat{\mathbf{y}}^{\mathrm{p}}$ that minimizes the total fixed and variable costs over the whole tactical planning horizon. The objective function (4.9) hence depends on the design and flow variables from the lower levels **MCND** and **wMCND**, respectively.

In **wMCND** we introduce the flow variables x_{tpk} for commodity k on path p in period t and determine flows for each period minimizing variable cost (4.16) for a fixed design solution z given by **MCND**. Constraints (4.17) ensure that the demand is satisfied for each commodity in each period. The set of paths $\mathcal{P}_k$ includes outsourcing paths $\mathcal{P}_k^{\mathrm{out}}$ for a commodity k, so constraints (4.12) and (4.17) can always be satisfied. Constraints (4.18) enforce flows in each period to be only on selected paths and smaller than the capacity of the path. We draw the attention to the time series forecasts that occur in **wMCND** while the periodic demand estimates occur in **MCND**.

The decision variables of **wMCND** do not occur in the objective function (4.11) of **MCND**. Thus, for (z^*, x^*) an optimal solution of **MCND**, if z^* is feasible for **wMCND**, then z^* is

an optimal solution to **MCND-wMCND**. We can therefore make the following claim.

Claim 1 *If z^* is feasible for **wMCND**, then **MCND-wMCND** can be solved sequentially to optimality for a fixed $\hat{\mathbf{y}}^p$.*

In this work, we consider the four following mappings:

$$\mathbf{y}^{\mathrm{p}}_{\max} = h_1(\mathbf{y}_1, \ldots, \mathbf{y}_T) = \max_{t=1,\ldots,T} \mathbf{y}_t, \tag{4.20}$$

$$\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}} = h_2(\mathbf{y}_1, \ldots, \mathbf{y}_T) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t, \tag{4.21}$$

$$\mathbf{y}^{\mathrm{p}}_{\mathrm{q2}} = h_3(\mathbf{y}_1, \ldots, \mathbf{y}_T) = Q_2(\mathbf{y}_t, t = 1 \ldots, T), \tag{4.22}$$

$$\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}} = h_4(\mathbf{y}_1, \ldots, \mathbf{y}_T) = Q_3(\mathbf{y}_t, t = 1 \ldots, T) \tag{4.23}$$

which represent the maximum, mean, second quartile Q_2 and third quartile Q_3, respectively. The corresponding estimates from the forecasts are denoted $\hat{\mathbf{y}}^{\mathrm{p}}_{\max}$, $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$, $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{q2}}$ and $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{q3}}$. Given that we consider a discrete $\mathcal{Y}$ of small cardinality, we find the solution to **PDE** by solving **MCND-wMCND** for each $\mathbf{y}^{\mathrm{p}} \in \mathcal{Y}$. In the following section, we describe a specific instance of (4.9)-(4.19) in the context of tactical planning for intermodal rail transportation.

4.4 Application

We illustrate our approach on the intermodal network of CN, composed of 24 main intermodal terminals and 133 origin-destination (OD) pairs. Figure 4.3 depicts a map of the network. The railtracks extend from East to West of Canada and from Canada to South of the United States and gather 25 intermodal terminals. The railroad carries a variety of container types (20, 40, 45, 48 and 53-feet long), yet for tactical planning purposes they can be aggregated into either 40-feet or 53-feet containers. Indeed, 20-feet containers can be considered as half-40-feet containers. Other sizes (45, 48 and 53-feet) occupy the same locations on railcars (so-called slots Mantovani et al., 2018) as 53-feet containers. A commodity is defined by an origin, a destination and a type of container, and we consider a total of $K = 170$ commodities. Two commodities can hence have the same OD pair but differ by the type of container. The tactical period is a week and a tactical horizon lasts $T = 10$ weeks. The train schedule is repeated each week and CN operates so that demand over a week is satisfied. Assumption 1 therefore holds. More precisely, in case of insufficient capacity on intermodal trains, they use general cargo trains or additional ad-hoc intermodal trains to satisfy demand.

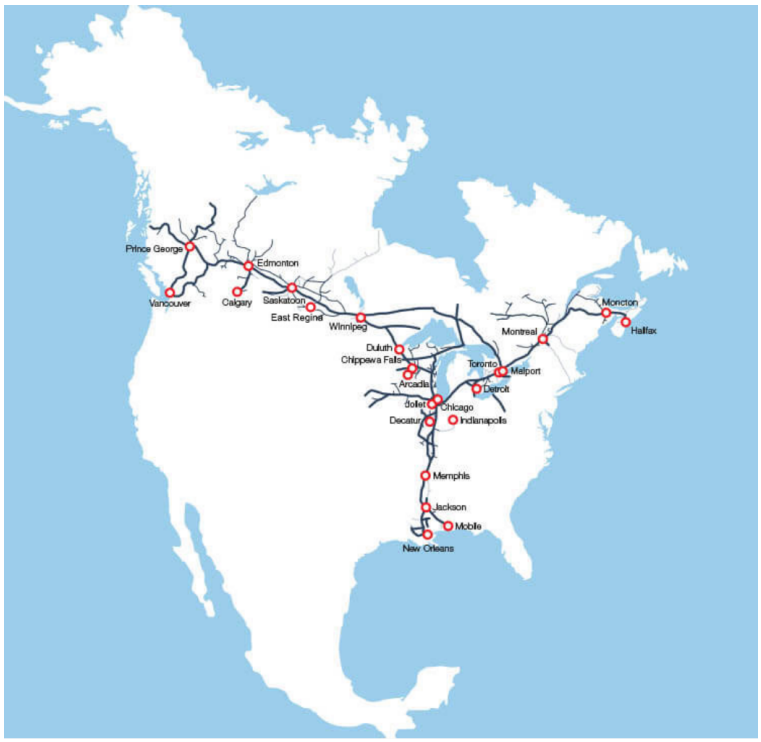


Figure 4.3 Intermodal Network of the Canadian National Railway Company. Source: www.cn.ca

The specific MCND problem to the intermodal network of CN is the tactical block planning problem (Morganti et al., 2020). A block refers to a consolidation of railcars. In this context, a set of railcars flowing as a single unit between a given OD pair and where containers loaded on the railcars have the same OD. Morganti et al. (2020) introduce a path-based Block Planning formulation (**BP**). It is defined using a space-time graph generated based on a schedule of intermodal trains. The graph contains 28,854 arcs and 15,269 nodes. A block is a path in this graph and the set is denoted $\mathcal{B}$ with $|\mathcal{B}| = 2,208$. We keep this notation to be consistent with Morganti et al. (2020) but note that the set $\mathcal{B}$ corresponds to the set of paths $\mathcal{P}$ in **MCND**. The set $\mathcal{B}$ contains a subset of *artificial blocks* $\mathcal{B}^{\text{artif}}$ whose role is to transport demand exceeding capacity. They hence correspond to the outsourcing paths. They are built without design cost, i.e., $C_b^{\text{design}} = 0, \forall b \in \mathcal{B}^{\text{artif}}$. Similarly to **MCND**, $\mathcal{B}_k$ and $\mathcal{B}_k^{\text{artif}}$ denote respectively the set of blocks and the set of artificial blocks for commodity k, and $\mathcal{K}_b$ the set of commodities that can use b.

Below we briefly describe **BP** and refer to Morganti et al. (2020) for more details. There are three categories of decision variables. First, the design variables $z_b, b \in \mathcal{B}$ where z_b equals one if block b is built, and zero otherwise. Second, integer flow variables $x_{bk}, k \in \mathcal{K}, b \in \mathcal{B}_k$ that equal the number of containers for commodity k transported on block b. Third, auxiliary

variables for the number of 40-feet v_b^{40} and 53-feet v_b^{53} double-stack platforms to carry the containers assigned to block $b \in \mathcal{B}$.

$$\textbf{BP} \quad \min_{x,z} \sum_{b \in \mathcal{B} \setminus \mathcal{B}^{\text{artif}}} C_b^{\text{design}} z_b + \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}_k \setminus \mathcal{B}_k^{\text{artif}}} C_{bk}^{\text{flow}} x_{bk} + \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}_k^{\text{artif}}} C_{bk}^{\text{out}} x_{bk} \tag{4.24}$$

$$\text{s.t.} \sum_{b \in \mathcal{B}_k} x_{bk} = y_k^{\text{p}}, \qquad k \in \mathcal{K}, \tag{4.25}$$

$$x_{bk} \leq y_k^{\text{p}} z_b, \qquad k \in \mathcal{K}, b \in \mathcal{B}_k, \tag{4.26}$$

$$v_b^{53} = \max \left[0, \left\lceil \frac{1}{2} \left(\sum_{k \in \mathcal{K}_b, \tau_k = 53} x_{bk} - \sum_{k \in \mathcal{K}_b, \tau_k = 40} x_{bk} \right) \right\rceil \right], b \in \mathcal{B}, \tag{4.27}$$

$$v_b^{40} = \left\lceil \frac{1}{2} \left(\sum_{k \in \mathcal{K}_b} x_{bk} \right) \right\rceil - v_b^{53}, \qquad b \in \mathcal{B}, \tag{4.28}$$

$$\sum_{b \in \mathcal{B}_a} \left(L^{40} v_b^{40} + L^{53} v_b^{53} \right) \leq u_a, \qquad a \in \mathcal{A}^{TM}, \tag{4.29}$$

$$z_b \in \{0, 1\}, \qquad b \in \mathcal{B}, \tag{4.30}$$

$$v_b^{40}, v_b^{53} \in \mathbb{N}, \qquad b \in \mathcal{B}, \tag{4.31}$$

$$x_{bk} \in \mathbb{N}, \qquad k \in \mathcal{K}, b \in \mathcal{B}_k. \tag{4.32}$$

The objective function (4.24) minimizes fixed and variable costs as well as a variable cost associated with outsourced demand (flow on artificial blocks). Constraints (4.25) ensure that the demand is satisfied by either the network capacity or outsourcing. Constraints (4.26) enforce flows to be on selected blocks only. Constraints (4.27) and (4.28) fix the number of platforms required to transport the demand. These constraints take into account how containers of different sizes can be double stacked. Since 40-feet platforms use less train capacity than 53-feet platforms, they are used whenever there are less 53-feet containers than 40-feet ones (40-feet container stacked in the bottom position and 53-feet container on top), and 53-feet platforms are used otherwise. Constraints (4.29) ensure that the train capacity, expressed in number of feet, is not exceeded. The platform lengths are denoted L^{40} and L^{53}, respectively. The train capacity, $u_a, a \in \mathcal{A}^{TM}$, is defined for the set of arcs in the space-time graph that represent moving trains, $\mathcal{A}^{TM}$. We denote by $\mathcal{B}_a$ the set of blocks that use train moving arc $a \in \mathcal{A}^{TM}$.

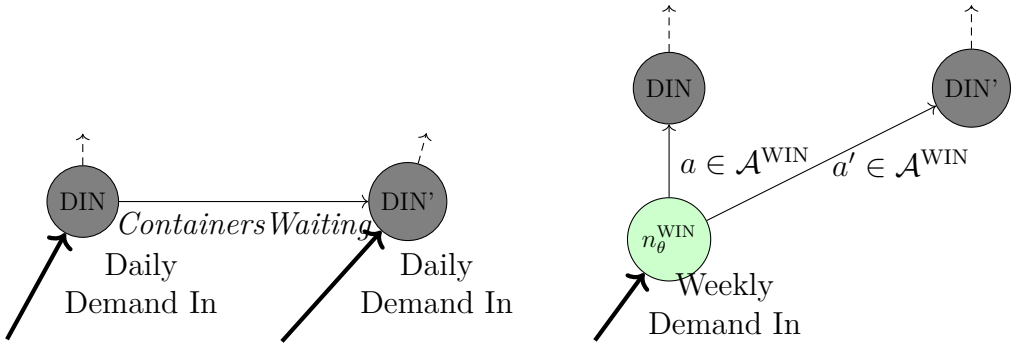
Finally, we give an order of magnitude of the size of the formulation. In the case of the instances we solve in this paper, there are over 386,000 variables and some 18,000 constraints.

4.4.1 Block Generation For Weekly Demand Inputs

While the tactical plan is computed for a weekly schedule, Morganti et al. (2020) assume that the time at which the demand arrives to the system within the week is given exogenously. We provide an illustrative example in Figure 4.4a. Demand enters the network via a node noted DIN which is associated to one admissible train departure node. Containers are either assigned to a block, or wait at the terminal, represented by flow on arcs called *Containers Waiting*.

Under Assumption 2, we propose a slightly different block generation so that the model optimally distributes the weekly demand over the train departures. For this purpose, we introduce a new set of nodes $\mathcal{N}^{\text{WIN}}$ such that there is one node WIN $n_\theta^{\text{WIN}} \in \mathcal{N}^{\text{WIN}}$ per terminal $\theta \in \Theta$, where Θ is the set of terminals in the network. We also introduce a new set of arcs $\mathcal{A}^{\text{WIN}} = \{(n_\theta^{\text{WIN}}, j) \mid \theta \in \Theta, j \in \mathcal{N}_\theta^{\text{DIN}}\}$, where $\mathcal{N}_\theta^{\text{DIN}}$ is the set of DIN nodes for terminal θ.

We illustrate this change compared to Morganti et al. (2020) in Figure 4.4b. For each terminal, the node WIN receives the weekly demand input and splits the commodity flow to the different DIN nodes on the arcs $\mathcal{A}^{\text{WIN}}$ at no cost. Instead of arriving at different points in time, demand now arrives in one source node and the model selects the optimal distribution over the week.



(a) Morganti et al. (2020)

(b) Our model

Figure 4.4 Illustration of the difference between Morganti et al. (2020) and our model

4.4.2 Periodic Demand Estimation Problem

We present below the formulation **PDE** specific to our application. The **MCND** formulation is replaced by **BP** and we introduce a weekly **BP** formulation, **wBP**. For the latter, we

introduce flow variables and auxiliary platform variables for each week t, $x_{tbk}, v_{tb}^{40}, v_{tb}^{53}, t \in \mathcal{T}, k \in \mathcal{K}, b \in \mathcal{B}$.

$$\textbf{PDE} \min_{\hat{\mathbf{y}}^{\mathrm{p}}} C^{\mathrm{PDE}} = \sum_{t=1}^{T} \left[\sum_{b \in \mathcal{B} \setminus \mathcal{B}^{\mathrm{artif}}} C_b^{\mathrm{design}} z_b + \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}_k \setminus \mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{flow}} x_{tbk} + \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{out}} x_{tbk} \right] \tag{4.33}$$

$$\text{s.t. } \hat{\mathbf{y}}^{\mathrm{p}} \in \mathcal{Y}, \tag{4.34}$$

$$\textbf{BP} \min_{x,z} \sum_{b \in \mathcal{B} \setminus \mathcal{B}^{\mathrm{artif}}} C_b^{\mathrm{design}} z_b + \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}_k \setminus \mathcal{B}_k^{\mathrm{artif}}} C_{bk}^{\mathrm{flow}} x_{bk} + \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}_k^{\mathrm{artif}}} C_{bk}^{\mathrm{out}} x_{bk} \tag{4.35}$$

$$\text{s.t. } \sum_{b \in \mathcal{B}_k} x_{bk} = \hat{y}_k^{\mathrm{p}}, \qquad k \in \mathcal{K}, \tag{4.36}$$

$$x_{bk} \leq \hat{y}_k^{\mathrm{p}} z_b, \qquad k \in \mathcal{K}, b \in \mathcal{B}_k, \tag{4.37}$$

$$v_b^{53} = \max \left[0, \left\lceil \frac{1}{2} \left(\sum_{k \in \mathcal{K}_b, \tau_k = 53} x_{bk} - \sum_{k \in \mathcal{K}_b, \tau_k = 40} x_{bk} \right) \right\rceil \right], \qquad b \in \mathcal{B}, \tag{4.38}$$

$$v_b^{40} = \left\lceil \frac{1}{2} \left(\sum_{k \in \mathcal{K}_b} x_{bk} \right) \right\rceil - v_b^{53}, \qquad b \in \mathcal{B}, \tag{4.39}$$

$$\sum_{b \in \mathcal{B}_a} \left(L^{40} v_b^{40} + L^{53} v_b^{53} \right) \leq u_a, \qquad a \in \mathcal{A}^{TM}, \tag{4.40}$$

$$z_b \in \{0, 1\}, \qquad b \in \mathcal{B}, \tag{4.41}$$

$$v_b^{40}, v_b^{53} \in \mathbb{N}, \qquad b \in \mathcal{B}, \tag{4.42}$$

$$x_{bk} \in \mathbb{N}, \qquad k \in \mathcal{K}, b \in \mathcal{B}_k, \tag{4.43}$$

$$\textbf{wBP} \quad \min_{x_1, \ldots, x_T} \sum_{t=1}^{T} \sum_{k \in \mathcal{K}} \left[\sum_{b \in \mathcal{B}_k \setminus \mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{flow}} x_{tbk} + \sum_{b \in \mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{out}} x_{tbk} \right] \tag{4.44}$$

$$\text{s.t. } \sum_{b \in \mathcal{B}_k} x_{tbk} = \hat{y}_{tk}, \qquad t \in \mathcal{T}, k \in \mathcal{K}, \tag{4.45}$$

$$x_{tbk} \leq \hat{y}_{tk} z_b, \qquad t \in \mathcal{T}, k \in \mathcal{K}, b \in \mathcal{B}_k, \tag{4.46}$$

$$v_{tb}^{53} = \max \left[0, \left\lceil \frac{1}{2} \left(\sum_{k \in \mathcal{K}_b, \tau_k = 53} x_{tbk} - \sum_{k \in \mathcal{K}_b, \tau_k = 40} x_{tbk} \right) \right\rceil \right], t \in \mathcal{T}, b \in \mathcal{B}, \tag{4.47}$$

$$v_{tb}^{40} = \left\lceil \frac{1}{2} \left(\sum_{k \in \mathcal{K}_b} x_{tbk} \right) \right\rceil - v_{tb}^{53}, \qquad t \in \mathcal{T}, b \in \mathcal{B}, \tag{4.48}$$

$$\sum_{b \in \mathcal{B}_a} \left(L^{40} v_{tb}^{40} + L^{53} v_{tb}^{53} \right) \leq u_a, \qquad t \in \mathcal{T}, a \in \mathcal{A}^{TM}, \tag{4.49}$$

$$v_{tb}^{40}, v_{tb}^{53} \in \mathbb{N}, \qquad t \in \mathcal{T}, b \in \mathcal{B}, \tag{4.50}$$

$$x_{tbk} \in \mathbb{N}, \qquad t \in \mathcal{T}, k \in \mathcal{K}, b \in \mathcal{B}_k. \tag{4.51}$$

The objective function (4.33) has the same structure as (4.9), with design costs, flow costs and outsourcing costs. We note that we define **BP** (4.35)-(4.43) using periodic demand

$\hat{\mathbf{y}}^{\mathrm{p}}$, while we define **wBP** (4.44)-(4.51) for fixed design variables and weekly demand $\hat{\mathbf{y}}_t$. Furthermore, there is no fixed cost associated with $b \in \mathcal{B}^{\mathrm{artif}}$. Therefore, a solution for **BP** is always feasible for **wBP**. Using Claim 1, we can solve **BP-wBP** sequentially.

4.5 Computational Results

Our dataset contains the observed daily container shipments of all types of containers on each origin-destination pair of the rail network, collected over 6 years, from December 2013 to November 2019. The tactical plan is weekly. To ensure the observed demand is not constrained by the supply, we do a weekly aggregation of 2,226 observations of daily demand for each commodity. This results in 318 observations per commodity. Hence, y_{tk} refers to the number of containers of type τ_k to be carried from origin o_k to destination d_k during week t.

In this section, we first provide a descriptive analysis of the data. We report the results of the time series forecasting models in Section 4.5.2. Finally, in Section 4.5.3 we report results for the periodic demand estimation problem. Note that, for confidentiality reasons, we only report relative numbers in all results.

4.5.1 Descriptive Analysis

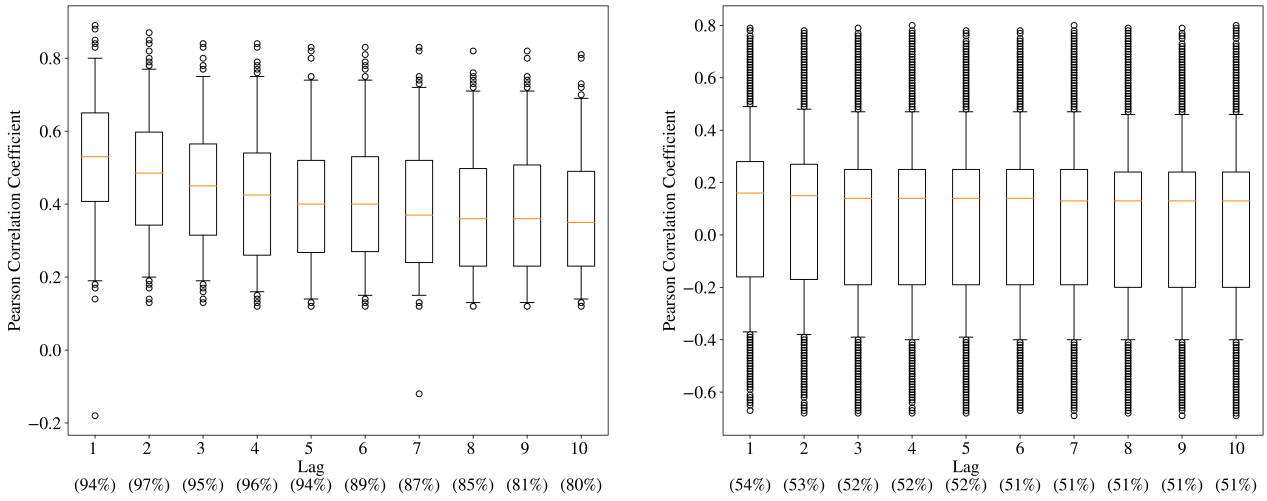
The large size of the transportation network and the large number of commodities suggest that they can be spatiotemporal correlated. We provide here an analysis of the different types of correlations we identified: between commodities and between commodities and weather.

Correlation Between Commodities

We start by analyzing interweek correlation by computing, for each commodity, estimates of the Pearson correlation coefficient between weekly shipments over successive weeks. Figure 4.5a presents the distribution of the coefficient over the commodities for lags from 1 to 10 weeks. It shows that there are substantial positive correlations between weeks for all commodities. Correlations are strong between successive weeks and are slightly weaker for longer lags.

We now turn our attention to intercommodity correlations. For each pair of commodities, we compute estimates of the Pearson correlation coefficient between weekly shipments over successive weeks. We present in Figure 4.5b the distribution of those correlations for lags from 1 to 10 weeks. There are 170 commodities, hence 28,730 intercommodity correlation coefficients to examine at each lag. Correlations vary across pairs of commodities. Some are

large (positive or negative) for short and longer lags, while 50% of the pairs have a weak correlation between -0.2 and 0.2. Outliers at each time lag represent 9% of the significant coefficients.



(a) Interweek correlation

(b) Intercommodity correlation

Figure 4.5 Distribution of the Pearson correlation coefficient over commodities for the two different types of correlations. We consider only the coefficients which met the 95% confidence level threshold and we indicate their proportion in parenthesis.

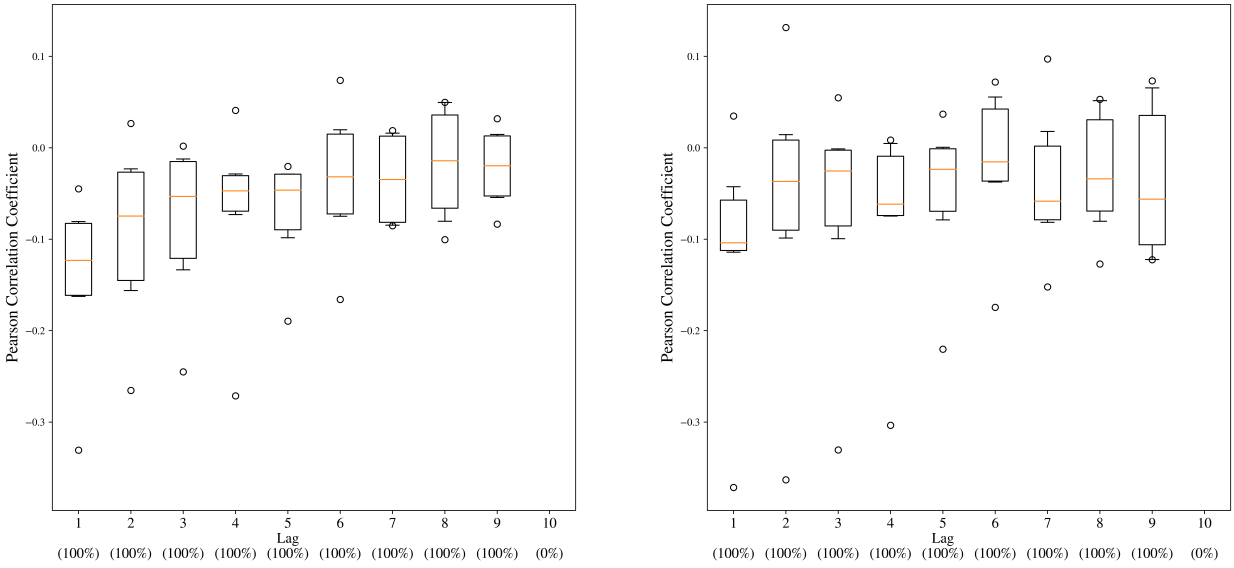
In summary, the data shows evidence of two types of correlation: strong positive interweek correlation between the weekly shipments over successive weeks, and strong intercommodity correlation between the weekly shipments of different commodities. Interweek correlation highlights the potential need for an autoregressive model while intercommodity correlation highlights a potential need for a model able to learn various dependence structures from the data.

Correlation Between Demand and Weather

Weather can be an important aspect for rail freight transportation, especially in North America where railways extend over the subcontinent which is subject to major weather disruptions such as snowstorms. To assess the importance of weather on shipments, we use meteorological data and estimate the Pearson correlation coefficient between weekly shipments and several weather indicators. The data comes from National Oceanic and Atmospheric Administration (NOAA, https://www.ncdc.noaa.gov/) for terminals in the United States and from Statistics Canada (https://www.statcan.gc.ca) for terminals in Canada. More precisely, we use the av-

erage daily temperature and total daily snowfall in centimeters for the main 17 terminals for the complete time range covered by the data. We compute the weekly temperature as the average temperature over the week and the accumulated snow (cm) as the sum of the daily values. For each terminal, at each week t, we sum the total departing and arriving demand over all commodities.

Figure 4.6 shows the distribution over the terminals of the Pearson correlation coefficient between accumulated snow over week t and departing and arriving demand at week $t + \text{lag}$. We note a negative correlation for both arriving and departing demand for successive weeks. It is weaker for longer lags.

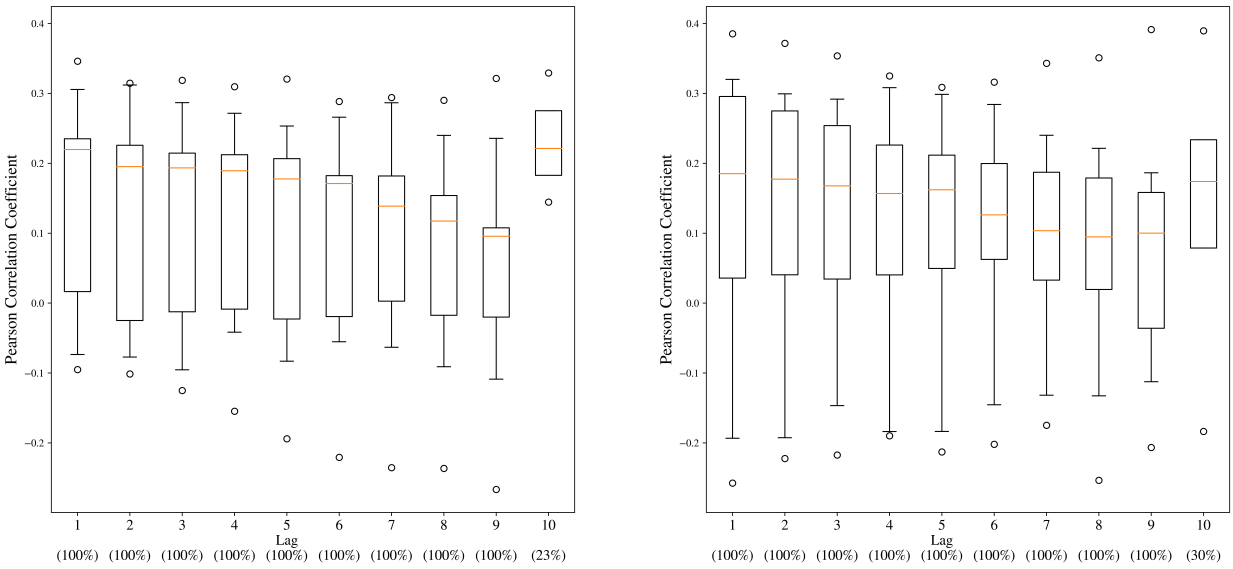


(a) Departing Demand

(b) Arriving Demand

Figure 4.6 Distribution of the Pearson correlation coefficient between accumulated snow over week t and demand (arriving and departing) at week $t + \text{lag}$ over all the terminals. We consider only the coefficients which met the 95% confidence level threshold and we indicate their proportion in parenthesis.

We present in Figure 4.7 the distribution over the terminals of the Pearson correlation coefficient between the average temperature and the demand arriving or departing over terminals. It shows an average positive correlation between demand and temperature which is weaker for longer lags. Some terminals have, however, negative correlations. This can be explained by a seasonality effect. On the one hand, summer and spring are busier periods for most ODs and temperature are higher than the rest of the year. On the other hand, for import terminals, fall and the Chinese New Year are busier periods when temperatures are lower.



(a) Departing Demand

(b) Arriving Demand

Figure 4.7 Distribution of the Pearson correlation coefficient between average temperature at week t and demand (arriving and departing) at week $t + \text{lag}$ over all the terminals. We consider only the coefficients which met the 95% confidence level threshold and we indicate their proportion in parenthesis.

Correlation between demand and weather highlights the potential need for a forecasting model which can include weather features. In the following section we present forecasting results for different forecasting models, ranging from very simple ones based on strong independence assumptions to neural networks that relax some of those assumptions.

4.5.2 Time Series Forecasting Results

We divide the dataset into a training, validation and test sets. Time series forecasting models require the last seen observed data before predicting the demand to come. Thus, we use the first 5 years of data for training (December 2013 - December 2018), the next 4 months of data for validation (January 2019 - April 2019) and the last 7 months (May 2019 - November 2019) for testing. At each week t_0 in the dataset, we forecast demand for all commodities for $t = t_0 + 1, \ldots, t_0 + T$. We consider $T = 10$ weeks. To simplify the notation, we assume $t_0 = 0$ and refer to the estimates by $t = 1, \ldots, T$.

We compare four types of models detailed below: a simple model that uses last observed values as prediction (CONSTANT) as well as autoregressive (AR), feedforward neural network (FFNN) and recurrent neural network (RNN) models.

CONSTANT This is the simplest possible model. It is based on the assumptions that commodities $k \in \mathcal{K}$ are independent and that the demand from observed week t_0 is the forecasts for the next T weeks:

$$\hat{\mathbf{y}}_1 = \cdots = \hat{\mathbf{y}}_T = \mathbf{y}_0. \tag{4.52}$$

AR For each commodity, we fit an autoregressive $AR(p)$ process on the training data. This implies that the commodities are treated as independent. We use the estimated coefficients $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)^T$ to compute the forecasts on the test set. The multistep forecasts for $t > 1$ are obtained using

$$\hat{y}_{tk} = \hat{\phi}_{1k}\hat{y}_{t-1,k} + \cdots + \hat{\phi}_{tk}y_{0k} + \cdots + \hat{\phi}_{pk}y_{t-p,k}. \tag{4.53}$$

FFNN and RNN We leverage the capacity of these models to forecast demand for multiple commodities simultaneously. As opposed to CONSTANT and AR, they relax the independence assumption on the commodities. We build one main neural network architecture described in Figure 4.8. It is composed of 2 inputs layers, a stack of hidden dense or recurrent layers and one output layer. We consider several variants of this architecture. When the layers in the grey square are dense (or feed-forward), it forms the feed-forward archi-

tecture (FFNN). When the layers are LSTM, it forms the recurrent architecture (RNN). The dimension of the output layer is equal to the number of commodities that we predict simultaneously.

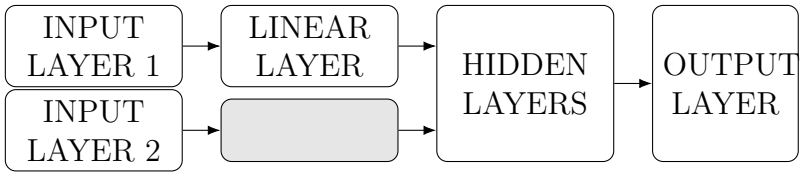


Figure 4.8 Neural network architecture

At each period t_0, the output layer computes the forecast for the next time step $t_0 + 1$. Input Layer 1 is dedicated to external data such as weather features to model the correlation between demand and weather. Input Layer 2 is dedicated to the autoregressive modeling. It takes as input the demand of the previous weeks of all commodities, either observed or forecasted. When we generate the multistep demand forecasts, for Input Layer 2, we use $\hat{\mathbf{y}}_{t_0+1}$ to forecast demand at $t_0 + 2$, $\hat{\mathbf{y}}_{t_0+1}$ and $\hat{\mathbf{y}}_{t_0+2}$ for $t_0 + 3$, until $t_0 + T$ with $\hat{\mathbf{y}}_{t_0+1}, \ldots, \hat{\mathbf{y}}_{t_0+T-1}$.

We evaluate several variants of the architecture to model the spatiotemporal correlations and the correlation with weather. They differ depending on the considered input data: lagged observed/forecasted demand and/or weather features. To train, validate and test the model, we use real observed weather data described in Section 4.5.1. At prediction time, such information is not available and we would then rely on weather forecasts. These results are hence designed to assess the potential for weather related features in an optimistic setting in regards to their accuracy.

Forecasting 170 commodities simultaneously requires a rich and large dataset. Our dataset is fairly limited as it contains only 318 weekly demand data points for each commodity. To facilitate the forecasting task, we create a partition of $\mathcal{K}$ and train a neural network for each set of the partition. For this purpose, we split the set of commodities into 2 subsets: $\mathcal{K}_{53}$ which contains commodities of 53-feet container type and $\mathcal{K}_{40}$ which contains commodities of 40-feet container ($|\mathcal{K}_{53}| = 58$ and $|\mathcal{K}_{40}| = 112$). Table 4.1 summarizes the model variants. Their name include a letter "W" if weather features are used to train the model, and "SPLIT" to indicate a partition of $\mathcal{K}$.

The neural networks are trained with the backpropagation algorithm and the stochastic gradient descent using a Mean Squared Error (MSE) loss. For each model, we do a hyperparameter search with the Tree of Parzen Estimators implemented in the python library Hyperopt (Bergstra et al., 2013). We select the set of hyperparameters which minimizes the MSE on the validation dataset. We report the detailed input features and the chosen set of

Table 4.1 Features and set of commodities for each variant of the neural network architecture evaluated

	Commodities	Weather features	Autoregressive Features
RNN	$\mathcal{K}$		✓
FFNN	$\mathcal{K}$		✓
RNN-W	$\mathcal{K}$	✓	✓
FFNN-W	$\mathcal{K}$	✓	✓
RNN-W-SPLIT1	$\mathcal{K}_{53}$	✓	✓
RNN-W-SPLIT2	$\mathcal{K}_{40}$	✓	✓
FFNN-W-SPLIT1	$\mathcal{K}_{53}$	✓	✓
FFNN-W-SPLIT2	$\mathcal{K}_{40}$	✓	✓

hyperparameters for each trained model in the Appendix (Table A.1).

Forecast Accuracy Measures

We measure the accuracy on the test set with two metrics: the Weighted Absolute Percentage Error (WAPE)

$$\text{WAPE}_k = \frac{\sum_{t_0 \in \mathcal{D}_{\text{test}}} \sum_{t=1}^{T} |y_{t_0+t,k} - \hat{y}_{t_0+t,k}|}{\sum_{t_0 \in \mathcal{D}_{\text{test}}} \sum_{t=1}^{T} y_{t_0+t,k}} \times 100, \ k = 1, \ldots, K, \tag{4.54}$$

where $\mathcal{D}_{\text{test}}$ is the test set of size N, and the Root Mean Squared Error (RMSE)

$$\text{RMSE}_k = \sqrt{\frac{1}{N \times T} \sum_{t_0 \in \mathcal{D}_{\text{test}}} \sum_{t=1}^{T} (y_{t_0+t,k} - \hat{y}_{t_0+t,k})^2}, \ k = 1, \ldots, K. \tag{4.55}$$

The WAPE is a weighted version of the Mean Absolute Percentage Error (MAPE) which handles small or zero demand values. Low demands are frequent in our data for some commodities with sparse demand over the year. Hence the importance of having a metric independent to the scale of the time series such as WAPE. The RMSE puts a high weight on large errors, which is also an important metric to consider.

Results

Table 4.2 reports the performance metrics averaged over all commodities. We note that the metrics for models based on a partition of the commodities (SPLIT) are averaged over the partitions. The results show that the AR has the best performance and it is considerably better than the NN models. The CONSTANT baseline has a WAPE close to AR but a

considerably worse RMSE. The descriptive statistics in Section 4.5.1 show strong intercommodity correlations. Nevertheless, the AR model, based on the assumption that demands for commodities are independent, performs better than the NN models where this assumption is relaxed. Neural networks have more parameters to fit on the same limited data. While they have the capacity to model non-linear relationships, they also require more data to be trained. We believe that our data containing only 318 observations for each commodity is too limited for training the NN models. Moreover, we note that these findings are consistent with other studies (e.g., Makridakis et al., 2018). That is, basic time series models outperform deep learning models on difficult time series forecasting tasks, such as this one.

Table 4.2 Performance metrics of the forecasting models averaged over all commodities. The best and second best metric values are highlighted in bold.

Model	RMSE	WAPE
CONSTANT	86.0	**34.7%**
AR	**78.0**	**34.0%**
FFNN	105.2	38.7%
FFNN-W	**84.8**	37.1%
FFNN-W-SPLIT	90.4	37.2%
RNN	105.1	37.8%
RNN-W	86.3	37.8%
RNN-W-SPLIT	85.2	38.6%

The results for the deep learning models confirm that adding weather features and considering all commodities simultaneously help to improve the performance. This is consistent with the descriptive statistics reported in Section 4.5.1. Previous demands for all commodities constitute relevant information to consider.

Henceforth, we keep two forecasting models when analyzing periodic demand results: the overall best performing model (AR) as well as the best deep learning model (FFNN-W).

4.5.3 Periodic Demand Estimation

We divide the results related to periodic demand estimation into two parts. The purpose is to disentangle the errors associated with the periodic demand estimation from those associated with the demand forecasts. The two parts are briefly described in the following:

- **Analysis 1: the impact of periodic demand estimation**. We assume we have no forecast errors, i.e., the carrier knows perfectly the demand to come for the planning horizon. The periodic demand is estimated with the mappings (4.20)-(4.23) from

historical data (ground truth values) and we compute the tactical costs generated by those periodic demands.

- **Analysis 2: the impact of imperfect demand forecasts**. We estimate the periodic demands from the forecasts obtained with the AR and FFNN-W models. We compute the associated tactical costs and compare them to Analysis 1.

We consider two demand instances – $I1$ and $I2$ – from the test set of the forecasting models. They are from two distinct periods: $I1$ corresponds to end of spring and beginning of summer, from May 6th to July 14th, 2019 and I2 corresponds to end of summer and beginning of fall, from July 29th to October 6th, 2019. Both instances have $K = 170$ commodities and a tactical planning horizon of $T = 10$ weeks. Instance $I2$ corresponds to a busier period for the carrier: Figure 4.9 shows the difference of the total demand summed over all commodities of $I2$ at each week relative to the total demand of $I1$. We note that the total demand for $I2$ can be up to 20% higher than its $I1$ counterpart.

All the results are generated by fixing the variable $\hat{\mathbf{y}}^{\mathrm{p}}$ in **PDE** and sequentially solving **BP-wBP**. We recall that the periodic demand $\mathbf{y}^{\mathrm{p}}$ is a mapping from the real demand values, $\hat{\mathbf{y}}^{\mathrm{p}}$ is a mapping from the demand forecasts, $\mathbf{Y}$ is the matrix of real demand values and $\hat{\mathbf{Y}}$ is the matrix of demand forecasts. We denote by $\mathbf{BP}(\mathbf{y}^{\mathrm{p}})$ and $\mathbf{BP}(\hat{\mathbf{y}}^{\mathrm{p}})$ when **BP** is solved with $\mathbf{y}^{\mathrm{p}}$ and $\hat{\mathbf{y}}^{\mathrm{p}}$, respectively, in constraints (4.36). Furthermore, we denote by $\mathbf{wBP}(\mathbf{Y})$ and $\mathbf{wBP}(\hat{\mathbf{Y}})$ when **wBP** is solved with demand values $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, respectively, in constraints (4.45).

Analysis 1: The Impact of Periodic Demand Estimation

The analysis in this section is based on two sets of results for demand instances $I1$ and $I2$:

- **Reference**: We solve $\mathbf{BP}(\mathbf{y}_t)$ for $t = 1, \ldots, T$. In other words, we do not restrict the demand to be periodic and we obtain a tactical cost $C^{\mathrm{PDE}}_{\mathrm{ref}}$ (4.33).
- **Periodic**: We solve $\mathbf{BP}(\mathbf{y}^{\mathrm{p}}_{\cdot}) - \mathbf{wBP}(\mathbf{Y})$ with the four periodic demand vectors $\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$, $\mathbf{y}^{\mathrm{p}}_{\mathrm{max}}$, $\mathbf{y}^{\mathrm{p}}_{\mathrm{q2}}$ and $\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}}$. For each solution, we compute the percentage gap to the reference cost $C^{\mathrm{PDE}}_{\mathrm{ref}}$.

Table 4.3 reports the results where each number is the percentage gap to the reference cost. The cost C^{PDE} is a non-trivial trade-off between the three components C^{design}, C^{flow} and C^{out}, that are linked with demand. Hence we cannot analyze each component independently of the others. The parameter values in the objective function (4.33) have been chosen with

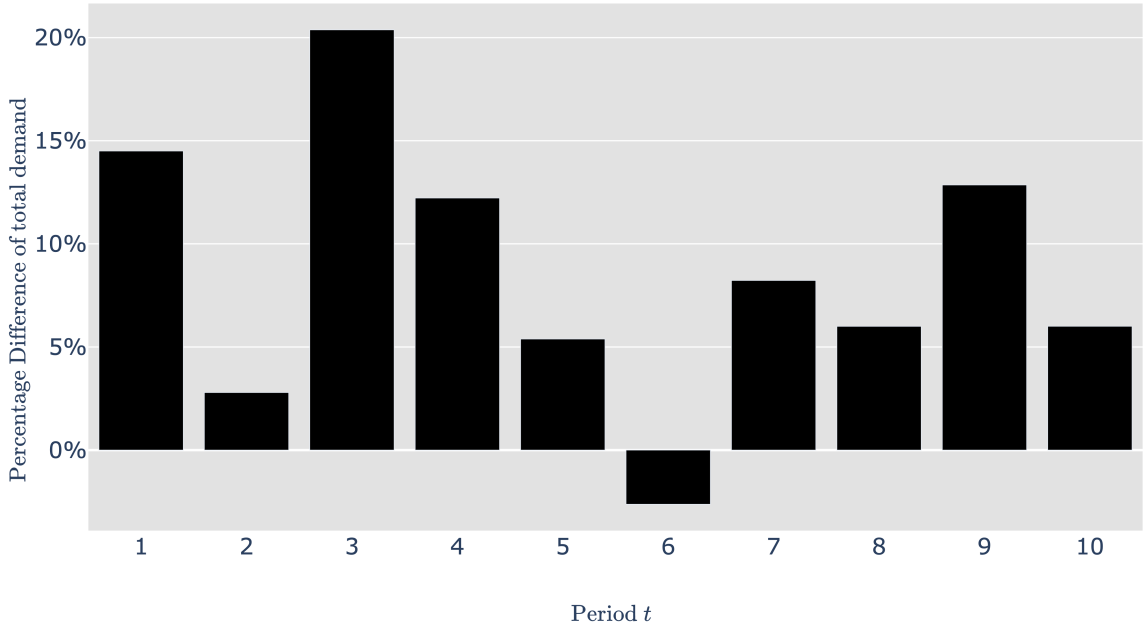


Figure 4.9 Percentage difference of the total demand at each week of $I2$ relative to $I1$

Table 4.3 Tactical costs relative to the reference. We use historical data for the demand at each week, that is $\mathbf{wBP}(\mathbf{Y})$. The minimum value for C^{PDE} indicates the optimal periodic demand estimate in the subset $\mathcal{Y}$ we consider.

	Periodic Definition $\mathbf{y}^{\text{p}}$ in **BP**	Percentage difference of costs C^{PDE}	C^{design}	C^{flow}	C^{out}
$I1$	$\mathbf{y}^{\text{p}}_{\text{max}}$	77.3%	-2.8%	-3.4%	84.0%
	$\mathbf{y}^{\text{p}}_{\text{mean}}$	93.3%	-16.9%	-4.4%	101.3%
	$\mathbf{y}^{\text{p}}_{\text{q2}}$	93.8%	-16.8%	-3.3%	101.8%
	$\mathbf{y}^{\text{p}}_{\text{q3}}$	**28.4%**	-7.5%	-1.2%	30.8%
$I2$	$\mathbf{y}^{\text{p}}_{\text{max}}$	73.0%	-14.1%	-5.7%	76.6%
	$\mathbf{y}^{\text{p}}_{\text{mean}}$	40.6%	-20.0%	-2.8%	42.6%
	$\mathbf{y}^{\text{p}}_{\text{q2}}$	41.6%	-23.2%	-2.8%	43.6%
	$\mathbf{y}^{\text{p}}_{\text{q3}}$	**23.1%**	-14.8%	-2.6%	24.3%

CN to best represent their costs. We note that $C^{\text{PDE}}_{\text{ref}}$ (4.33) constitutes a lower bound: at each week, the blocks built are specific for the demand to best exploit the network capacity.

Two important findings emerge. First, the tactical cost has an important variation over the different periodic demand estimates. This underlines the importance of the periodic demand estimation problem. Second, using another estimate than the commonly used mean periodic demand $\mathbf{y}^{\text{p}}_{\text{mean}}$ can lead to an important cost reduction. In our case using the third quartile $\mathbf{y}^{\text{p}}_{\text{q3}}$ reduces the total costs by 33.6% in $I1$, and by 12.4% in $I2$. When a smaller periodic demand estimate is used, $\mathbf{y}^{\text{p}}_{\text{mean}}$ and $\mathbf{y}^{\text{p}}_{\text{q2}}$ for instance, less blocks are built, at the expense of outsourced demand.

We analyze the total periodic demand relative to the total actual demand to explain the results. We report the values in Table 4.4. The total periodic demand of $\mathbf{y}^{\text{p}}_{\text{max}}$ is almost three times larger than the one of $\mathbf{y}^{\text{p}}_{\text{q3}}$ in both instances which explains the large gap in their costs. The design made from solving $\mathbf{BP}(\mathbf{y}^{\text{p}}_{\text{q3}})$ is based on a lower estimation of the periodic demand than $\mathbf{BP}(\mathbf{y}^{\text{p}}_{\text{max}})$ which generates fewer blocks in the network. The advantage is that they are better used: the design from solving $\mathbf{BP}(\mathbf{y}^{\text{p}}_{\text{max}})$ is made of a large number of small blocks which cannot accommodate all demands. Hence the increase in outsourced demand. We highlight that while the increase in outsourced demand seems large, it represents only few percent of the total demand.

Table 4.4 Total periodic demand. The point of reference is indicated by a dash.

	$I1$	$I2$
$\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} y_{tk}$	-	-
$\mathbf{y}^{\text{p}}_{\text{max}}$	34.8%	31.8%
$\mathbf{y}^{\text{p}}_{\text{mean}}$	0.0%	0.0%
$\mathbf{y}^{\text{p}}_{\text{q2}}$	-1.0%	-0.5%
$\mathbf{y}^{\text{p}}_{\text{q3}}$	11.8%	11.7%

Analysis 2: The Impact of Imperfect Demand Forecasts

In practice, carriers rely on demand forecasts to estimate a periodic demand and build the design for the tactical planning horizon. Then, at each week of the horizon, they adapt operationally the tactical plan based on observed demand. In this section, we follow our methodology and analyze the quality of the solution a posteriori.

We proceed in two steps: First, we estimate the periodic demand by solving $\mathbf{BP}(\hat{\mathbf{y}}^{\text{p}}) - \mathbf{wBP}(\hat{\mathbf{Y}})$ for each mapping h and select the one minimizing C^{PDE} (4.33). Second, we assess the tactical cost associated with the periodic demand estimate. For this purpose we solve

$\mathbf{BP}(\hat{\mathbf{y}}^{\mathrm{p}}) - \mathbf{wBP}(\mathbf{Y})$. Note that we use real demand values $\mathbf{Y}$ to accurately assess this cost. Hence, it is affected by two combined sources of error: the one of the periodic demand estimation and the forecast error, both discussed separately in previous sections.

Step 1: Periodic demand estimation We report results in Table 4.5 and indicate by a dash the point of reference ($\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$). We note that the value of $\hat{\mathbf{Y}}$ depends on both the forecasting model and the instance. In other words, we can only compare results from the same forecasting model and the same instance. The results show that, for both forecasting models and both instances, the tactical costs are minimized with $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$. In Analysis 1 with historical data, tactical costs were minimized with $\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}}$. This is because forecasting models smooth demand and struggle to forecast accurately the peaks. With historical data, we have access to the maximum of demand which can be an outlier, thus expensive. Following the methodology, we choose the periodic demand $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$ for Step 2 for both forecasting models.

Table 4.5 Percentage difference of tactical cost C^{PDE} resulting from solving $\mathbf{BP}(\hat{\mathbf{y}}^{\mathrm{p}}) - \mathbf{wBP}(\hat{\mathbf{Y}})$ with forecasts of demand from two models: AR and FFNN-W. For each model and each instance, we compare the value with the one from $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$.

		Forecasting Model	
	Periodic Demand $\mathbf{y}^{\mathrm{p}}$ in **BP**	AR	FFNN-W
$I1$	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	**-38.4%**	**-32.2%**
	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$	-	-
	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{q2}}$	17.5%	28.0%
	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{q3}}$	-30.8%	-18.8%
$I2$	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	**-7.4%**	**-28.4%**
	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$	-	-
	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{q2}}$	12.7%	29.2%
	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{q3}}$	8.8%	-11.7%

Step 2: Assessment of the tactical costs Table 4.6 reports the relative costs resulting from solving $\mathbf{BP}(\hat{\mathbf{y}}^{\mathrm{p}}) - \mathbf{wBP}(\mathbf{Y})$. We use the same reference as in Analysis 1, that is, the lower bound on C^{PDE}. In addition to the best periodic demand identified by our methodology, we report the result for $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$. Consistent with the findings in Analysis 1, we note that using the latter leads to a large increase in costs. Despite the relatively large forecast errors reported

in Table 4.2, the results show that estimating periodic demand using our methodology can even reduce the costs compared to $\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$ computed on *perfect information* (historical data).

Table 4.6 Percentage difference of tactical costs resulting from solving $\mathbf{BP}(\hat{\mathbf{y}}^{\mathrm{p}}) - \mathbf{wBP}(\mathbf{Y})$. The point of reference is the reference used in Analysis 1 ($\mathbf{BP}(\mathbf{y}_t)$ for $t = 1, \ldots, T$).

		Periodic Demand $\mathbf{y}^{\mathrm{p}}$ in **BP**	Percentage difference of costs C^{PDE}	C^{design}	C^{flow}	C^{out}
	Historical data	$\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$	93.3%	-16.9%	-4.4%	101.3%
	Historical data	$\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}}$	28.4%	-7.5 %	-1.2%	30.8%
$I1$	AR	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	55.9%	-17.8%	-1.9%	60.6%
	AR	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$	125.1%	-14.1%	-3.9%	135.8%
	FFNN-W	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	48.2%	-12.3%	-2.5%	52.4%
	FFNN-W	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$	117.4%	-16.2%	-6.0%	127.5%
	Historical data	$\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$	40.6%	-20.0%	-2.8 %	42.6%
	Historical data	$\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}}$	23.1%	-14.8%	-2.6 %	24.3%
$I2$	AR	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	50.1%	-23.6%	-10.0%	52.8%
	AR	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$	137.6%	-31.8%	-12.1%	144.4%
	FFNN-W	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	80.0%	-25.3%	-7.1 %	83.9%
	FFNN-W	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{mean}}$	164.4%	-31.1%	-17.8%	172.6%

We report the total demand values in Table 4.7. For demand instance $I1$, both periodic demand $\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$ overestimate the reference demand. The design built is capable of handling more demand, which results in less outsourced demand. However, they underestimate the periodic demand $\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}}$, which lead to a less important decrease in outsourced demand. In instance $I2$, the periodic $\hat{\mathbf{y}}^{\mathrm{p,AR}}_{\mathrm{max}}$ overestimates the total demand yet it leads to an increase in outsourced demand. This is because $\hat{\mathbf{y}}^{\mathrm{p,AR}}_{\mathrm{max}}$ either overestimates demand for commodities that already lack capacity with $\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$, or it underestimates demand for some commodities for which blocks are then not built in the design and are consequently outsourced at each week.

Table 4.7 Total periodic demand summed over commodities. The point of reference is indicated by a dash.

		$I1$	$I2$
Historical data	$\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$	-	-
Historical data	$\mathbf{y}^{\mathrm{p}}_{\mathrm{q3}}$	11.9%	11.7%
AR	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	10.2%	0.7%
FFNN-W	$\hat{\mathbf{y}}^{\mathrm{p}}_{\mathrm{max}}$	10.0%	-1.4%

4.6 Conclusion and Future Research

Tactical planning is essential to freight carriers as it allows to, e.g., design the service network to meet expected demand while minimizing cost. In this work we focused on large-scale tactical planning that is restricted to deterministic models for the sake of computational tractability. Even though estimates of periodic demand is a central input to such models, the associated estimation problem has not been studied in the literature. In this paper we addressed this gap: We formally introduced the periodic demand estimation problem and we proposed a methodology that proceeds in two steps. The first step consists in using a time series forecasting model to predict demand for each period in the tactical planning horizon. The second step defines periodic demand as a solution to a multilevel mathematical program that explicitly connects the estimation problem to the tactical planning problem of interest. This allows to estimate periodic demand such that the costs are minimized. Since the origin-destination demand matrices typically are unbalanced, this can be of importance as the cost of forecast errors is not evenly distributed across commodities.

We reported results for a real large-scale application at the Canadian National Railway Company. The results clearly showed the importance of the periodic demand estimation problem when compared to the approach commonly used in practice. The latter consists in averaging the time series forecasts over the tactical planning horizon. Compared to this practice, the results showed that using another estimate can lead to a substantial reduction in cost. As expected, the results also showed that the time series forecasting problem is difficult and the forecast errors hence are relatively large. Nevertheless, the periodic demand estimates that resulted from the proposed methodology still led to costs that were comparable, or even better, than those obtained by using the average demand baseline computed on *perfect information* (i.e., no forecast error). Moreover, the costs were substantially reduced compared to averaging the forecasts.

In terms of exposition, we chose to limit the methodology to the MCND formulation. However, the methodology applies to other cyclic network design formulations. Similarly, adaptation of the tactical plan in each period (**wMCND** and **wBP** formulations) can also be represented differently from this paper. The methodology hinges on the separation between the design variables that are fixed for all periods in the tactical planning horizon while the flow decisions are not. The adaptation of the flow decisions serves as a proxy for operational costs.

Given that we introduced a new problem in this paper, it opens up a number of directions for future research. First, improving the time series forecasts (step one in the methodology).

Second, extending the feasible set of periodic demand values to more general mappings and devise an effective solution approach for this case. The work reported in this paper constituted a first step in addressing the periodic demand estimation problem that hitherto has been overlooked in the literature. We showed that adequately addressing it can lead to important cost reductions.

Acknowledgments

This research was funded by the Canadian National Railway Company Chair in Optimization of Railway Operations at Université de Montréal and a Collaborative Research and Development Grant from the Natural Sciences and Engineering Research Council of Canada (CRD-477938-14). Moreover, we gratefully acknowledge the close collaboration with personnel from different divisions of CN.

CHAPTER 5 ARTICLE 2: A TWO-STEP HEURISTIC FOR THE PERIODIC DEMAND ESTIMATION PROBLEM

The text of this chapter is the one of the research paper *Solution Algorithms for the Periodic Demand Estimation Problem* to be submitted to the journal *Computers & Operations Research*.

Authors Greta Laage, Emma Frejinger, Gilles Savard

Abstract The Periodic Demand Estimation (PDE) problem aims at finding the periodic demand minimizing the tactical costs, and is important for freight carriers. The periodic demand is the demand expected to repeat in cyclic and deterministic formulations of service network design problems. It is defined as a mapping of the demand forecasts over the tactical horizon. The PDE problem has been introduced as a multilevel mathematical programming formulation yet an efficient solution method has not been introduced in the literature and we aim at addressing this gap. We present a new formulation of the problem, where the periodic demand is defined as a deviation from the average of the demand forecasts, and the variables are the deviation coefficients for all commodities transported in the network. We develop two local search metaheuristics to solve the PDE problem and compare with NOMAD, an off-the-shelf blackbox optimization software that performs best for problems with few variables. They all exploit the sequential property, that is, when the first-level variables are fixed, the lower levels of PDE can be solved sequentially. Large-scale applications widespread in practice carry hundreds of commodities, and the three algorithms might be challenged by the large number of variables. To address this issue, we propose heuristic approaches which reduce the size of the feasible set while keeping high-quality solutions. It consists in creating clusters of commodities that have an equal deviation coefficient, hence reducing the number of variables. We report results in an extensive empirical study of a real large-scale application from the Canadian National Railway Company. Two main findings emerge. First, the solutions obtained outperform the approach commonly used in practice which simply consists in using the mean of the demand forecasts. Second, the clustering heuristic allows to obtain the best solution and enables the use of off-the-shelf softwares even for large-scale applications.

Key words Freight transportation, tactical planning, large-scale, periodic demand, heuristic, clustering.

5.1 Introduction

Service Network Design (SND) problems are an important class of planning problems for freight carriers and they aim at designing a plan satisfying demand in a cost-effective way. Deterministic SND formulations are mostly used for real large-scale applications for the sake of computational tractability. Such formulations rely on an accurate representation of *periodic demand*, that is the demand expected to repeat at each period (e.g. a week) of the tactical planning horizon (e.g. a few months). A tactical plan minimizing costs while satisfying the periodic demand is then defined over the period and repeated over the planning horizon. However, time series forecasting models produce one point estimate of demand at each period.

Our previous work (Laage et al., 2021b) introduces a methodology for the Periodic Demand Estimation problem (PDE). It presents a multilevel formulation whose solution is the periodic demand minimizing the tactical costs over the horizon. The periodic demand is defined as a mapping from the demand forecasts estimated for each transported commodity at each period. The formulation is composed of three levels, however, when the periodic demand is fixed and when the second level variables are feasible for the third level, the lower levels can be solved sequentially. The PDE problem can be formulated as an optimization problem $\min_{s \in S} f(s, g^*(s))$, where s is the first level variables, i.e. the periodic demand, and S its feasible set, f the tactical costs and g^* the optimal solution of the lower-level program. Laage et al. (2021b) evaluate the methodology by exploring a restricted feasible set of four periodic demands and find the solution by enumeration. An effective solution approach for the PDE problem with any periodic demand has not yet been studied in the literature, and our work addresses this gap.

The PDE optimization problem is challenging for several reasons. Lower levels are combinatorial problems that are non-convex and not differentiable. Moreover, it aims at addressing large-scale applications. Hence metaheuristics and blackbox algorithms are promising solution methods that we consider in this work.

Metaheuristics are essentially divided in two types: local search methods iterating over one solution and population methods working with a population of solutions (Talbi, 2009). Solving the lower levels might be computationally expensive, hence local search methods are more adapted for our problem. For the same reason, blackbox optimization methods are also related to our work (Audet and Hare, 2017). The black box is the lower-level program whose solution is used to compute the objective function. Those approaches are, however, limited when the number of variables is large, which is the case for large-scale freight networks.

By using a mapping of the forecasts in the first-level, PDE aims at integrating forecasting and planning problems. The general idea of integrating prediction and optimization has been explored in the literature, yet to the best of our knowledge, we are the first to consider the notion of defining a good periodic demand for real-life applications of SND problems.

Contributions. In this research, we extend the work from Laage et al. (2021b) by allowing a broad and continuous feasible set of periodic demands. The paper offers both methodological and empirical contributions. First, we formalize the mapping from the demand forecasts to the periodic demand as a deviation from the average of the forecasts. The first-level variables are then the deviation coefficient for each commodity. We develop two local search metaheuristics to solve the new formulation of the PDE problem. To address the challenge of limited performances due to a large number of first-level variables, we propose heuristic approaches that reduce the set of feasible solutions by creating clusters of commodities having equal deviation coefficients. This in turn allows to reduce the number of variables. The clustering heuristics exploit the characteristics of the problem considered, defined with new metrics. We report results in an extensive empirical study of a real large-scale application from the Canadian National Railway Company. Finding a good periodic demand by solving the PDE problem lead to substantial cost reductions compared to the common practice, which consists in taking the average of the demand forecasts. We show that the clustering step is crucial for capacitated networks, as it allows to reach the best costs. Moreover, it leverages off-the-shelf solvers even for large-scale networks.

Paper Organization. The remainder paper is structured as follows. Next we present the related work to this research: we briefly describe the PDE problem and the solution approaches. In Section 5.3, we introduce our formulation, the metaheuristics and the heuristic developed to solve PDE. Then, we outline our large-scale application in Section 5.4. Finally, we report empirical results in Section 5.5 and conclude with directions for future research in Section 5.6.

5.2 Related Works

We start by briefly summarizing the context and describe the PDE formulation introduced in Laage et al. (2021b). Then, we provide a review of the solution approaches developed in the literature.

5.2.1 The Periodic Demand Estimation Problem

The PDE formulation translates the following planning process into a multilevel optimization problem. First, a time series forecasting model produces the demand forecasts for each commodity at each week of the tactical planning horizon. Second, a periodic demand is estimated for each commodity over the tactical planning horizon, as a mapping from the forecasts. Finally, design and flow decisions follow. While the former are fixed over the tactical horizon, the latter are adjusted according to the observed demand. Hence two sources of costs occur at the tactical level: the tactical plan yield fixed costs and the flow and outsourcing adjustments are responsible for the variable costs. The objective of the PDE problem is to minimize the tactical costs over the tactical planning horizon by finding a good estimate of the periodic demand.

The tactical planning horizon is decomposed into T periods indexed by $t = 1, \ldots, T$. The network carries a set of commodities $\mathcal{K}$, and each commodity k is characterized by its origin o_k, destination d_k and type γ_k. We denote K the cardinality of $\mathcal{K}$. The demand vector of period t is designated by $\mathbf{y}_t$, such that $\mathbf{y}_t = (y_{t1}, \ldots, y_{tK})^\top$, where y_{tk} is the quantity of commodity k to be transported during period t. Let $\mathbf{Y} \in \mathbb{R}_+^{T \times K}$ be the demand matrix, with $[\mathbf{Y}]_{tk} = y_{tk}$. The periodic demand vector is designated by $\mathbf{y}^{\mathrm{p}}$, such that $\mathbf{y}^{\mathrm{p}} = (y_1^{\mathrm{p}}, \ldots, y_K^{\mathrm{p}})^\top$ where y_k^{p} is the periodic demand for commodity k. Let us note $\mathcal{Y}$ the set of feasible values for $\mathbf{y}^{\mathrm{p}}$. In practice, the demand is not known in advance and the demand values are forecasts.

We consider a class of SND problems, the Multicommodity Capacitated Fixed-charge Network Design (MCND) problems (Magnanti and Wong, 1984) for illustration purposes. The generic formulation of a path-based MCND relies on a space-time graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is the set of nodes and $\mathcal{A}$ is the set of arcs. A path p is a sequence of arcs in $\mathcal{G}$, and $\mathcal{P}$ denote the set of paths. Let $\mathcal{P}_k$ denote the set of paths for commodity k such that the source node of the first arc of $p \in \mathcal{P}_k$ is o_k and the sink node of the last arc is d_k. We denote $\mathcal{K}_p$ the set of commodities that can use path $p \in \mathcal{P}$. Let $\mathcal{P}^{\mathrm{out}}$ designate the paths corresponding to outsourcing options that transport demand in the case of insufficient capacity, such that $\mathcal{P}^{\mathrm{out}} \subset \mathcal{P}$. The set $\mathcal{P}_k^{\mathrm{out}}$ designates the outsourcing paths for commodity k, such that $\mathcal{P}_k^{\mathrm{out}} \subset \mathcal{P}_k$. We present below the formulation **PDE**, where **MCND** and **wMCND** constitute the lower levels. The upper level aims at minimizing the total fixed and variable costs over the tactical planning horizon, and $\mathbf{y}^{\mathrm{p}}$ is the decision variable. The objective of **MCND** is to satisfy the periodic demand at minimum cost. It has two categories of decision variables: Binary design variables z_p, $\forall p \in \mathcal{P}$, equal to 1 if path p is used and 0 otherwise, and flow variables $x_{pk} \geq 0, p \in \mathcal{P}_k$, $\forall k \in \mathcal{K}$. Depending on the type of freight, the flow variables can be integers. Finally the third level **wMCND** aims at satisfying demand at

each period of the horizon at minimum cost for a fixed design solution z given by **MCND**. The variables x_{tpk} designate the flow for commodity k on path p in period t.

$$\textbf{PDE} \quad \min_{\mathbf{y}^{\mathrm{p}}} \quad C^{\mathrm{PDE}} = \sum_{t=1}^{T} \left[\sum_{p \in \mathcal{P}} C_p^{\mathrm{design}} z_p + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \setminus \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{flow}} x_{tpk} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{out}} x_{tpk} \right] \tag{5.1}$$

$$\text{s.t.} \quad \mathbf{y}^{\mathrm{p}} \in \mathcal{Y}, \tag{5.2}$$

$$\textbf{MCND} \quad \min_{z,x} \sum_{p \in \mathcal{P}} C_p^{\mathrm{design}} z_p + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \setminus \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{flow}} x_{pk} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{out}} x_{pk} \tag{5.3}$$

$$\text{s.t.} \sum_{p \in \mathcal{P}_k} x_{pk} = y_k^{\mathrm{p}}, \qquad k \in \mathcal{K}, \tag{5.4}$$

$$\sum_{k \in \mathcal{K}_p} x_{pk} \leq u_p z_p, \qquad p \in \mathcal{P}, \tag{5.5}$$

$$x_{pk} \geq 0, \qquad k \in \mathcal{K}, p \in \mathcal{P}_k, \tag{5.6}$$

$$z_p \in \{0,1\}, \qquad p \in \mathcal{P}, \tag{5.7}$$

$$\textbf{wMCND} \quad \min_{x_1,\ldots,x_T} \sum_{t=1}^{T} \left[\sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \setminus \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{flow}} x_{tpk} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k^{\mathrm{out}}} C_p^{\mathrm{out}} x_{tpk} \right] \tag{5.8}$$

$$\text{s.t.} \sum_{p \in \mathcal{P}_k} x_{tpk} = y_{tk}, \qquad t = 1,\ldots,T, k \in \mathcal{K}, \tag{5.9}$$

$$\sum_{k \in \mathcal{K}_p} x_{tpk} \leq u_p z_p, \qquad t = 1,\ldots,T, p \in \mathcal{P}, \tag{5.10}$$

$$x_{tpk} \geq 0, \qquad t = 1,\ldots,T, k \in \mathcal{K}, p \in \mathcal{P}_k. \tag{5.11}$$

The design and flow variables from the lower levels **MCND** and **wMCND** are used in the definition of the tactical cost in the objective function (5.1) of **PDE**.

The objective function (5.3) of **MCND** includes three terms. The first term designates the fixed design cost $C_p^{\mathrm{design}} \geq 0$ and account for the paths built to transport demand. The second term is the variable flow cost $C_p^{\mathrm{flow}} \geq 0$ for satisfied demand and the third term with $C_p^{\mathrm{out}} \geq 0$ designates the flow cost of outsourced demand. The objective function (5.8) contains the last two terms, that are the variable flow cost and the outsourcing cost at each period.

Constraints (5.4) and (5.9) ensure that respectively the periodic demand and the demand at each period are satisfied for each commodity. Constraints (5.5) and (5.10) enforce flows on selected paths only, and that respectively the flow from the periodic demand and the flow at each period do not exceed the path capacity u_p.

An important property of **PDE** that we call the *sequential property* is that for (z^*, x^*) an optimal solution of **MCND**, if z^* is feasible for **wMCND**, then **MCND-wMCND** can be

solved sequentially for a fixed $\mathbf{y}^{\mathrm{p}}$. By always allowing outsourcing, i.e., either fixing $z_p = 1$ for $p \in \mathcal{P}^{\mathrm{out}}$ or $C_p^{\mathrm{design}} = 0$, the property holds.

The periodic demand is defined as a mapping of the demand values at each week of the tactical horizon. Then the set of feasible periodic demand $\mathcal{Y}$ is a set of feasible mappings h, where h is defined as following:

$$\begin{aligned} h\colon \quad & \mathbb{R}_+^{T\times K} \to \mathbb{R}_+^K \\ & \mathbf{Y} \mapsto \mathbf{y}^{\mathrm{p}} = h(\mathbf{Y}). \end{aligned} \tag{5.12}$$

5.2.2 Solution Approaches

The formulation **PDE** (5.1)-(5.11) appears as a multilevel optimization problem, yet it does not belong to this class of problems due to the sequential property when outsourcing is allowed. It is nonetheless challenging, because lower levels are non-convex, not differentiable and combinatorial, and it aims at addressing large-scale problems.

The PDE problem is then defined as a generic optimization problem with no derivative information. Local search methods are attractive for this type of problems. They build a trajectory in the space of solutions trying to move towards optimal solutions (Talbi, 2009). The first-level constraints of **PDE** can be integrated in the set of solutions to visit and the lower levels are solved sequentially. Their solution is used to compute C^{PDE} (5.1), and the latter will give information to continue the search. Local search methods is an important and mature research topic and many metaheuristics have been proposed in the literature: the Tabu Search (Glover, 1986), the Simulated Annealing (Kirkpatrick et al., 1983; Černỳ, 1985), the Variable Neighborhood Search (Mladenović and Hansen, 1997), the Greedy Randomized Adaptive Search Procedure (Feo and Resende, 1989, 1995), to name a few. They differ by the definition of the solutions visited at each iteration, the use of memory during the search and the diversification and intensification procedures to explore the set of solutions.

Blackbox optimization methods are also related to our work (Audet and Hare, 2017). The black box is the lower level program whose solution is used to compute the objective function. We take a particular interest in NOMAD (Le Digabel, 2011), a software implementing the Nonsmooth Optimization by Mesh Adaptive Direct Search (Abramson et al., 2004, 2009). Here, the black box problem is **MCND-wMCND**, when the first-level decision variables are fixed. However, NOMAD performs best for problems with fewer than 50 variables (Le Digabel, 2011) and large-scale applications might contain hundreds of variables.

5.3 Methodology

In this section, we present our methodological contributions. In Section 5.3.1 we introduce the new formulation of **PDE**, where the periodic demand is defined as a deviation from the average of the demand forecasts. In Section 5.3.2, we describe the local search metaheuristics developed to solve the new formulation. Finally, in Section 5.3.3, we present our clustering heuristic that reduces the size of the set of feasible solutions of **PDE** $\mathcal{Y}$.

5.3.1 Model

We write **PDE** as following to simplify the notation:

$$\textbf{PDE} \quad \min_{\mathbf{y}^{\mathrm{p}}} \quad C^{\mathrm{PDE}}(\mathbf{y}^{\mathrm{p}}, \mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) \tag{5.13}$$

$$\text{s.t.} \quad \mathbf{y}^{\mathrm{p}} \in \mathcal{Y}, \tag{5.14}$$

$$(\mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) \in \underset{\mathbf{z'}, \mathbf{x'}, \mathbf{x_1}', \ldots, \mathbf{x_T}'}{\operatorname{argmin}} \textbf{MCND-wMCND}(\mathbf{y}^{\mathrm{p}}, \mathbf{z'}, \mathbf{x'}, \mathbf{x_1}', \ldots, \mathbf{x_T}') \tag{5.15}$$

Constraint (5.14) defines the feasible space where $\mathcal{Y}$ is a set of mappings h defined in (5.12), and solving **PDE** consists in finding the mapping h such that C^{PDE} (5.13) is minimized. Common practices take h as the average of the demand values at each week t for each commodity:

$$\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{y}_t. \tag{5.16}$$

We propose an extension of the common practice where the periodic demand is defined as a deviation from $\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$:

$$\mathbf{y}^{\mathrm{p}} = \boldsymbol{\alpha} \odot \mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}, \tag{5.17}$$

where $\odot$ designates the element-wise multiplication. The vector $\boldsymbol{\alpha}$ is defined such that, for each commodity k, $y^{\mathrm{p}}_k = [\boldsymbol{\alpha}]_k y^{\mathrm{p}}_{\mathrm{mean},k}$.

We present our new formulation:

$$\textbf{PDE} \quad \min_{\boldsymbol{\alpha}} \quad C^{\mathrm{PDE}}(\mathbf{y}^{\mathrm{p}}, \mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) \tag{5.18}$$

$$\text{s.t.} \quad \mathbf{y}^{\mathrm{p}} = \boldsymbol{\alpha} \odot \mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}, \tag{5.19}$$

$$\boldsymbol{\alpha} \leq \boldsymbol{\alpha}_{\max} \tag{5.20}$$

$$\boldsymbol{\alpha} \geq \boldsymbol{\alpha}_{\min} \tag{5.21}$$

$$(\mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) \in \underset{\mathbf{z'}, \mathbf{x'}, \mathbf{x_1}', \ldots, \mathbf{x_T}'}{\operatorname{argmin}} \textbf{MCND-wMCND}(\mathbf{y}^{\mathrm{p}}, \mathbf{z'}, \mathbf{x'}, \mathbf{x_1}', \ldots, \mathbf{x_T}') \tag{5.22}$$

Constraint (5.20) and Constraint (5.21) respectively impose an upper bound and a lower bound on the variable $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}_{\max}$ is defined such that $\mathbf{y}^{\mathrm{p}}_{\max} = \boldsymbol{\alpha}_{\max} \odot \mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$, with $\mathbf{y}^{\mathrm{p}}_{\max} = \max_{t=1,\ldots,T}\{\mathbf{y}_t\}$ and $\boldsymbol{\alpha}_{\min}$ is defined such that $\mathbf{y}^{\mathrm{p}}_{\min} = \boldsymbol{\alpha}_{\min} \odot \mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}$, with $\mathbf{y}^{\mathrm{p}}_{\min} = \min_{t=1,\ldots,T}\{\mathbf{y}_t\}$.

We aim at solving the formulation **PDE** (5.18)-(5.22). Let us illustrate its importance and the reduction of tactical costs it can lead to with the following small problem. The network transports one commodity, from its origin O_1 to its destination D_1 over a tactical horizon of 3 periods. Figure 5.1 presents the three different possible paths for the commodity, along with their capacity u and cost c. The path 3 corresponds to the outsourcing path: it is always built in the design, not constrained by capacity, but expensive.

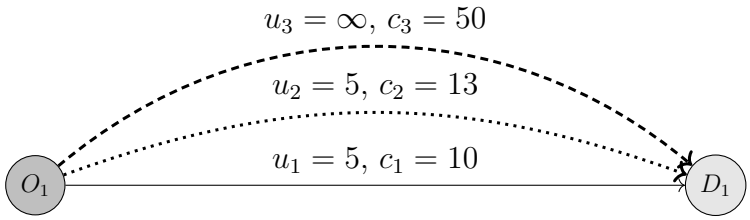


Figure 5.1 Illustration of the PDE problem on a small network

The time-series demand forecasts $\hat{\mathbf{Y}}$ and the observed demand $\mathbf{Y}$ of the commodity at the three periods are

$$\hat{\mathbf{Y}} = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 4 \\ 6 \\ 7 \end{pmatrix}. \tag{5.23}$$

Common practice is to take the average of the time-series forecasts as periodic demand, that is, $\hat{y}^{\mathrm{p}}_{\mathrm{mean}} = 5$. In this case, path 1 is built to minimize costs while satisfying the periodic demand. The cost of the fixed plan is $c_1 \cdot 5 = 50$. At each week, the maximum offered capacity is 5 on path 1, and the exceeding demand is allocated to path 3. The tactical cost estimated

with the forecasts $\hat{\mathbf{Y}}$ are $c_1 \cdot (3+5+5) + c_3 \cdot (0+2+0) = 230$ and the actual tactical costs, estimated after the demand realizations $\mathbf{Y}$ are $c_1 \cdot (4+5+5) + c_3 \cdot (0+1+2) = 290$. If the periodic demand were estimated as a deviation from the average of the forecasts, with for instance $\alpha = 1.2$ and $\hat{y}^{\mathrm{p}} = \alpha \hat{y}^{\mathrm{p}}_{\mathrm{mean}} = 6$, then paths 1 and 2 are built. The cost of the fixed plan is then $c_1 \cdot 5 + c_2 \cdot 1 = 63$. The tactical costs from the forecasts are $c_1 \cdot (3+5+5) + c_2 \cdot (0+2+0) = 156$ and $c_1 \cdot (3+5+5) + c_2 \cdot (0+1+2) = 179$ after the demand realizations.

5.3.2 Metaheuristics

The vector $\boldsymbol{\alpha}$ of size K describes the first-level decision variables in the formulation **PDE** (5.18)-(5.22). Local search methods are perfectly fitted to solve **PDE**: the first-level constraints can be integrated in the set of solutions to visit and we can exploit the sequential property of **PDE** at each iteration, when a new vector $\boldsymbol{\alpha}$ is selected. We fix the upper-level variable to $\boldsymbol{\alpha}$, then solve **MCND-wMCND** sequentially and finally use the solution to compute C^{PDE}, which value indicates where to continue the search. We present next the two local search metaheuristics we develop to solve **PDE**.

Both require a feasible set of vectors that can be visited, a neighborhood N indicating the movements allowed in the feasible set and a stopping criterion. We designate by $C^{\mathrm{PDE}}(\boldsymbol{\alpha})$ the value of the objective function (5.18) for a fixed $\boldsymbol{\alpha}$. We present below their pseudo-code and parameters.

Neighborhood Search

We name Neighborhood Search (NS) the first metaheuristic, which is a simple local search method. It iterates over a current solution, explores the neighborhood around it at each iteration and stops when the value of C^{PDE} is not improved. We present in Algorithm 1 the pseudo-code for NS and in Algorithm 2 the pseudo-code for the neighborhood definition.

Algorithm 1 Neighborhood Search (NS)

1: **Input**: Initial solution $\boldsymbol{\alpha}$
2: Stop = False
3: **while** not Stop **do**
4: Define $N(\boldsymbol{\alpha})$ the neighborhood around the solution $\boldsymbol{\alpha}$
5: Find the solution $\boldsymbol{\alpha'}$ minimizing C^{PDE} in $N(\boldsymbol{\alpha})$
6: **If** $C^{\text{PDE}}(\boldsymbol{\alpha'}) < C^{\text{PDE}}(\boldsymbol{\alpha})$
7: **Update** $\boldsymbol{\alpha} := \boldsymbol{\alpha'}$
8: **Else** Stop := True

The elements of the neighborhood $N(\boldsymbol{\alpha})$ for a solution $\boldsymbol{\alpha}$ are randomly generated following a normal distribution. We note V the number of neighbors, i.e., $V = |N(\boldsymbol{\alpha})|$, β a distance parameter and $\mathbf{I}_K$ the identity matrix of size K. Step 5 in Algorithm 2 ensures that the potential solution vectors satisfy Constraints (5.20) and (5.21).

Algorithm 2 Neighborhood Definition

1: **Input**: $\boldsymbol{\alpha}$
2: **Parameters**: β, V
3: **Initialization** $N(\boldsymbol{\alpha}) = \emptyset$
4: **while** $|N(\boldsymbol{\alpha})| < V$ **do**
5: Generate random $\boldsymbol{\alpha'}$ from the normal distribution $\mathcal{N}(\boldsymbol{\alpha}, \beta\mathbf{I}_K)$
6: $\boldsymbol{\alpha'} := \max(\boldsymbol{\alpha'}, \boldsymbol{\alpha}_{\min})$, $\boldsymbol{\alpha'} := \min(\boldsymbol{\alpha'}, \boldsymbol{\alpha}_{\max})$
7: **Update** $N(\boldsymbol{\alpha}) := N(\boldsymbol{\alpha}) \cup \{\boldsymbol{\alpha'}\}$

Neighborhood Search with Diversification and Intensification

The well-known limitation of simple algorithms such as NS is their potential to get stuck at the first local minimum. To address this challenge, we use a metaheuristic allowing at each iteration to move to a neighbor that might not improve the best found objective function. We call it the Neighborhood Search with Diversification and Intensification (NSDI). The neighborhood N is defined as in NS, described in Algorithm 2. We add a diversification and intensification procedure, which consists in updating the parameters β and V at each iteration whether a better solution was found or not. This metaheuristic, presented next

in Algorithm 3, stops when a maximum number of iterations without improvements M is reached.

Algorithm 3 Neighborhood Search with Diversification and Intensification (NSDI)

1: **Input**: Initial solution $\boldsymbol{\alpha}$ and initial best known solution $\boldsymbol{\alpha}^*$
2: **Parameters** β, V, M, v^+, b^-, b^+
3: Stop = 0
4: **while** Stop $< M$ **do**
5: Define $N(\boldsymbol{\alpha})$ the neighborhood around the solution $\boldsymbol{\alpha}$
6: Find the solution $\boldsymbol{\alpha'}$ which minimizes C^{PDE} in $N(\boldsymbol{\alpha})$
7: **If** $C^{\text{PDE}}(\boldsymbol{\alpha'}) < C^{\text{PDE}}(\boldsymbol{\alpha}^*)$
8: **Update** $\boldsymbol{\alpha}^* := \boldsymbol{\alpha'}$
9: **Update** Stop:= 0
10: **Intensification** $\beta := b^-\beta$
11: **Else**
12: **Update** Stop := Stop + 1
13: **Diversification** $\beta := b^+\beta$, $V := v^+V$
14: **Update** $\boldsymbol{\alpha} := \boldsymbol{\alpha'}$

In Algorithm 3, at each iteration, the current solution is updated by the best solution in its neighborhood, even if it does not improve the current value of objective function. Step 10 and 13 consists in the diversification and intensification procedure. When the best found solution is improved, we intensify the search in this direction: $b^- < 1$ and the neighbors generated are closer. On the contrary, when the best found solution is not improved, we look for new solutions that are further, with $v^+ > 1$ and $b^+ > 1$. More neighbors that are further are generated.

By definition, NSDI might be stuck between 2 solutions, going back and forth from one to the other. The Tabu Search (Glover, 1986) avoid this problem by defining a tabu list, which contains solutions that cannot be visited for a few iterations. Here, the randomness in the definition of N ensures that this is unlikely to happen, and we do not need to define a tabu list. For the same current solution, at different iterations, the set N will likely be different and the algorithm will move to a different neighbor.

Both metaheuristics NS and NSDI require an initial solution. Since $\boldsymbol{\alpha}$ represents the deviation from the mean, we take as initial solution $\boldsymbol{\alpha} = \mathbf{1}_K$, the K-vector of ones.

5.3.3 Clustering to Reduce the Set of Feasible Mappings

For large-scale applications with multiple commodities, the number of first-level decision variables, i.e., the dimension of the vector $\boldsymbol{\alpha}$, K, might be large. In our application, $K = 170$. To address this challenge, we restrict the set of feasible mappings h, which can also be seen as imposing constraints on $\boldsymbol{\alpha}$. One simple example is to constrain all variables to be equal. This is equivalent to having a single decision variable α and the following formulation:

$$\begin{aligned}
\textbf{PDE} \quad \min_{\boldsymbol{\alpha}} \quad & C^{\mathrm{PDE}}(\mathbf{y}^{\mathrm{p}}, \mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) && (5.24)\\
\text{s.t.} \quad & \mathbf{y}^{\mathrm{p}} = \alpha \mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}, && (5.25)\\
& \alpha \leq \max_k \boldsymbol{\alpha}_{\mathrm{max},k} && (5.26)\\
& \alpha \geq \min_k \boldsymbol{\alpha}_{\mathrm{min},k} && (5.27)\\
& (\mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) \in \operatorname*{argmin}_{\mathbf{z}', \mathbf{x}', \mathbf{x_1}', \ldots, \mathbf{x_T}'} \textbf{MCND-wMCND}(\mathbf{y}^{\mathrm{p}}, \mathbf{z}', \mathbf{x}', \mathbf{x_1}', \ldots, \mathbf{x_T}'). && (5.28)
\end{aligned}$$

Before solving **PDE** with the metaheuristics proposed in Section 5.3.2, we propose a first step which consists in creating clusters of commodities that have the same deviation coefficients. In other words, clusters of components of $\boldsymbol{\alpha}$ which have the same value. The objective of clustering is to reduce the number of variables while keeping high-quality solutions, to improve the performance of the metaheuristics introduced in Section 5.3.2 and to make the use of the state-of-the-art possible for large-scale applications with hundreds of commodities.

We denote $\mathcal{C} = \{C_1, \ldots, C_{n_C}\}$ the set of n_C clusters. It is a partition of $\mathcal{K}$ such that for $C_i \in \mathcal{C}$ with $C_i = \{k_1, .., k_c\}$, the coefficients $\alpha_{k'}$ for the commodities $k' \in C_i$ are equal. We present below the formulation **PDE** with clusters:

$$\begin{aligned}
\textbf{PDE} \quad \min_{\boldsymbol{\alpha}} \quad & C^{\mathrm{PDE}}(\mathbf{y}^{\mathrm{p}}, \mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) && (5.29)\\
\text{s.t.} \quad & \mathbf{y}^{\mathrm{p}} = \boldsymbol{\alpha} \odot \mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}, && (5.30)\\
& \boldsymbol{\alpha} \leq \boldsymbol{\alpha}_{\max} && (5.31)\\
& \boldsymbol{\alpha} \geq \boldsymbol{\alpha}_{\min} && (5.32)\\
& [\boldsymbol{\alpha}]_{k_i} = [\boldsymbol{\alpha}]_{k_j}, \qquad C \in \mathcal{C}, \forall k_i, k_j \in C && (5.33)\\
& (\mathbf{z}, \mathbf{x}, \mathbf{x_1}, \ldots, \mathbf{x_T}) \in \operatorname*{argmin}_{\mathbf{z}', \mathbf{x}', \mathbf{x_1}', \ldots, \mathbf{x_T}'} \textbf{MCND-wMCND}(\mathbf{y}^{\mathrm{p}}, \mathbf{z}', \mathbf{x}', \mathbf{x_1}', \ldots, \mathbf{x_T}'). && (5.34)
\end{aligned}$$

The set of clusters should take into account the application considered. For instance, an

increase of the tactical costs can be due to the outsourcing of commodities that could not be loaded. This happens when the periodic demand used to define the plan is low and the network has not enough capacity allocated for them. One of the cluster could gather the commodities that have high risks of being outsourced, so that **PDE** assigns them a large deviation coefficient. In the following, we describe two approaches to define the set of clusters $\mathcal{C}$ that use characteristics from the network and the commodities.

Variance-based Clustering

If the demand forecasts for a commodity have a large variance over the planning horizon, then the periodic demand from the average forecasts is low at certain weeks compared to the demand forecasts, resulting in turn in outsourcing.

We propose a first clustering based on the coefficient of variation, i.e., the standard deviation scaled to the average demand values. We denote σ^{s}_k the coefficient of variation for a commodity k, such that

$$\sigma^{\mathrm{s}}_k = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_{tk} - y^{\mathrm{p}}_{\text{mean, k}})^2}}{y^{\mathrm{p}}_{\text{mean, k}}}, \quad k = 1, \ldots, K. \tag{5.35}$$

When K is small, we can create the clusters from analyzing the distribution of σ^{s} over the commodities. More generally, the n_C clusters are created by decomposing the set $\{\sigma^{\mathrm{s}}_k, k = 1, \ldots, K\}$ into n_C intervals. Let Q_e designates the e-th percentile of $\{\sigma^{\mathrm{s}}_k, k = 1, \ldots, K\}$. We could create for instance $n_C = 2$ clusters, where the first cluster contains all $\sigma^{\mathrm{s}}_k \leq Q_{0.25}$ and the second cluster contains all $\sigma^{\mathrm{s}}_k > Q_{0.25}$. In our application, we consider $n_C = 5$ clusters and describe them in Section 5.5.

Resource-based Clustering

We propose a second approach built from the analysis of the shared resources in the network. Increasing the periodic demand for commodities that share resources lead to an increased dedicated capacity in the design. This in turn might help to handle the sudden changes in forecasted and observed demand, when they are higher than their respective average over the periods of the planning horizon. We propose the following steps to identify the set of clusters.

1. We start from the common practice where the periodic demand is the average of the demand forecasts, i.e., we compute $C^{\mathrm{PDE}}(\alpha_{\mathrm{mean}})$, where $\alpha_{\mathrm{mean}} = 1$.

2. For each commodity k, we identify the commodities that share at least one train on their route with k. We call a *group* the set formed by k and the related commodities.

3. The first cluster is the largest group over all commodities.

4. The second cluster is the second largest group of cardinality higher than 1 that has no commodities in common with the first cluster. We iterate until there is no more group satisfying the criteria.

5. The remaining commodities not yet assigned to a cluster are gathered to form the last cluster.

Unlimited Resource-based Clustering We propose another set of clusters built similarly, except for the first step. Instead of computing $C^{\text{PDE}}(\alpha_{\text{mean}})$ from the formulation (5.36)-(5.56), we relax the capacity constraints (5.45) and (5.54) and solve this new formulation with α_{mean}. This aims to identify the best paths to transport each commodity, and analyze the bottlenecks in the network. We then follow the steps 2 to 5 to obtain the clusters.

5.4 Application

We proceed with the application from Laage et al. (2021b) of the intermodal network of Canadian National (CN). It is composed of 24 main intermodal terminals and 133 origin-destination pairs. A commodity is defined by an origin, a destination and a type of container. There are two main types of containers, the 40-feet and the 53-feet long, resulting in $K = 170$ commodities. When two commodities have the same OD pair, they differ by the type of container. The tactical period is a week and a tactical horizon lasts $T = 10$ weeks.

CN faces a specific MCND problem for its intermodal network, referred to as the *block planning problem*. A block designates a consolidation of railcars that move together between a given OD pair, where containers loaded on the railcars share the same OD. Morganti et al. (2020) introduce a path-based formulation, **BP**, for this problem where the periodic demand and the schedule of intermodal trains are given as inputs. **BP** is based on a space-time graph which contains 28,854 arcs and 15,269 nodes. A block is a path in the graph and $\mathcal{B}$ designates the set of blocks such that $|\mathcal{B}| = 2208$. The set $\mathcal{B}$ hence corresponds to the set of paths $\mathcal{P}$ in **MCND**. It contains a subset of outsourcing paths denoted $\mathcal{B}^{\text{artif}}$, also referred to as *artificial blocks* whose role is to transport demand exceeding capacity.

Below we briefly describe **PDE** and refer to Morganti et al. (2020) and Laage et al. (2021b) for more details. The lower level formulations are replaced by **BP** and its weekly formulation,

wBP. There are three categories of decision variables at the second level. First, the design variables $z_b, b \in \mathcal{B}$ where $z_b = 1$ if block b is built. Second, x_{bk} designates the integer flow variable, that is, the number of containers for commodity k transported on block b. Third, to write the capacity constraints we need the auxiliary variables for the number of 40-feet v_b^{40} and 53-feet v_b^{53} double-stack platforms on block b. At the third level, we introduce flow variables and auxiliary platform variables for each week t, $x_{tbk}, v_{tb}^{40}, v_{tb}^{53}, t \in \mathcal{T}, k \in \mathcal{K}, b \in \mathcal{B}$.

$$\textbf{PDE}\ \min_{\boldsymbol{\alpha}} C^{\mathrm{PDE}} = \sum_{t=1}^{T}\left[\sum_{b\in\mathcal{B}\setminus\mathcal{B}^{\mathrm{artif}}} C_b^{\mathrm{design}} z_b + \sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}_k\setminus\mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{flow}} x_{tbk} + \sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{out}} x_{tbk}\right] \tag{5.36}$$

$$\text{s.t. } \mathbf{y}^{\mathrm{p}} = \boldsymbol{\alpha}\odot\mathbf{y}^{\mathrm{p}}_{\mathrm{mean}}, \tag{5.37}$$

$$\boldsymbol{\alpha} \le \boldsymbol{\alpha}_{\max} \tag{5.38}$$

$$\boldsymbol{\alpha} \ge \boldsymbol{\alpha}_{\min} \tag{5.39}$$

$$\textbf{BP}\ \min_{x,z} \sum_{b\in\mathcal{B}\setminus\mathcal{B}^{\mathrm{artif}}} C_b^{\mathrm{design}} z_b + \sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}_k\setminus\mathcal{B}_k^{\mathrm{artif}}} C_{bk}^{\mathrm{flow}} x_{bk} + \sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}_k^{\mathrm{artif}}} C_{bk}^{\mathrm{out}} x_{bk} \tag{5.40}$$

$$\text{s.t. } \sum_{b\in\mathcal{B}_k} x_{bk} = y_k^{\mathrm{p}}, \qquad k\in\mathcal{K}, \tag{5.41}$$

$$x_{bk} \le y_k^{\mathrm{p}} z_b, \qquad k\in\mathcal{K}, b\in\mathcal{B}_k, \tag{5.42}$$

$$v_b^{53} = \max\left[0, \left\lceil \frac{1}{2}\left(\sum_{k\in\mathcal{K}_b,\tau_k=53} x_{bk} - \sum_{k\in\mathcal{K}_b,\tau_k=40} x_{bk}\right)\right\rceil\right], \qquad b\in\mathcal{B}, \tag{5.43}$$

$$v_b^{40} = \left\lceil \frac{1}{2}\left(\sum_{k\in\mathcal{K}_b} x_{bk}\right)\right\rceil - v_b^{53}, \qquad b\in\mathcal{B}, \tag{5.44}$$

$$\sum_{b\in\mathcal{B}_a}\left(L^{40}v_b^{40} + L^{53}v_b^{53}\right) \le u_a, \qquad a\in\mathcal{A}^{TM}, \tag{5.45}$$

$$z_b \in \{0,1\}, \qquad b\in\mathcal{B}, \tag{5.46}$$

$$v_b^{40}, v_b^{53} \in \mathbb{N}, \qquad b\in\mathcal{B}, \tag{5.47}$$

$$x_{bk} \in \mathbb{N}, \qquad k\in\mathcal{K}, b\in\mathcal{B}_k, \tag{5.48}$$

$$\textbf{wBP}\quad \min_{x_1,\ldots,x_T} \sum_{t=1}^{T}\sum_{k\in\mathcal{K}}\left[\sum_{b\in\mathcal{B}_k\setminus\mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{flow}} x_{tbk} + \sum_{b\in\mathcal{B}_k^{\mathrm{artif}}} C_{tbk}^{\mathrm{out}} x_{tbk}\right] \tag{5.49}$$

$$\text{s.t. } \sum_{b\in\mathcal{B}_k} x_{tbk} = y_{tk}, \qquad t\in\mathcal{T}, k\in\mathcal{K}, \tag{5.50}$$

$$x_{tbk} \le y_{tk} z_b, \qquad t\in\mathcal{T}, k\in\mathcal{K}, b\in\mathcal{B}_k, \tag{5.51}$$

$$v_{tb}^{53} = \max\left[0, \left\lceil \frac{1}{2}\left(\sum_{k\in\mathcal{K}_b,\tau_k=53} x_{tbk} - \sum_{k\in\mathcal{K}_b,\tau_k=40} x_{tbk}\right)\right\rceil\right], t\in\mathcal{T}, b\in\mathcal{B}, \tag{5.52}$$

$$v_{tb}^{40} = \left\lceil \frac{1}{2}\left(\sum_{k\in\mathcal{K}_b} x_{tbk}\right)\right\rceil - v_{tb}^{53}, \qquad t\in\mathcal{T}, b\in\mathcal{B}, \tag{5.53}$$

$$\sum_{b\in\mathcal{B}_a}\left(L^{40}v_{tb}^{40} + L^{53}v_{tb}^{53}\right) \le u_a, \qquad t\in\mathcal{T}, a\in\mathcal{A}^{TM}, \tag{5.54}$$

$$v_{tb}^{40}, v_{tb}^{53} \in \mathbb{N}, \qquad t\in\mathcal{T}, b\in\mathcal{B}, \tag{5.55}$$

$$x_{tbk} \in \mathbb{N}, \qquad t\in\mathcal{T}, k\in\mathcal{K}, b\in\mathcal{B}_k. \tag{5.56}$$

Constraints (5.41) and (5.50) enforce that the demand is satisfied by either the network capacity or outsourcing. Constraints (5.42) and (5.51) ensure flows to be on selected blocks only. Constraints (5.43), (5.44), (5.52) and (5.53) fix the number of platforms required to

transport the demand. These constraints model the double-stacking possibilities for the containers of different sizes. The 40-foot platforms are preferred to 53-foot platforms, and the remaining 53-foot containers are stacked on top of 40-feet containers. The platform lengths are denoted L^{40} and L^{53}, respectively. Constraints (5.45) and (5.54) enforce that the train capacity is not exceeded. The latter, denoted $u_a, a \in \mathcal{A}^{TM}$ is expressed in number of feet, and defined for the set of arcs that represent moving trains, $\mathcal{A}^{TM}$. The set $\mathcal{B}_a$ designates the set of blocks that use train moving arc $a \in \mathcal{A}^{TM}$.

The main costs from C^{PDE} come from the flow of commodity on blocks and the outsourcing flows. The cost of building the blocks is relatively low compared to the containers being moved on said blocks. The value of the parameters in the objective function, namely $C^{\text{design}}, C^{\text{flow}}$ and C^{out} was defined after a thorough analysis with CN of the paths produced by **BP** for each commodity.

5.5 Results

In order to validate the approach, we test the performance of the algorithms on instances that differ by the number of commodities, the resource sharing and the capacity constraints. We present below three metrics that allow to quantify the resource sharing and the capacity constraints. We report in Section 5.5.1 the results of the metaheuristics on the different instances without the clustering heuristic, and with the clustering heuristic in Section 5.5.2.

Let us consider an illustration to explain the aforementioned metrics, with four commodities k_1, k_2, k_3, k_4 and two trains A_1, A_2. Commodities k_1 and k_2 take A_1 and A_2 on their route, k_3 takes A_1 and k_4 takes A_2. To quantify the resource sharing, we consider two metrics. We designate by τ the average number of commodities that a train can carry. In the example, $\tau = 3$ since both trains carry a total of 3 commodities. We also introduce κ, a metric which quantifies, for each commodity, the number of other commodities sharing a train with the former. In other words, κ is the average number of commodities with at least a common train, per commodity. In the example, $\kappa = 2.5$ since k_1 and k_2 share the capacity with 3 other commodities and k_3 and k_4 with 2.

To quantify the capacity limits, we analyze the potentially outsourced commodities. We first solve **BP** with the real life instance with $K = 170$ commodities and $\boldsymbol{\alpha} = \mathbf{1}_K$, i.e, the periodic demand in **BP** is the average of the demand forecasts. We partition the set of commodities $\mathcal{K}$ in two subsets: the commodities with completely satisfied demand and the commodities with partial or total outsourced demand. We designate the latter by $\mathcal{K}_L$.

Table 5.1 reports the characteristics of each instance. They have different sizes, and present

various levels of resource sharing and capacity constraints. We also indicate the number of blocks $B = |\mathcal{B}|$ in Table 5.1, that is the number of paths on the space-time graph for each instance. The instance IC corresponds to the real life instance. Even though the other instances seem small relative to IC, there are still fairly difficult. Instances $I1$ and $I3$ are not constrained by capacity, as $\mathcal{K}_L = 0$, and instances $I4$ and IC have large capacity sharing metrics.

Table 5.1 Description of the instances and their characteristics

Instance	$\lvert\mathcal{K}\rvert$	$\lvert\mathcal{K}_L\rvert$	B	τ	κ
$I1$	28	0	328	2	4
$I2$	26	17	487	2	3
$I3$	48	0	501	3	4
$I4$	55	12	991	4	11
IC	170	84	2208	9	22

Solving BP and wBP In the remainder of the paper, we present results where **PDE** is solved either with the metaheuristics NS and NSDI presented in Section 5.3.2, or with the off-the-shelf solver NOMAD. All use the sequential property and **BP** and **wBP** are solved sequentially with the commercial solver CPLEX 12.10.0 for a fixed periodic demand. For our application, we need to solve them almost to optimality (the best feasible integer solution has to be within 0.4% of optimal) to properly orient the search algorithms. This is because the value of C^{PDE} is large and a larger gap for **BP** would yield a non-optimal design. This might in turn lead to large outsourcing costs in **wBP**, and a high value of the objective function, finally yielding a wrong search direction in the algorithms.

Parameters for NS and NSDI The metaheuristics rely on several parameters, namely the number of neighbors V, the distance parameter β, and for NSDI, the diversification and intensification parameters b^-, b^+, v^+ and the maximum number of iterations without improvements M. We tested various sets of parameters for the metaheuristics, however we only report in the following the best results. For all instances, $b^- = 0.7$, $b^+ = 1.3$, $v^+ = 1.1$. For instances $I1$ to $I4$, $V = 15$, $\beta = 0.05$ and $M = 15$. Finally, for instance IC, $V = 10$, $\beta = 0.02$ and $M = 7$.

5.5.1 Results Without Clustering

We define $\boldsymbol{\alpha}_{q3}$ such that $\mathbf{y}^{\text{p}}_{\text{q3}} = \boldsymbol{\alpha}_{\text{q3}} \odot \mathbf{y}^{\text{p}}_{\text{mean}}$, where $y^{\text{p}}_{\text{q3},k} = Q_{0.75}(y_{1k}, \ldots, y_{Tk})$. The vector $\boldsymbol{\alpha}_{q3}$ is the vector of deviation from the third quartile of the demand forecasts to the mean of the

forecasts.

For each instance, we run four experiments, with either K variables, i.e. one α_k per commodity k, or one variable when the components of $\boldsymbol{\alpha}$ are constrained to be equal:

- **PDE** solved with $\boldsymbol{\alpha}$ fixed to either $\boldsymbol{\alpha}_{\max}$ or $\boldsymbol{\alpha}_{q3}$.
- Formulation (5.36)-(5.56) solved with NS, NSDI and NOMAD. We note that for the instance IC, NOMAD cannot be used as the number of variables is too large.
- **PDE** solved with the periodic demand fixed to either $\alpha_{\text{mean}} = 1$, $\alpha_{\max} = \max\{\alpha_{\max,k}, k = 1, \ldots, K\}$ or $\alpha_{q3} = \max\{\alpha_{q3,k}, k = 1, \ldots, K\}$.
- Formulation (5.24)-(5.28) solved with NS, NSDI and NOMAD.

Solving **PDE** with $\boldsymbol{\alpha}$ fixed to either $\boldsymbol{\alpha}_{\max}$, $\boldsymbol{\alpha}_{q3}$ or α_{mean} constitutes the enumeration approach proposed in Laage et al. (2021b). Table 5.2 reports the gap to best known solution for the different experiments and for each instance.

Table 5.2 Gap to best known value

		$I1$	$I2$	$I3$	$I4$	IC
scalar α Formulation (5.24)-(5.28)	α_{mean}	0%	125%	361%	137%	51%
	$\alpha_{\max}$	0%	80%	0%	1010%	885%
	α_{q3}	0%	22%	8%	338%	228%
	NS	0%	0%	8%	14%	16%
	NSDI	0%	0%	0%	3%	0%
	NOMAD	0%	2%	0%	3%	10%
K-vector $\boldsymbol{\alpha}$ Formulation (5.36)-(5.56)	$\boldsymbol{\alpha}_{\max}$	0%	0%	0%	210%	78%
	$\boldsymbol{\alpha}_{\text{q3}}$	0%	42%	16%	27%	21%
	NS	0%	125%	119%	64%	51%
	NSDI	0%	0%	0%	0%	51%
	NOMAD	0%	31%	16%	3%	-

Several findings emerge. Defining the periodic demand as a deviation from the average of the demand values allows to obtain good solutions and can reduce substantially the costs compared to simply taking the mean of the demand values, i.e., $\alpha_k = 1, k \in \mathcal{K}$. Moreover, having a single variable α and hence a restricted feasible set of first-level variables allows to obtain good solutions, even the best found solution except for instance $I4$.

When the capacity sharing metrics are low, for instances $I1$, $I2$ and $I3$, the best objective function value is found with the periodic demand $\mathbf{y}^{\text{p}} = \boldsymbol{\alpha}_{\max}$. In other words, when the network is not constrained, or slightly constrained by capacity, we should use the maximum

of the demand values as periodic demand. For the instance $I1$, C^{PDE} does not vary with the periodic demand. Having one block per commodity is sufficient, so as long as the periodic demand is higher than 0, **BP** will create one block per commodity. If the periodic demand is large, more blocks are built in **BP** but not used in **wBP**. The main part of the tactical costs comes from the flow costs, hence the increase of C^{PDE} due to the higher number of blocks is not significant. The variations of C^{PDE} are not visible in Table 5.2 because of their small magnitude.

The differences in costs resulting from a single common $\alpha_{\max}$ or a vector $\boldsymbol{\alpha}_{\max}$ are due to the limited capacity. When **PDE** is solved with a single α fixed to $\alpha_{\max}$, the periodic demand is overestimated for the commodities where $\alpha_{\max,k} < \alpha_{\max}$. As a result, **BP** builds too many blocks for one commodity, resulting in a lack of capacity for the others, and in turn, outsourcing.

For the real life size instance IC, best performances are reached with the formulation (5.24)-(5.28), i.e., with a single common deviation coefficient α. The metaheuristic NSDI finds this solution. With a vector of $K = 170$ variables, both heuristics do not improve the starting solution because the set of feasible solutions is large and finding a good deviation coefficient for each commodity simultaneously is challenging.

5.5.2 Results With Clustering

In this section, we restrict the analysis to instances $I4$ and IC as they both have high resource sharing and limited capacity metrics. In the other instances, the network has enough capacity and the best solution is obtained with $\alpha_{\max}$. We perform the following clustering described in Section 5.3.3 and report in Table 5.3 the number of clusters for each instance:

- CV: Variance-based clustering
- CR: Resource-based clustering
- CRU: Unlimited resource-based clustering

Table 5.3 Number of clusters created in each clustering step

	$I4$	IC
CV	5	5
CR	4	12
CRU	5	16

For the variance-based clustering, we ran several experiments varying the number of clusters, but report only the best results found with $n_C = 5$. The five clusters are following:

- $C_1 = \{k \in \mathcal{K}, \sigma^{\mathrm{s}}_{\min} \leq \sigma^{\mathrm{s}}_k \leq Q_{0.25}\}$
- $C_2 = \{k \in \mathcal{K}, Q_{0.25} < \sigma^{\mathrm{s}}_k \leq Q_{0.5}\}$
- $C_3 = \{k \in \mathcal{K}, Q_{0.5} < \sigma^{\mathrm{s}}_k \leq Q_{0.75}\}$
- $C_4 = \{k \in \mathcal{K}, Q_{0.75} < \sigma^{\mathrm{s}}_k \leq Q_{0.9}\}$
- $C_5 = \{k \in \mathcal{K}, Q_{0.9} < \sigma^{\mathrm{s}}_k \leq \sigma^{\mathrm{s}}_{\max}\}$,

where $\sigma^{\mathrm{s}}_{\min} = \min\{\sigma^{\mathrm{s}}_k, k = 1, \ldots, K\}$, $\sigma^{\mathrm{s}}_{\max} = \max\{\sigma^{\mathrm{s}}_k, k = 1, \ldots, K\}$ and Q_e designates the e-th percentile of $\{\sigma^{\mathrm{s}}_k, k = 1, \ldots, K\}$.

Table 5.4 reports the gap to the best found solution and includes the results from Table 5.2 for comparison.

Table 5.4 Gap to best known value with clustering

		$I4$	IC
scalar α Formulation (5.24)-(5.28)	α_{mean}	143%	59%
	α_{max}	1037%	938%
	α_{q3}	349%	246%
	NS	17%	22%
	NSDI	5%	5%
	NOMAD	5%	16%
K-vector $\boldsymbol{\alpha}$ Formulation (5.36)-(5.56)	$\boldsymbol{\alpha}_{\text{max}}$	218%	87%
	$\boldsymbol{\alpha}_{\text{q3}}$	30%	27%
	NS	68%	59%
	NSDI	2%	59%
	NOMAD	6%	-
CR and (5.29)-(5.34)	NS	1%	59%
	NSDI	1%	2%
	NOMAD	1%	**0%**
CRU and (5.29)-(5.34)	NS	6%	13%
	NSDI	1%	4%
	NOMAD	2%	2%
CV and (5.29)-(5.34)	NS	21%	25%
	NSDI	**0%**	12%
	NOMAD	**0%**	1%

Several findings emerge. Combining clustering and metaheuristics allows to further improve the costs. It reduces the costs by 2% for the instance $I4$, and by 5% for the instance IC.

Clusters created from CV lead to a flexible plan able to handle sudden changes in demand at each week. They define a higher periodic demand for commodities whose demand forecasts have a large variance over the tactical planning periods. Then, when the forecasts at one period are higher than the average over the periods, the design is better adapted. Clusters from CR allow to increase the capacity for commodities that compete for the same trains, which lead to a reduction of the outsourcing.

Best performance are obtained when solving the formulation (5.29)-(5.34) with the clustering step for both the metaheuristics and NOMAD. Therefore, while clustering reduces the feasible set, it allows to leverage the state-of-the-art blackbox optimization solvers for large-scale

applications and produces high-quality solutions.

Computing Times Since tactical planning concerns medium-term decisions, the computing times of the solution method for **PDE** is not critical. It is nonetheless important but highly depends on the application considered, the time required to solve both **BP** and **wBP** and the optimal gap criteria. We report in Table 5.5 the number of evaluations, that is the number of time the algorithms solve the lower levels **BP-wBP** before reaching the best found solution. We note that on average, **BP** and **wBP** reached the optimality gap criteria in 1.97s and 5.06s for instance $I4$ and in 109.1s and 85.8s for instance IC. Therefore we can impose a time limit for both **BP** and **wBP** such that the CPLEX MIP procedure stops when a maximum of 900-s computational time is exhausted.

Table 5.5 Number of evaluations. We indicate by 1 when the initial solution is not improved, and by a dash when the algorithm could not solved the problem.

		$I4$	IC
scalar α Formulation (5.24)-(5.28)	NS	31	46
	NSDI	46	222
	NOMAD	26	52
K-vector $\boldsymbol{\alpha}$ Formulation (5.36)-(5.56)	NS	31	1
	NSDI	79	1
	NOMAD	168	-
CR and (5.29)-(5.34)	NS	31	1
	NSDI	61	69
	NOMAD	107	225
CRU and (5.29)-(5.34)	NS	76	31
	NSDI	107	247
	NOMAD	126	450
CV and (5.29)-(5.34)	NS	16	46
	NSDI	367	11
	NOMAD	214	184

The results show that while NOMAD finds the best solution, the method NSDI gives a good solution in less evaluations, hence less time. For the instance IC with CR clustering, NSDI gives a solution 2% more expensive than NOMAD, but in less than a third of the evaluations.

5.6 Conclusion

In this work we focused on the periodic demand estimation problem for large-scale tactical planning. The latter are often modelled by SND problems, and most studies considering

large-scale real-life instances have been focusing on deterministic approaches for the sake of computational tractability. Therefore, this is the approach we took in this paper. We have considered a setting where the demand forecasts are known and we aimed at estimating the periodic demand. We proposed a new formulation of the PDE problem where the periodic demand is defined as a deviation from the average of the forecasts, and the first-level variables are the deviation coefficients. Hence, there are K variables, the number of commodities transported in the network. Our approach allowed interpretability, as the periodic demand is a comparison to the mean, which is crucial for carriers.

We developed two local search metaheuristics to solve PDE taking advantage of its sequential property: when the first level variables are fixed, the lower levels can be solved sequentially. The first metaheuristic is a simple neighborhood search which stops when the solution is not improved while the second one contains diversification and intensification steps to explore the set of solutions. We have compared the metaheuristics with NOMAD, a turnkey blackbox optimization software that performs best for problems with few variables.

We presented a heuristic to reduce the size of the set of feasible solutions while keeping high-quality solutions. It consists in defining clusters of commodities having the same deviation coefficient, hence reducing the number of variables. To form the clusters, we proposed to either exploit the properties of the network or the distribution of the demand forecasts.

We reported results for a real large-scale application at the Canadian National Railway Company. The results showed that defining the periodic demand as a deviation from the commonly used periodic demand, i.e., the average of the time series forecasts, lead to substantial cost reductions. Moreover, the combined steps of clustering and local search algorithm to solve PDE allowed to reach the best costs. By reducing the number of variables, the clustering step makes the use of NOMAD possible for large-scale applications with hundreds of variables. While they might not find the best solution, the metaheuristics proposed in this paper allow to reach good solutions that are 1%-2% more expensive than the best one, in substantially smaller computing times.

We solved the lower levels of the PDE problem sequentially to optimality. This is critical for our application to orient the local search in good directions, however it resulted in significant computing times. There might be other applications, where the optimality gap can be relaxed, resulting in a speed up of the metaheuristics.

Future work should investigate two main avenues. The first one would focus on improving the computing times. Even though it is not a high-priority criteria for tactical planning, it is still important and there are two main directions: fasten the time to solve the lower levels to optimality and reduce the number of iterations required by each algorithm to reach

the best solution. The second avenue would focus on the clustering step and we want to develop a learning algorithm which learns the set of clusters from the network structure and the instance considered.

CHAPTER 6 ARTICLE 3: ASSESSING THE IMPACT: DOES AN IMPROVEMENT TO A REVENUE MANAGEMENT SYSTEM LEAD TO AN IMPROVED REVENUE ?

The text of this chapter is the one of the research paper *Assessing the Impact: Does an Improvement to a Revenue Management System Lead to an Improved Revenue?* (Laage et al., 2021a), submitted to the journal *Omega, The International Journal of Management Science*.

Authors Greta Laage, Emma Frejinger, Andrea Lodi, Guillaume Rabusseau.

Abstract Airlines and other industries have been making use of sophisticated Revenue Management Systems to maximize revenue for decades. While improving the different components of these systems has been the focus of numerous studies, estimating the impact of such improvements on the revenue has been overlooked in the literature despite its practical importance. Indeed, quantifying the benefit of a change in a system serves as support for investment decisions. This is a challenging problem as it corresponds to the difference between the generated value and the value that would have been generated keeping the system as before. The latter is not observable. Moreover, the expected impact can be small in relative value.

In this paper, we cast the problem as counterfactual prediction of unobserved revenue. The impact on revenue is then the difference between the observed and the estimated revenue. The originality of this work lies in the innovative application of econometric methods proposed for macroeconomic applications to a new problem setting. Broadly applicable, the approach benefits from only requiring revenue data observed for origin-destination pairs in the network of the airline at each day, before and after a change in the system is applied. We report results using real large-scale data from Air Canada. We compare a deep neural network counterfactual predictions model with econometric models. They achieve respectively 1% and 1.1% of error on the counterfactual revenue predictions, and allow to accurately estimate small impacts (in the order of 2%).

Key words Data analytics, Decision support systems, Performance evaluation, Air transport, Revenue Management

6.1 Introduction

Airlines have been making use of sophisticated Revenue Management Systems (RMSs) to maximize revenue for decades. Through interacting prediction and optimization components, such systems handle demand bookings, cancellations and no-shows, as well as the optimization of seat allocations and overbooking levels. Improvements to existing systems are made by the airlines and solution providers in an iterative fashion, aligned with the advancement of the state-of-the-art where studies typically focus on one or a few components at a time (Talluri and Van Ryzin, 2005). The development and maintenance of RMSs require large investments. In practice, incremental improvements are therefore often assessed in a proof of concept (PoC) prior to full deployment. The purpose is then to assess the performance over a given period of time and limited to certain markets, for example, a subset of the origin-destination pairs offered for the movement of passengers on the airline's network. We focus on a crucial question in this context: *Does the improvement to the RMS lead to a significant improvement in revenue?* This question is difficult to answer because the value of interest is not directly observable. Indeed, it is the difference between the value generated during the PoC and *the value that would have been generated* keeping business as usual. Moreover, the magnitude of the improvement can be small in a relative measure (for example, 1-3%) while still representing important business value. Small relative values can be challenging to detect with statistical confidence.

Considering the wealth of studies aiming to improve RMSs, it is surprising that the literature focused on assessing quantitatively the impact of such improvements is scarce. We identify two categories of studies in the literature: First, those assessing the impact in a simplified setting leveraging simulation (Weatherford and Pölt, 2002; Fiig et al., 2019). These studies provide valuable information but are subject to the usual drawback of simulated environments. Namely, the results are valid assuming that the simulation behaves as the real system. This is typically not true for a number of reasons, for instance, assumptions on demand can be inaccurate and in reality there can be a human in the loop adjusting the system. Statistical field experiments do not have this drawback as they can be used to assess impacts in a real setting. Studies focusing on field experiments constitute our second category. There are, however, few applications in revenue management (Lopez Mateos et al., 2021; Koushik et al., 2012; Pekgün et al., 2013) and even less focus on the airline industry (Cohen et al., 2019). Each application presents its specific set of challenges. Our work can be seen as a field experiment whose aim is to assess if a PoC is a success or not with respect to a given success criteria. In practice, airlines often take a pragmatic approach and compare the value generated during a PoC to a simple baseline: either the revenue generated at the same time

of the previous year, or the revenue generated by another market with similar behavior as the impacted market. This approach has the advantage of being simple. However, finding an adequate market is difficult, and the historical variation between the generated revenue and the baseline can exceed the magnitude of the impact that we aim to measure. In this case, the answer to the question of interest would be inconclusive.

We propose casting the problem as counterfactual prediction of the revenue without changing the RMS, and we compare it to the observed revenue generated during the PoC. Before providing background on counterfactual prediction models, we introduce some related vocabulary in the context of our application. Consider a sample of *units* and observations of *outcomes* for all units over a given time period. In our case, an example of a unit is an origin-destination (OD) pair and the observed outcome is the associated daily revenue. Units of interest are called *treated units* and the other (untreated) units are referred to as *control units*. In our case, the *treatment* is a change to the RMS and it only impacts the treated units (in our example a subset of the ODs in the network). The goal is to estimate the *untreated outcomes* of *treated units* defined as a function of the outcome of the control units. In other words, the goal is to estimate what would have been the revenue for the treated OD pairs without the change to the RMS. We use the observed revenue of the untreated ODs for this purpose.

Brief background on counterfactual prediction models Doudchenko and Imbens (2016) and Athey et al. (2021) review different approaches for imputing missing outcomes which include the three we consider for our application: (i) synthetic controls (Abadie and Gardeazabal, 2003; Abadie et al., 2010) (ii) difference-in-differences (Ashenfelter and Card, 1985; Card, 1990; Card and Krueger, 1994; Athey and Imbens, 2006) and (iii) matrix completion methods (Mazumder et al., 2010; Candès and Recht, 2009; Candès and Plan, 2010). Doudchenko and Imbens (2016) propose a general framework for difference-in-differences and synthetic controls where the counterfactual outcome for the treated unit is defined as a linear combination of the outcomes of the control units. Methods (i) and (ii) differ by the constraints applied to the parameters of the linear combination. Those models assume that the estimated patterns across units are stable before and after the treatment while models from the unconfoundedness literature (Imbens and Rubin, 2015; Rosenbaum and Rubin, 1983) estimate patterns from before treatment to after treatment that are assumed stable across units. Athey et al. (2021) qualify the former as vertical regression and the latter as horizontal regression.

Amjad et al. (2018) propose a robust version of synthetic controls based on de-noising the matrix of observed outcomes. Poulos (2017) proposes an alternative to linear regression

methods, namely a non-linear recurrent neural network. Athey et al. (2021) propose a general framework for counterfactual prediction models under matrix completion methods, where the incomplete matrix is the one of observed outcomes without treatment for all units at all time periods and the missing data patterns are not random. They draw on the literature on factor models and interactive fixed effects (Bai, 2003; Bai and Ng, 2002) where the untreated outcome is defined as the sum of a linear combination of covariates, that is, a low rank matrix and an unobserved noise component.

The studies in the literature are mainly focused on macroeconomic applications. For example, estimating the economic impact on West Germany of the German reunification in 1990 (Abadie et al., 2015), the effect of a state tobacco control program on per capita cigarette sales (Abadie et al., 2010) and the effect of a conflict on per capita GDP (Abadie and Gardeazabal, 2003). In comparison, our application exhibits some distinguishing features. First, the number of treated units can be large since airlines may want to estimate the impact on a representative subset of the network. Often there are hundreds, if not thousands of ODs in the network. Second, the number of control units is potentially large but the network structure leads to potential spillover effects that need to be taken into account. Third, even if the number of treated units can be large, the expected treatment effect is typically small. In addition, airline networks are affected by other factors, such as weather and seasonality. Their impact on the outcome needs to be disentangled from that of the treatment.

Contributions This paper offers three main contributions. First, we formally introduce the problem and provide a comprehensive overview of existing counterfactual prediction models that can be used to address it. Second, based on real data from Air Canada, we provide an extensive computational study showing that the counterfactual predictions accuracy is high when predicting revenue. We focus on a setting with multiple treated units and a large set of controls. We present a non-linear deep learning model to estimate the missing outcomes that takes as input the outcome of control units as well as time-specific features. The deep learning model achieves less than 1% error for the aggregated counterfactual predictions over the treatment period. Third, we present a simulation study of treatment effects showing that we can accurately estimate the effect even when it is relatively small.

Paper Organization. The remainder of the paper is structured as follows. Next we present a thorough description of the problem. We describe in Section 6.3 the different counterfactual prediction models. In Section 6.4, we describe our experimental setting and the results of an extensive computational study. Finally, we provide some concluding remarks in Section 6.5.

6.2 Problem Description

In this section, we provide a formal description of the problem and follow closely the notation from Doudchenko and Imbens (2016) and Athey et al. (2021).

We are in a panel data setting with N units covering time periods indexed by $t = 1, \ldots, T$. A subset of units is exposed to a binary treatment during a subset of periods. We observe the realized outcome for each unit at each period. In our application, a unit is an OD pair and the realized outcome is the *booking issue date revenue* at time t, that is, the total revenue yielded at time t from bookings made at t. The methodology described in this paper is able to handle various types of treatments, assuming it is applied to a subset of units. The set of *treated units* receive the treatment and the set of *control units* are not subject to any treatment. The treatment effect is the difference between the observed outcome under treatment and the outcome without treatment. The latter is unobserved and we focus on estimating the missing outcomes of the treated units during the treatment period.

We denote T_0 the time when the treatment starts and split the complete observation period into a pre-treatment period $t = 1, \ldots, T_0$ and a treatment period $t = T_0 + 1, \ldots, T$. We denote $T_1 = T - T_0$ the length of the treatment period. Furthermore, we partition the set of units into treated $i = 1, \ldots, N^{\mathrm{t}}$ and control units $i = N^{\mathrm{t}} + 1, \ldots, N$, where the number of control units is $N^{\mathrm{c}} = N - N^{\mathrm{t}}$.

In the pre-treatment period, both control units and treated units are untreated. In the treatment period, only the control units are untreated and, importantly, we assume that they are unaffected by the treatment. The set of treated pairs (i, t) is

$$\mathcal{M} = \{(i,t)\ i = 1, \ldots, N^{\mathrm{t}}, t = T_0 + 1, \ldots, T\}, \tag{6.1}$$

and the set of untreated pairs (i, t) is

$$\mathcal{O} = \{(i,t)\ i = 1, \ldots, N^{\mathrm{t}}, t = 1, \ldots, T_0\} \cup \{(i,t)\ i = N^{\mathrm{t}} + 1, \ldots, N, t = 1, \ldots, T\}. \tag{6.2}$$

Moreover, the treatment status is denoted by W_{it} and is defined as

$$W_{it} = \begin{cases} 1 & \text{if } (i,t) \in \mathcal{M} \\ 0 & \text{if } (i,t) \in \mathcal{O}. \end{cases} \tag{6.3}$$

For each unit i in period t, we observe the treatment status W_{it} and the realized outcome $Y_{it}^{\mathrm{obs}} = Y_{it}(W_{it})$. Our objective is to estimate $\hat{Y}_{it}(0)\ \forall (i,t) \in \mathcal{M}$. Counterfactual prediction models define the latter as a mapping of the outcome of the control units.

The observation matrix, denoted by $\mathbf{Y}^{\text{obs}}$ is a $N \times T$ matrix whose components are the observed outcomes for all units at all periods. The first N^{t} rows correspond to the outcomes for the treated units and the first T_0 columns to the pre-treatment period. The matrix $\mathbf{Y}^{\text{obs}}$ hence has a block structure,

$$\mathbf{Y}^{\text{obs}} = \begin{pmatrix} \mathbf{Y}^{\text{obs,t}}_{\text{pre}} & \mathbf{Y}^{\text{obs,t}}_{\text{post}} \\ \mathbf{Y}^{\text{obs,c}}_{\text{pre}} & \mathbf{Y}^{\text{obs,c}}_{\text{post}} \end{pmatrix},$$

where $\mathbf{Y}^{\text{obs,c}}_{\text{pre}}$ (respectively $\mathbf{Y}^{\text{obs,t}}_{\text{pre}}$) represents the $N^{\text{c}} \times T_0$ (resp. $N^{\text{t}} \times T_0$) matrix of observed outcomes for the control units (resp. treated units) before treatment. Similarly, $\mathbf{Y}^{\text{obs,c}}_{\text{post}}$ (respectively $\mathbf{Y}^{\text{obs,t}}_{\text{post}}$) represents the $N^{\text{c}} \times T_1$ (resp. $N^{\text{t}} \times T_1$) matrix of observed outcomes for the control units (resp. treated units) during the treatment.

Synthetic control methods have been developed to estimate the average causal effect of a treatment (Abadie and Gardeazabal, 2003). Our focus is slightly different as we aim at estimating the total treatment effect during the treatment period $T_0 + 1, \ldots, T$,

$$\tau = \sum_{i=1}^{N^{\text{t}}} \sum_{t=T_0+1}^{T} Y_{it}(1) - Y_{it}(0). \tag{6.4}$$

We denote by $\hat{\tau}$ the estimated treatment effect,

$$\hat{\tau} = \sum_{i=1}^{N^{\text{t}}} \sum_{t=T_0+1}^{T} Y^{\text{obs}}_{it} - \hat{Y}_{it}(0). \tag{6.5}$$

6.3 Counterfactual Prediction Models

In this section, we describe counterfactual prediction models from the literature that can be used to estimate the missing outcomes $Y_{it}(0)$ $\forall (i,t) \in \mathcal{M}$. Namely, grouped under synthetic control methods (Section 6.3.1), we describe the constrained regressions in Doudchenko and Imbens (2016) which include difference-in-differences and synthetic controls from Abadie et al. (2010). In Section 6.3.2, we delineate the robust synthetic control estimator from Amjad et al. (2018) followed by the matrix completion with nuclear norm minimization from Athey et al. (2021) in Section 6.3.3. Note that we present all of the above with one single treated unit, i.e., $N^{\text{t}} = 1$. This is consistent with our application as we either consider the units independently, or we sum the outcome of all treated units to form a single one. Finally, in Section 6.3.4, we propose a feed-forward neural network architecture that either considers a single treated unit or several to relax the independence assumption.

6.3.1 Synthetic Control Methods

Doudchenko and Imbens (2016) propose the following linear structure for estimating the unobserved $Y_{it}(0)$, $(i,t) \in \mathcal{M}$, arguing that several methods from the literature share this structure. More precisely, it is a linear combination of the control units,

$$Y_{it}(0) = \mu + \sum_{j=N^{\mathrm{t}}+1}^{N} \omega_j Y_{jt}^{\mathrm{obs}} + e_{it} \quad \forall (i,t) \in \mathcal{M}, \tag{6.6}$$

where μ is the intercept, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{N^{\mathrm{c}}})^\top$ a vector of N^{c} parameters and e_{it} an error term. Synthetic control methods differ in the way the parameters of the linear combination are chosen depending on specific constraints and the observed outcomes $\mathbf{Y}_{\mathrm{pre}}^{\mathrm{obs,t}}$, $\mathbf{Y}_{\mathrm{pre}}^{\mathrm{obs,c}}$ and $\mathbf{Y}_{\mathrm{post}}^{\mathrm{obs,c}}$. We write it as an optimization problem with an objective function minimizing the sum of least squares

$$\min_{\mu,\boldsymbol{\omega}} \left\| \mathbf{Y}_{\mathrm{pre}}^{\mathrm{obs,t}} - \mu \mathbf{1}_{T_0}^\top - \boldsymbol{\omega}^\top \mathbf{Y}_{\mathrm{pre}}^{\mathrm{obs,c}} \right\|^2, \tag{6.7}$$

potentially subject to one or several of the following constraints

$$\mu = 0 \tag{6.8}$$

$$\sum_{j=N^{\mathrm{t}}+1}^{N} \omega_j = 1 \tag{6.9}$$

$$\omega_j \geq 0, \quad j = N^{\mathrm{t}}+1, \ldots, N \tag{6.10}$$

$$\omega_j = \bar{\omega}, \quad j = N^{\mathrm{t}}+1, \ldots, N. \tag{6.11}$$

In the objective (6.7), $\mathbf{1}_{T_0}$ denotes a T_0 vector of ones. Constraint (6.8) enforces no intercept and (6.9) constrains the sum of the weights to equal one. Constraints (6.10) impose non-negative weights. Finally, constraints (6.11) force all the weights to be equal to a constant. If $T_0 \gg N$, Doudchenko and Imbens (2016) argue that the parameters μ and $\boldsymbol{\omega}$ can be estimated by least squares, without any of the constraints (6.8)-(6.11) and we may find a unique solution $(\mu, \boldsymbol{\omega})$. As we further detail in Section 6.4, this is the case in our application. We hence ignore all the constraints and estimate the parameters by least squares.

Difference-in-Differences

The Difference-In-Differences (DID) methods (Ashenfelter and Card, 1985; Card, 1990; Card and Krueger, 1994; Meyer et al., 1995; Angrist and Krueger, 1999; Bertrand et al., 2004;

Angrist and Pischke, 2008; Athey and Imbens, 2006) consist in solving

$$(DID) \quad \min_{\mu,\boldsymbol{\omega}} \left\| \mathbf{Y}_{\text{pre}}^{\text{obs,t}} - \mu \mathbf{1}_{T_0}^{\top} - \boldsymbol{\omega}^{\top} \mathbf{Y}_{\text{pre}}^{\text{obs,c}} \right\|^2 \tag{6.7}$$
$$\text{s.t. (6.9), (6.10), (6.11).}$$

With one treated unit and $N^{\text{c}} = N - 1$ control units, solving (*DID*) leads to the following parameters and counterfactual predictions:

$$\hat{\omega}_j^{\text{DID}} = \frac{1}{N-1}, \quad j = 2, \ldots, N \tag{6.12}$$

$$\hat{\mu}^{\text{DID}} = \frac{1}{T_0} \sum_{t=1}^{T_0} Y_{1t} - \frac{1}{(N-1)T_0} \sum_{t=1}^{T_0} \sum_{j=2}^{N} Y_{jt} \tag{6.13}$$

$$\hat{Y}_{1t}^{\text{DID}}(0) = \hat{\mu}^{\text{DID}} + \sum_{j=2}^{N} \hat{\omega}_j^{\text{DID}} Y_{jt}. \tag{6.14}$$

Abadie-Diamond-Hainmueller Synthetic Control Method

Introduced in Abadie and Gardeazabal (2003) and Abadie et al. (2010), the synthetic control approach consists in solving

$$(SC) \quad \min_{\mu,\boldsymbol{\omega}} \left\| \mathbf{Y}_{\text{pre}}^{\text{obs,t}} - \mu \mathbf{1}_{T_0}^{\top} - \boldsymbol{\omega}^{\top} \mathbf{Y}_{\text{pre}}^{\text{obs,c}} \right\|^2 \tag{6.7}$$
$$\text{s.t. (6.8), (6.9), (6.10).}$$

Constraints (6.8), (6.9) and (6.10) enforce that the treated unit is defined as a convex combination of the control units with no intercept.

The (SC) model is challenged in the presence of non-negligible levels of noise and missing data in the observation matrix $\mathbf{Y}^{\text{obs}}$. Moreover, it is originally defined for a small number of control units and relies on having deep domain knowledge to identify the controls.

Constrained Regressions

The estimator proposed by Doudchenko and Imbens (2016) consists in solving

$$(CR\text{-}EN) \quad \min_{\mu,\boldsymbol{\omega}} \left\| \mathbf{Y}_{\text{pre}}^{\text{obs,t}} - \mu \mathbf{1}_{T_0}^{\top} - \boldsymbol{\omega}^{\top} \mathbf{Y}_{\text{pre}}^{\text{obs,c}} \right\|_2^2 + \lambda^{\text{CR}} \left(\frac{1 - \alpha^{\text{CR}}}{2} ||\boldsymbol{\omega}||_2^2 + \alpha^{\text{CR}} ||\boldsymbol{\omega}||_1 \right), \tag{6.15}$$

while possibly imposing a subset of the constraints (6.8)-(6.11).

The second term of the objective function (6.15) serves as regularization. This is an *elastic-net*

regularization that combines the Ridge term which forces small values of weights and Lasso term which reduces the number of weights different from zero. It requires two parameters α^{CR} and λ^{CR}. To estimate their values, the authors propose a cross-validation procedure, where each control unit is alternatively considered as a treated unit and the remaining control units keep their role of control. They are used to estimate the counterfactual outcome of the treated unit. The parameters chosen minimize the mean-squared-error (MSE) between the estimations and the ground truth (real data) over the N^{c} validations sets.

The chosen subset of constraints depends on the application and the ratio of the number of time periods over the number of control units. In our experimental setting, we have a large number of pre-treatment periods, i.e., $T_0 \gg N^{\mathrm{c}}$ and we focus on solving *(CR-EN)* without constraints.

6.3.2 Robust Synthetic Control

To overcome the challenges of (SC) described in Section 6.3.1, Amjad et al. (2018) propose the Robust Synthetic Control algorithm. It consists in two steps: The first one de-noises the data and the second step learns a linear relationship between the treated units and the control units under the de-noising setting. The intuition behind the first step is that the observation matrix contains both the valuable information and the noise. The noise can be discarded when the observation matrix is approximated by a low rank matrix, estimated with singular value thresholding (Chatterjee et al., 2015). Only the singular values associated with valuable information are kept. The authors posit that for all units without treatment,

$$Y_{it}(0) = M_{it} + \epsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{6.16}$$

where M_{it} is the mean and ϵ_{it} is a zero-mean noise independent across all (i, t) (recall that for $(i, t) \in \mathcal{O}$, $Y_{it}(0) = Y_{it}^{\mathrm{obs}}$). A key assumption is that a set of weights $\{\beta_{N^{\mathrm{t}}+1}, \ldots, \beta_N\}$ exist such that

$$M_{it} = \sum_{j=N^{\mathrm{t}}+1}^{N} \beta_j M_{jt}, \quad i = 1, \ldots, N^{\mathrm{t}}, \quad t = 1, \ldots, T. \tag{6.17}$$

Before treatment, for $t \leq T_0$, we observe $Y_{it}(0)$ for all treated and control units. In fact, we observe $M_{it}(0)$ with noise. The latent matrix of size $N \times T$ is denoted $\mathbf{M}$. We follow the notation in Section 6.2: $\mathbf{M}^{\mathrm{c}}$ is the latent matrix of control units and $\mathbf{M}^{\mathrm{c}}_{\mathrm{pre}}$ the latent matrix of the control units in the pre-treatment period. We denote $\hat{\mathbf{M}}^{\mathrm{c}}$ the estimate of $\mathbf{M}^{\mathrm{c}}$ and $\hat{\mathbf{M}}^{\mathrm{c}}_{\mathrm{pre}}$ the estimate of $\mathbf{M}^{\mathrm{c}}_{\mathrm{pre}}$. With one treated unit, $i = 1$ designates the treated unit and the objective is to estimate $\hat{\mathbf{M}}^{\mathrm{t}}$, the latent vector of size T of treated units. The two-steps

algorithm is described in Algorithm 4. It takes two hyperparameters: the singular value threshold γ and the regularization coefficient η.

Algorithm 4 Robust Synthetic Control (Amjad et al., 2018)

Input: γ, η
Step 1: De-noising the data with singular value threshold
    Singular value decomposition of $\mathbf{Y}^{\text{obs,c}}$: $\mathbf{Y}^{\text{obs,c}} = \sum_{i=2}^{N} s_i u_i v_i^\top$
    Select the set of singular values above γ : $S = \{i : s_i \geq \gamma\}$
    Estimator $\hat{\mathbf{M}}^{\text{c}} = \frac{1}{\hat{p}} \sum_{i \in S} s_i u_i v_i^\top$, where $\hat{p}$ is the fraction of observed data
Step 2: Learning the linear relationship between controls and treated units
    $\hat{\boldsymbol{\beta}}(\eta) = \arg\min_{\mathbf{b} \in \mathbb{R}^{N-1}} \left\| \mathbf{Y}_{\text{pre}}^{\text{obs,t}} - \hat{\mathbf{M}}_{\text{pre}}^{\text{c}\top} \boldsymbol{b} \right\|^2 + \eta ||\boldsymbol{b}||_2^2$.
    Counterfactual means for the treatment unit: $\hat{\mathbf{M}}^{\text{t}} = \hat{\mathbf{M}}^{\text{c}\top} \hat{\boldsymbol{\beta}}(\eta)$
Return $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}(\eta) = \left(\hat{\mathbf{M}}_{\text{pre}}^{\text{c}} (\hat{\mathbf{M}}_{\text{pre}}^{\text{c}\top} + \eta \mathbf{I}) \right)^{-1} \hat{\mathbf{M}}_{\text{pre}}^{\text{c}} \mathbf{Y}^{\text{t}}{}_{\text{pre}} \tag{6.18}$$

Amjad et al. (2018) prove that the first step of the algorithm (which de-noises the data) allows to obtain a consistent estimator of the latent matrix. Hence, the estimate $\hat{\mathbf{M}}^{\text{c}}$ obtained with Algorithm 4 is a good estimate of $\mathbf{M}^{\text{c}}$ when the latter is low rank.

The threshold parameter γ acts as a way to trade-off the bias and the variance of the estimator. Its value can be estimated with cross-validation. The regularization parameter $\eta \geq 0$ controls the model complexity. To select its value, the authors recommend to take the forward chaining strategy, which maintains the temporal aspect of the pre-treatment data. It proceeds as follows. For each η, for each t in the pre-treatment period, split the data into 2 sets: $1, \ldots, t-1$ and t, where the last point serves as validation and select as value for η the one that minimizes the MSE averaged over all validation sets.

6.3.3 Matrix Completion with Nuclear Norm Minimization

Athey et al. (2021) propose an approach inspired by matrix completion methods. They posit a model similar to (6.16),

$$Y_{it}(0) = L_{it} + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{6.19}$$

where ε_{it} is a measure of error. This means that during the pre-treatment period, we observe L_{it} with some noise. The objective is to estimate the $N \times T$ matrix $\mathbf{L}$. Athey et al. (2021)

assume that the matrix $\mathbf{L}$ is low rank and hence can be estimated with a matrix completion technique. The estimated counterfactual outcomes of treated units without treatment $\hat{Y}_{it}(0), (i,t) \in \mathcal{M}$ is given by the estimate $\hat{L}_{it}, (i,t) \in \mathcal{M}$.

We use the following notation from Athey et al. (2021) to introduce their estimator. For any matrix $\mathbf{A}$ of size $N \times T$ with missing entries $\mathcal{M}$ and observed entries $\mathcal{O}$, $P_{\mathcal{O}}(\mathbf{A})$ designates the matrix with values of $\mathbf{A}$, where the missing values are replaced by 0 and $P^{\perp}_{\mathcal{O}}(\mathbf{A})$ the one where the observed values are replaced by 0.

They propose the following estimator of $\mathbf{L}$ from Mazumder et al. (2010), for a fixed value of λ^{mc}, the regularization parameter:

$$\hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left\{ \frac{1}{|\mathcal{O}|} ||P_{\mathcal{O}}(\mathbf{Y}^{\text{obs}} - \mathbf{L})||_F^2 + \lambda^{\text{mc}} ||\mathbf{L}||_* \right\}, \tag{6.20}$$

where $||\mathbf{L}||_F$ is the Fröbenius norm defined by

$$||\mathbf{L}||_F = \left(\sum_i \sigma_i(\mathbf{L})^2 \right)^2 = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} L_{it}^2 \right)^2 \tag{6.21}$$

with σ_i the singular values and $||\mathbf{L}||_*$ is the nuclear norm such that $||\mathbf{L}||_* = \sum_i \sigma_i(\mathbf{L})$. The first term of the objective function (6.20) is the distance between the latent matrix and the observed matrix. The second term is a regularization term encouraging $\mathbf{L}$ to be low rank.

Athey et al. (2021) show that their proposed method and synthetic control approaches are matrix completion methods based on matrix factorization. They rely on the same objective function which contains the Fröbenius norm of the difference between the unobserved and the observed matrices. Unlike synthetic controls that impose different sets of restrictions on the factors, they only use regularization.

Athey et al. (2021) use the convex optimization program SOFT-IMPUTE from Mazumder et al. (2010) described in Algorithm 5 to estimate the matrix $\mathbf{L}$. With the singular value decomposition $\mathbf{L} = \mathbf{S}\mathbf{\Sigma}\mathbf{R}^{\top}$, the matrix shrinkage operator is defined by $\text{shrink}_{\lambda^{\text{mc}}}(\mathbf{L}) = \mathbf{S}\tilde{\mathbf{\Sigma}}\mathbf{R}^{\top}$, where $\tilde{\mathbf{\Sigma}}$ is equal to $\mathbf{\Sigma}$ with the i-th singular value replaced by $\max(\sigma_i(\mathbf{L}) - \lambda^{\text{mc}}, 0)$.

The value of λ^{mc} can be selected via cross-validation as follows: For K subsets of data among the observed data with the same proportion of observed data as in the original observation matrix, for each potential value of λ_j^{mc}, compute the associated estimator $\hat{\mathbf{L}}(\lambda_j^{\text{mc}}, \mathcal{O}_k)$ and the MSE on the data without $\mathcal{O}_k$. Select the value of λ that minimizes the MSE. To fasten the convergence of the algorithm, the authors recommend to use $\hat{\mathbf{L}}(\lambda_j^{\text{mc}}, \mathcal{O}_k)$ as initialization for $\hat{\mathbf{L}}(\lambda_{j+1}^{\text{mc}}, \mathcal{O}_k)$ for each j and k.

Algorithm 5 SOFT-IMPUTE (Mazumder et al., 2010) for Matrix Completion with Nuclear Norm Maximization (Athey et al., 2021)

Initialization: $\mathbf{L}_1(\lambda^{\text{mc}}, \mathcal{O}) = \mathbf{P}_{\mathcal{O}}(\mathbf{Y}^{\text{obs}})$
for $k = 1$ until $\{\mathbf{L}_k(\lambda^{\text{mc}}, \mathcal{O})\}_{k \geq 1}$ converges **do**
    $\mathbf{L}_{k+1}(\lambda^{\text{mc}}, \mathcal{O}) = \text{shrink}_{\frac{\lambda^{\text{mc}}|\mathcal{O}|}{2}}\left(\mathbf{P}_{\mathcal{O}}(\mathbf{Y}^{\text{obs}}) + \mathbf{P}^{\perp}_{\mathcal{O}}\left(\mathbf{L}_k(\lambda)\right)\right)$
end for
$\hat{\mathbf{L}}(\lambda^{\text{mc}}, \mathcal{O}) = \lim_{k \to \infty} \mathbf{L}_k(\lambda^{\text{mc}}, \mathcal{O})$

6.3.4 Feed-forward Neural Network

In this section, we propose a deep learning model to estimate the missing outcomes and detail the training of the model. We consider two possible configurations: (i) when there is one treated unit and (ii) when there are multiple dependent treated units. In (i), the output layer of the model has one neuron. In (ii), the output layer contains N^{t} neurons. The model learns the dependencies between treated units and predicts simultaneously the revenue for all of them.

We define the counterfactual outcomes of the treated units as a non-linear function g of the outcomes of the control units with parameters θ^{ffnn} and matrix of covariates $\mathbf{X}$

$$\mathbf{Y}^{\text{t}}(0) = g\left(\mathbf{Y}^{\text{obs,c}}, \mathbf{X}, \theta^{\text{ffnn}}\right). \tag{6.22}$$

In the following subsections, we use terminology from the deep learning literature (Goodfellow et al., 2016) but keep the notations described in Section 6.2. We define g to be a feed-forward neural network (FFNN) architecture. We describe next the architecture in detail along with the training procedure.

Architecture

Barron (1994) shows that multilayer perceptrons (MLPs), also called FFNNs, are considerably more efficient than linear basis functions to approximate smooth functions. When the number of inputs I grows, the required complexity for an MLP only grows as $\mathcal{O}(I)$, while the complexity for a linear basis function approximator grows exponentially for a given degree of accuracy. When $N^{\text{t}} > 1$, the architecture is multivariate, i.e., the output layer has multiple neurons. It allows parameter sharing between outputs and thus considers the treated units as dependent.

Since historical observations collected prior to the beginning of the treatment period are

untreated, the counterfactual prediction problem can be cast as a supervised learning problem on the data prior to treatment. The features are the observed outcomes of the control units and the targets are the outcomes of the treated units. The pre-treatment period is used to train and validate the neural network and the treatment period forms the test set. This is a somewhat unusual configuration for supervised learning. Researchers usually know the truth on the test set also and use it to evaluate the ability to generalize. To overcome this difficulty, we describe in Section 6.3.4 a sequential validation procedure that aims at mimicking the standard decomposition of the dataset into training, validation and test sets.

We present in Figure 6.1 the model architecture. We use two input layers to differentiate features. Input Layer 1 takes external features, and Input Layer 2 takes the lagged outcomes of control units. Let us consider the prediction at day t as illustration. When t is a day, it is associated for instance to a day of the week dow_t, a week of the year woy_t and a month m_t. The inputs at Input Layer 1 could then be dow_t, woy_t, m_t. Lagged features of control units are $Y_{it'}, i = N^{\mathrm{t}}+1, \ldots, N$ and $t' = t, t-1, \ldots, t-l$, where l is the number of lags considered. They are fed into Input Layer 2. The output layer outputs N^{t} values, one for each treated unit.

Sequential Validation Procedure and Selection of Hyper-parameters

In standard supervised learning problems, the data is split into training, validation and test datasets, where the validation dataset is used for hyper-parameters search. Table 6.1 lists the hyper-parameters of our architecture and learning algorithm. For each potential set of hyper-parameters Θ, the model is trained on the training data and we estimate the parameters θ^{ffnn}. We compute the MSE between the predictions and the truth on the validation dataset. We select the set Θ which minimizes the MSE.

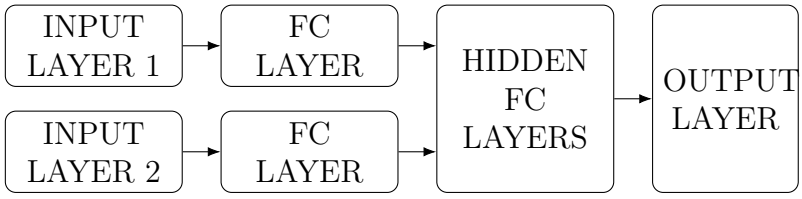


Figure 6.1 FFNN Architecture with Fully Connected (FC) layers

Table 6.1 Description of the hyper-parameters for the FFNN architecture

Name	**Description**
Hidden size	Size of the hidden layers
Hidden layers	Number of hidden layers after the concatenation of the dense layers from Input Layer 1 and Input Layer 2.
Context size	Size of the hidden FC layer after Input Layer 1
Batch size	Batch size for the stochastic gradient descent
Dropout	Unique dropout rate determining the proportion of neurons randomly set to zero for each output of the FC layers
Learning rate	Learning rate for the stochastic gradient descent.
Historical lags	Number of days prior to the date predicted considered for the control units outcomes.
Epochs Number	Number of epochs (iterations over the training dataset) required to train the model

One of the challenges of our problem is that the data have an important temporal aspect. While this is not a time series problem, for a test set period, we train the model with the last observed data, making the validation step for selecting hyper-parameters difficult. To overcome this challenge, we split chronologically the pre-treatment periods in two parts: $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{valid}}$. We train the model on $\mathcal{T}_{\text{train}}$ with the backpropagation algorithm using Early Stopping, a form of regularization to avoid overfitting that consists in stopping the training when the error on the validation set increases. We select Θ on $\mathcal{T}_{\text{valid}}$ and store $\hat{e}$, the number of epochs it took to train the model. As a final step, we train the model with hyper-parameters Θ for $\hat{e}$ epochs on $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{valid}}$, which gives an estimate $\hat{\theta}^{\text{ffnn}}$. Then, we compute the counterfactual predictions as $\hat{\mathbf{Y}}_t^{\text{t}}(0) = \hat{g}(\mathbf{Y}^{\text{obs,c}}, \mathbf{X}, \hat{\theta}^{\text{ffnn}})$ for $t = T_0 + 1, \ldots, T$.

Training Details

We present here some modeling and training tricks we used to achieve the best performance with the FFNN.

Data Augmentation Data augmentation is a well-known process to improve performances of neural networks and prevent overfitting. It is often used for computer vision tasks such as image classification (Shorten and Khoshgoftaar, 2019). It consists in augmenting the dataset by performing simple operations such as rotation, translation, symmetry, etc. We perform

one type of data augmentation, the homothety, which consists in increasing or reducing the (inputs, outputs) pair. We decompose it into the following steps. Let a denote the homothety maximum coefficient, typically an integer between 1 and 4. For each batch in the stochastic gradient descent algorithm, we multiply each sample, inputs and outputs, by a random number uniformly distributed between $1/a$ and a.

Ensemble Learning The ensemble learning algorithm relies on the intuition that the average performance of good models can be better than the performance of a single best model (Sagi and Rokach, 2018). We take a specific case of ensemble learning, where we consider as ensemble the 15 best models that provide the lowest MSE on the validation set from the hyper-parameter search. For each model $k = 1, \ldots, 15$, we store the set of hyper-parameters Θ_k and the number of training epochs $\hat{e}_k$. We train each model on the pre-treatment period to estimate $\hat{\theta}_k^{\text{ffnn}}$. We compute the counterfactuals $\hat{\mathbf{Y}}_t^{\text{t}k}(0) = \hat{g}_k(\mathbf{Y}^{\text{obs,c}}, \mathbf{X}, \hat{\theta}_k^{\text{ffnn}})$ and the predicted outcome is $\hat{\mathbf{Y}}_t^{\text{t}}(0) = \frac{1}{15}\sum_{k=1}^{15} \hat{\mathbf{Y}}_t^{\text{t}k}(0)$ for $t = T_0 + 1, \ldots, T$.

6.4 Application

This work was part of a large project with a major North American airline, Air Canada, operating a worldwide network. The objective of the overall project was to improve the accuracy of the demand forecasts of multiple ODs in the network. In this work, the new demand forecasting algorithm acts as the treatment. The details about the treatment is not part of this paper but it drove some of the decisions, especially regarding the selection of the treated and control units. The units correspond to the different ODs in the network and the outcome of interest is the revenue. In this paper, we present a computational study of a simulated treatment effect (ground truth impact is known). This was part of the validation work done prior to the PoC. Due to the uncertainty regarding the required duration of the treatment period, we planned for a period of 6 months in our validation study. For the sake of completeness, we also analyze the results for shorter treatment periods. Unfortunately, the Covid-19 situation hit the airline industry during the time of the PoC. It drastically changed the revenue and the operated flights making it impossible to assess the impact of the demand forecasts.

In the next section, we first provide details of our experimental setting. Next, in Section 6.4.2, we present the prediction performances of the models. In Section 6.4.3, we report results from a simulation study designed to estimate the revenue impact.

6.4.1 Experimental Setup and Data

Treatment Effect Definition There are two ways of considering the daily revenue yielded from bookings: by *flight date* or by *booking issue date*. The former is the total revenue at day t from bookings for flights departing at t, while the latter is the total revenue at day t from bookings made at t, for all possible departure dates for the flight booked. For our study, we consider the issue date revenue as it allows for a better estimation of the treatment effect. Indeed, as soon as the treatment starts at day $T_0 + 1$, all bookings are affected and thus the issue date revenue is affected. Hence, $Y_{it}(0)$ designates the untreated issue date revenue of OD i at day t. The treatment period is 6 months, i.e., $T_1 = 181$ days. The drawback of the flight date revenue is that only a subset of the flights is completely affected by the treatment, hence leading to an underestimation of the treatment effect. Only flights whose booking period starts at $T_0 + 1$ (or after) and for which the treatment period lasts for the full duration of the booking period, approximately a year, are completely affected.

Selection of Treated Units The selection of the treated ODs was the result of discussions with the airline. The objective was to have a sample of ODs representative of the North-American market, while satisfying constraints related to the demand managers in charge of those ODs. We select 15 non-directional treated ODs, i.e., 30 directional treated ODs ($N^{\mathrm{t}} = 30$). For instance, if Montreal-Boston was treated, then Boston-Montreal would be treated as well. The selected 30 ODs represent approximately 7% of the airline's yearly revenue.

Selection of Control Units The selection of control units depends on the treated units. Indeed, a change of the demand forecasts for an OD affects the RMS which defines the booking limits. Due to the network effect and the potential leg-sharing among ODs, this would in turn affect the demand for other ODs. With the objective to select control units that are *unaffected* by the treatment, we use the following restrictions:

- Geographic rule: for each treated OD, we consider two perimeters centered around the origin and the destination airports, respectively. We exclude all other OD pairs where either the origin or the destination is in one of the perimeters.
- Revenue ratio rule: for all ODs operated by the airline in the network, different from the treated ODs, we discard the ones where at least 5% of the itineraries have a leg identical to one of the treated ODs. This is because new pricing of OD pairs can affect the pricing of related itineraries, which in turn affects the demand.

- Sparse OD rule: we exclude seasonal ODs, i.e., those that operate only at certain times of the year. Moreover, we exclude all OD pairs that have no revenue on more than 85% of points in our dataset.

From the remaining set of ODs, we select the 40 most correlated ODs for each treated OD. The correlation is estimated with the Pearson correlation coefficient. These rules led to $N^{\mathrm{c}} = 317$ control units. We note that this selection is somewhat different from the literature, due to the network aspect of the airline operations and the abundance of potential control units. In Abadie et al. (2010), for instance, only a few controls are selected based on two conditions: (i) they have similar characteristics as the treated units and (ii) they are not affected by the treatment. The geographic restriction and the revenue ratio rule correspond to condition (ii). The sparse OD rule allows to partially ensure condition (i) as the treated ODs are frequent ODs from the airline's network. Considering a large number of controls has the advantage to potentially leverage the ability of deep learning models to capture the relevant information from a large set of features.

We ran several experiments with a larger set of control units, given that the geographic rule, the revenue ratio rule and the sparse OD rule were respected. In the following, we report results for the set of controls described above, as they provided the best performance.

Models and Estimators We compare the performance of the models and estimators detailed in Section 6.3:

- DID: Difference-in-Differences
- SC: Abadie-Diamond-Hainmueller Synthetic Controls
- CR-EN: Constrained Regressions with elastic-net regularization
- CR: CR-EN model with $\lambda^{\mathrm{CR}} = 0$ and $\alpha^{\mathrm{CR}} = 0$
- RSC: Robust Synthetic Controls
- MCNNM: Matrix Completion with Nuclear Norm Minimization
- FFNN: Feed-Forward Neural Network with Ensemble Learning. The external features of the FFNN are the day of the week and the week of the year. We compute a circular encoding of these two features using their polar coordinates to ensure that days 0 and 1 (respectively, week 52 and week 1 of the next year) are as distant as days 6 and days 0 (respectively, week 1 and week 2).

We started the analysis by investigating the approach often used in practice, which consists in comparing the year-over-year revenue. The counterfactual revenue is the revenue obtained in the same period of the previous year. We ruled out this approach due to its poor performance, both in terms of accuracy and variance. We provide details in Section 6.4.2, where we discuss the results.

Data The observed untreated daily issue date revenue covers the period from January 2013 to February 2020 for all control and treated units. This represents 907,405 data points. To test the performances of the different models, we select random periods of 6 months and predict the revenue values of the 30 treated ODs. In the literature, most studies use a random assignment of the pseudo-treated unit instead of a random assignment of treated periods. In our application, switching control units to treated units is challenging as the control set is specific to the treated units. Hence our choice of random assignment of periods. We refer to those periods as *pseudo-treated* as we are interested in retrieving the observed values. To overcome the challenges described in Section 6.3.4, we select random periods late in the dataset, between November 2018 and February 2020.

Two scenarios for the target variables. We consider two scenarios for the target variables: In the first – referred to as $S1$ – we aggregate the 30 treated units to a single one. In the second – referred to as $S2$ – we predict the counterfactual revenue for each treated unit separately. For both scenarios, our interest concerns the total revenue $Y_t = \sum_{i \in N^{\mathrm{t}}} Y_{it}$. In the following, we provide more details.

In $S1$, we aggregate the outcomes of the treated units to form one treated unit, even though the treatment is applied to each unit individually. The missing outcomes, i.e., the new target variables, are the values of Y_t^{agg}, where

$$\textit{(S1)} \quad Y_t^{\mathrm{agg}} = \sum_{i=1}^{N^{\mathrm{t}}} Y_{it}. \tag{6.23}$$

The models DID, SC, CR, CR-EN are in fact regressions on Y_t^{agg} with control unit outcomes as variates. For the models RSC and MCNNM, we replace in the observation matrix $\mathbf{Y}^{\mathrm{obs}}$ the N^{t} rows of the treated units revenue with the values of Y_t^{agg}, for $t = 1, \ldots, T$. All models estimate $\hat{Y}_t^{\mathrm{agg}}$, for $t = 1, \ldots, T$, and $\hat{Y}_t = \hat{Y}_t^{\mathrm{agg}}$.

In $S2$, we predict the counterfactual revenue for each treated OD. For models SC, DID, CR, CR-EN, MCNNM and RSC, this amounts to considering each treated unit as independent from the others and we estimate a model on each treated unit. For FFNN, we relax the

independence assumption so that the model can learn the dependencies and predict the revenue for each treated unit simultaneously. We have an estimate of the revenue for each pair (unit, day) in the pseudo-treatment period. Then, we estimate the total revenue at each period as the sum over each estimated pair, namely

$$\textit{(S2)} \quad \hat{Y}_t = \sum_{i \in N^{\mathrm{t}}} \hat{Y}_{it}. \tag{6.24}$$

Performance metrics We assess performance by analyzing standard Absolute Percentage Error (APE) and Root Mean Squared Error (RMSE). In addition, the bias of the counterfactual prediction model is an important metric as it, in turn, leads to a biased estimate of the impact. In our application, the observable outcome is the issue date net revenue from the bookings whose magnitude over a 6-month treatment period is measured in millions. A pseudo-period p has a length T_{1p} and we report for each p the percentage estimate of the total error

$$\mathrm{tPE}_p = \frac{\sum_{t=1}^{T_{1p}} \hat{Y}_t - \sum_{t=1}^{T_{1p}} Y_t}{\sum_{t=1}^{T_{1p}} Y_t} \times 100. \tag{6.25}$$

This metric allows us to have insights on whether the model tends to overestimate or underestimate the total revenue, which will be at use when estimating the revenue impact. We also report tAPE_p, the absolute values of tPE_p for a period p

$$\mathrm{tAPE}_p = \frac{|\sum_{t=1}^{T_{1p}} \hat{Y}_t - \sum_{t=1}^{T_{1p}} Y_t|}{\sum_{t=1}^{T_{1p}} Y_t} \times 100. \tag{6.26}$$

We present the results of $S1$ and $S2$ in the following. For confidentiality reasons, we only report relative numbers in the remainder of the paper with the focus of comparing the different models.

6.4.2 Prediction Performance

In this section, we start by analyzing the performance related to predicting daily revenue, followed by an analysis of total predicted revenue in Section 6.4.2.

Daily Predicted Revenue

We assess the performances of the models at each day t of a pseudo-treatment period, i.e., the prediction error on $\hat{Y}_t$ at each day t. We compute the errors for each t and report the

values average over all the pseudo-treatment period p, namely

$$\text{MAPE}_p = \frac{1}{T_1}\sum_{t=1}^{T_1}\frac{|\hat{Y}_t - Y_t|}{Y_t}, \quad \text{RMSE}_p = \sqrt{\frac{1}{T_1}\sum_{t=1}^{T_1}(\hat{Y}_t - Y_t)^2}. \tag{6.27}$$

For confidentiality reasons, we report a scaled version of RMSE_p for each p, which we refer to as RMSE^{s}_p. We use the average daily revenue of the first year of data as a scaling factor.

Figures 6.2 and 6.3 present MAPE_p and RMSE^{s}_p for $p = 1, \ldots, 15$, where the upper graph of each figure shows results for $S1$ and the lower the results for $S2$, respectively. We note that the performance is stable across pseudo-treated periods for all models. The values of MAPE_p at each period p of SC, RSC and CR models are below 5% while for FFNN it is only the case in $S2$. This is important, as the impact we wish to measure is less than this order of magnitude.

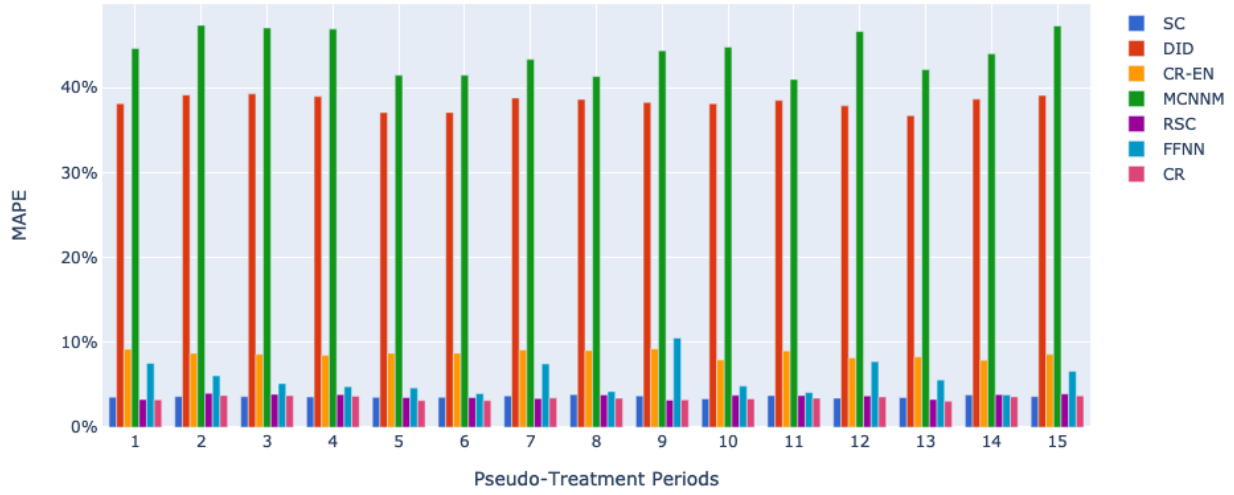
Table 6.2 reports the values of the metrics averaged over all pseudo-treatment periods for settings $S1$ and $S2$, i.e., $\text{MAPE} = \frac{1}{15}\sum_{p=1}^{15}\text{MAPE}_p$ and $\text{RMSE}^{\text{s}} = \frac{1}{15}\sum_{p=1}^{15}\text{RMSE}^{\text{s}}_p$. The results show that the best performance for both metrics and in both scenarios is achieved by CR model. On average, it reaches a MAPE of 3.4% and RMSE^{s} of 6.0. It achieves better results than CR-EN model. This is because we have $T \gg N$ and there are hence enough data to estimate the coefficients without regularization.

Table 6.2 Average of the daily MAPE and RMSE^{s} over all pseudo-treatment periods.

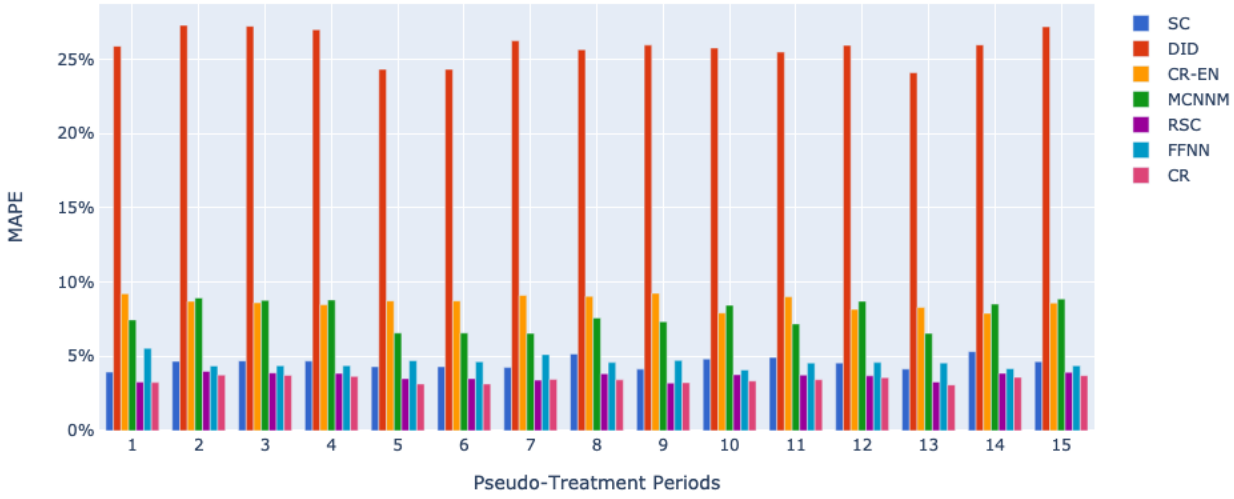
	$S1$		$S2$	
	MAPE	RMSE^{s}	MAPE	RMSE^{s}
CR	**3.4%**	**6.0**	**3.4%**	**6.0**
CR-EN	8.6%	15.0	8.6%	15.0
DID	38.3%	61.4	25.9%	39.2
FFNN	5.8%	9.4	4.6%	7.5
MCNNM	44.2%	70.0	7.8%	14.3
RSC	3.6%	6.5	3.6%	6.5
SC	3.6%	6.5	4.6%	8.3

Models DID and MCNNM have poor performance in $S1$. This is due to the difference in magnitude between the treated unit and the control units. In $S2$, the performance is improved because we build one model per treated unit. Each treated unit is then closer to the controls in terms of magnitude. Due to the constraint (6.11) of equal weights, DID model is not flexible enough and its performance does not reach that of the other models.

The FFNN model improves the MAPE by 1.2 points from $S1$ to $S2$. The neural network models the dependencies between the treated ODs and gain accuracy by estimating the

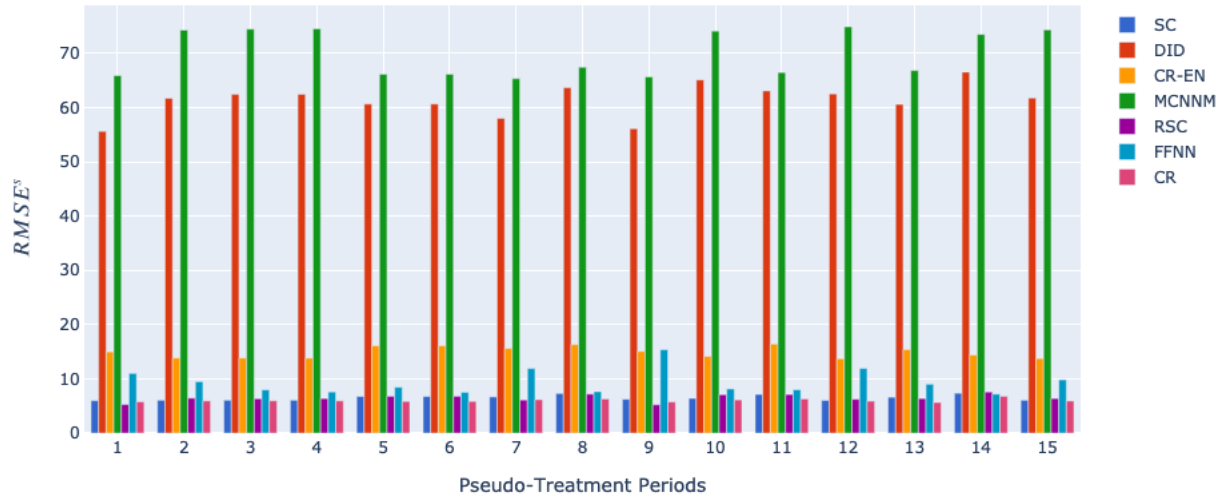


(a)

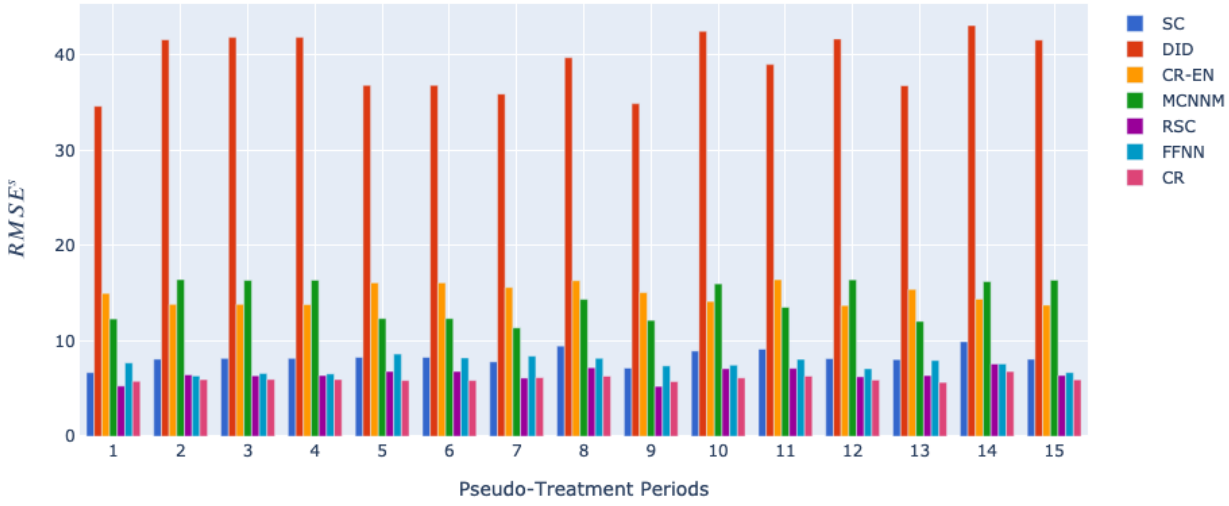


(b)

Figure 6.2 Values of daily error, MAPE_p, in each pseudo-treatment period (a) in Setting $S1$ with one model for a single aggregated unit and (b) in Setting $S2$ with one model per treated unit (note that the y-axis has a different scale in the two graphs).



(a)



(b)

Figure 6.3 Values of daily RMSE^s in each pseudo-treatment period (a) in Setting $S1$ with one model for a single aggregated unit and (b) in Setting $S2$ with one model per treated unit (note that the y-axis has a different scale in the two graphs).

revenue of each treated OD.

The advantage of $S2$ is that we predict separately the outcome for each unit at each day. In addition to computing the error between $\hat{Y}_t$ and Y_t for each pseudo-treatment period, we can also compute the error between $\hat{Y}_{it}$ and Y_{it}, for $i = 1, \ldots, N^{\mathrm{t}}$, and $t = 1, \ldots, T_1$, namely

$$\mathrm{MAPE}_i^{\mathrm{od}} = \frac{1}{T_1} \sum_{t=1}^{T_1} \frac{|\hat{Y}_{it} - Y_{it}|}{Y_{it}}, \quad \mathrm{MAPE}^{\mathrm{od}} = \frac{1}{N^{\mathrm{t}}} \sum_{i=1}^{N^{\mathrm{t}}} \mathrm{MAPE}_i^{\mathrm{od}}. \tag{6.28}$$

Figure 6.4 presents the values of $\mathrm{MAPE}^{\mathrm{od}}$ for each pseudo-treatment period, and Table 6.3 reports the average value of $\mathrm{MAPE}^{\mathrm{od}}$ over all pseudo-treatment periods. It shows that results are consistent across periods. Method SC reaches the best accuracy, with on average 13.1% of error for the daily revenue of one treated OD. The FFNN model has a similar performance with 13.3% of error on average. We conclude that estimating the counterfactual revenue of one OD is difficult and we gain significant accuracy by aggregating over the treated ODs. In the remainder of the paper, we only consider models CR, CR-EN, FFNN, RSC and SC as they perform best.

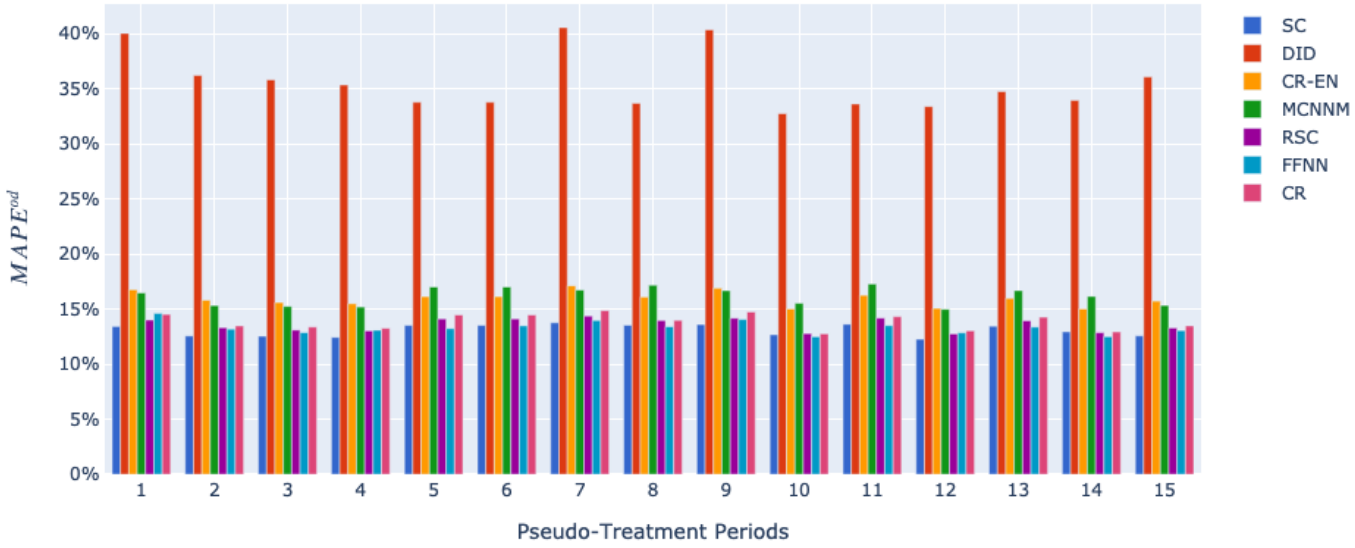


Figure 6.4 $\mathrm{MAPE}^{\mathrm{od}}$ for each pseudo-treatment period in $S2$.

Total Predicted Revenue

In this section, we analyze the models' performance over a complete pseudo-treatment period. We first consider a pseudo-treatment period of 6 months, and we then analyze the effect of

Table 6.3 MAPE^{od} averaged over all pseudo-treatment periods in $S2$.

	MAPE^{od}
CR	13.8%
CR-EN	16.0%
DID	35.6%
FFNN	13.3%
MCNNM	16.2%
RSC	13.6%
SC	**13.1**%

a reduced length.

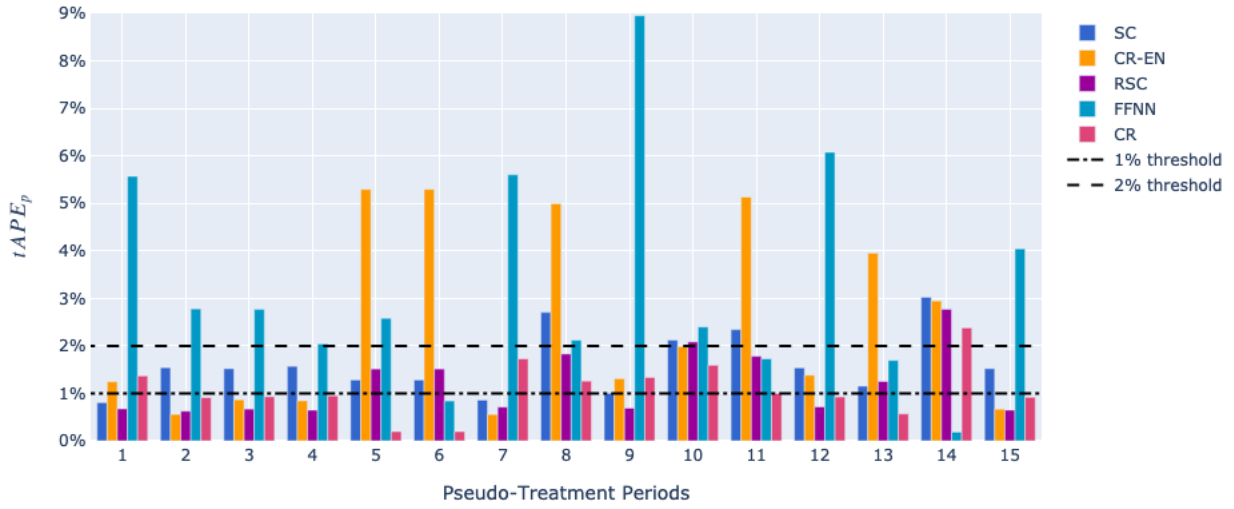
Figure 6.5 presents the value of tAPE_p defined in (6.26) for pseudo-treatment periods $p = 1, \ldots, 15$. The upper graph shows the results for $S1$ and the lower the results for $S2$, respectively. To illustrate treatment impacts' order of magnitude, we depict the 1% and 2% thresholds in dashed lines. We note that FFNN and CR-EN models have higher variance than SC, CR and RSC methods which stay below 3% of error at each period. Moreover, the model FFNN is stable across all periods for $S2$.

Table 6.4 reports the values of $\text{tAPE} = \frac{1}{15}\sum_{p=1}^{15} \text{tAPE}_p$ for each model. All models are able to predict the total 6-months counterfactual revenue with less than 3.5% of error on average, in both settings. For $S1$, the CR method reaches the best performance, with 1.1% error on average and, for $S2$, the best is the FFNN model with 1.0% average error.

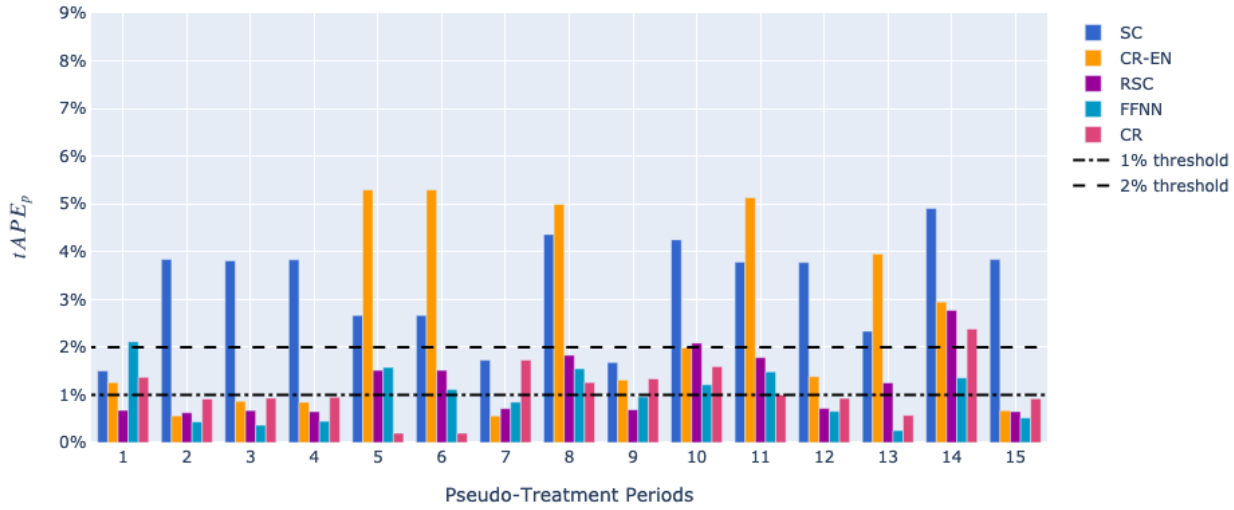
Table 6.4 tAPE over all pseudo-treatment periods

	$S1$	$S2$
	tAPE	tAPE
CR	**1.1**%	1.1%
CR-EN	2.5%	2.5%
FFNN	3.3%	**1.0%**
RSC	1.2%	1.2%
SC	1.6%	3.3%

We present in Figure 6.6 the values of tPE_p defined in (6.25) at each period $p = 1, \ldots, 15$. It shows that for $S1$, the FFNN model systematically overestimates the total counterfactual revenue while SC, CR-EN and RSC methods systematically underestimate it. For $S2$, we observe the same behavior for models SC, CR-EN and RSC while both FFNN and CR methods either underestimate or overestimate the counterfactual revenue.

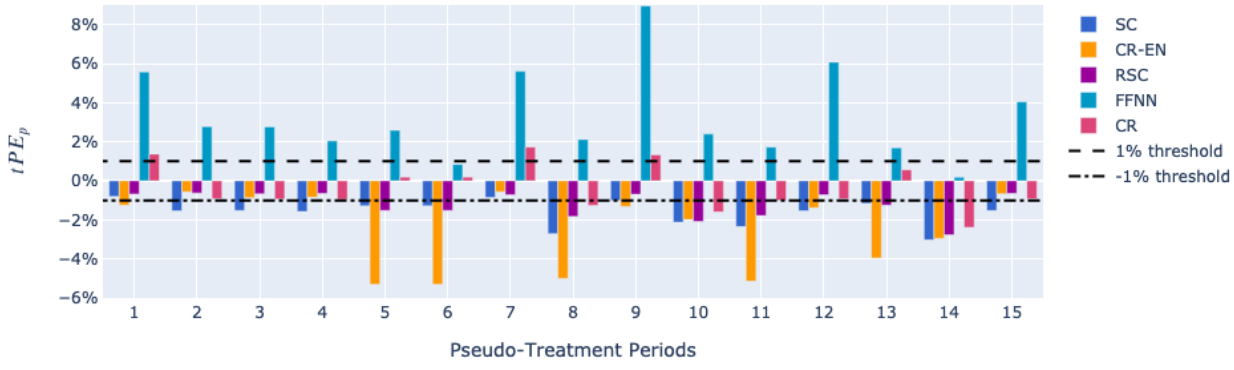


(a)

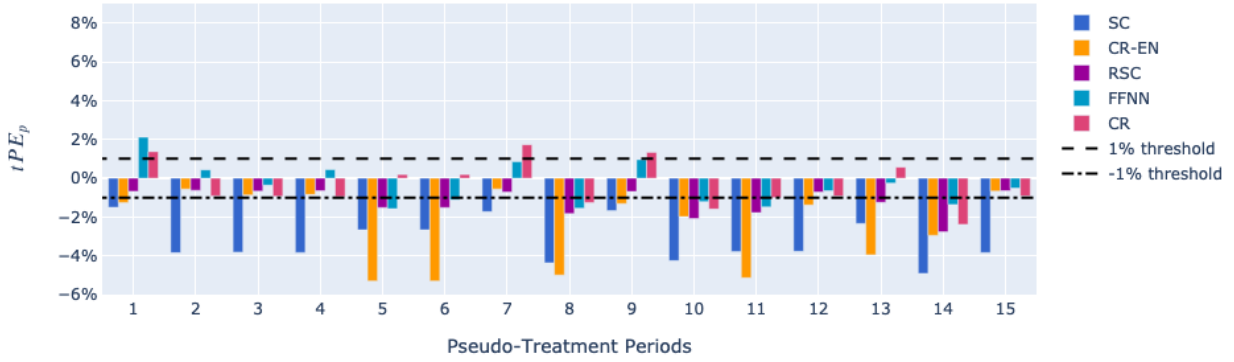


(b)

Figure 6.5 Values of tAPE_p for each pseudo-treatment period (a) in Setting $S1$ with one model for a single aggregated unit and (b) in Setting $S2$ with one model per treated unit.



(a)



(b)

Figure 6.6 Values of tPE_p for each pseudo-treatment period $p = 1, \ldots, 15$ (a) in Setting $S1$ with one model for a single aggregated unit and (b) in Setting $S2$ with one model per treated unit.

Length of the treatment period We now turn our attention to analyzing the effect of the treatment duration period on performance. For this purpose, we study the variations of tAPE_p for different values of T_1 for the pseudo-treatment periods $p = 1, \ldots, 15$. We analyze the results for each period but for illustration purposes we focus only on the second one. We report the values for all the other periods in Appendix B (the general observations we describe here remain valid).

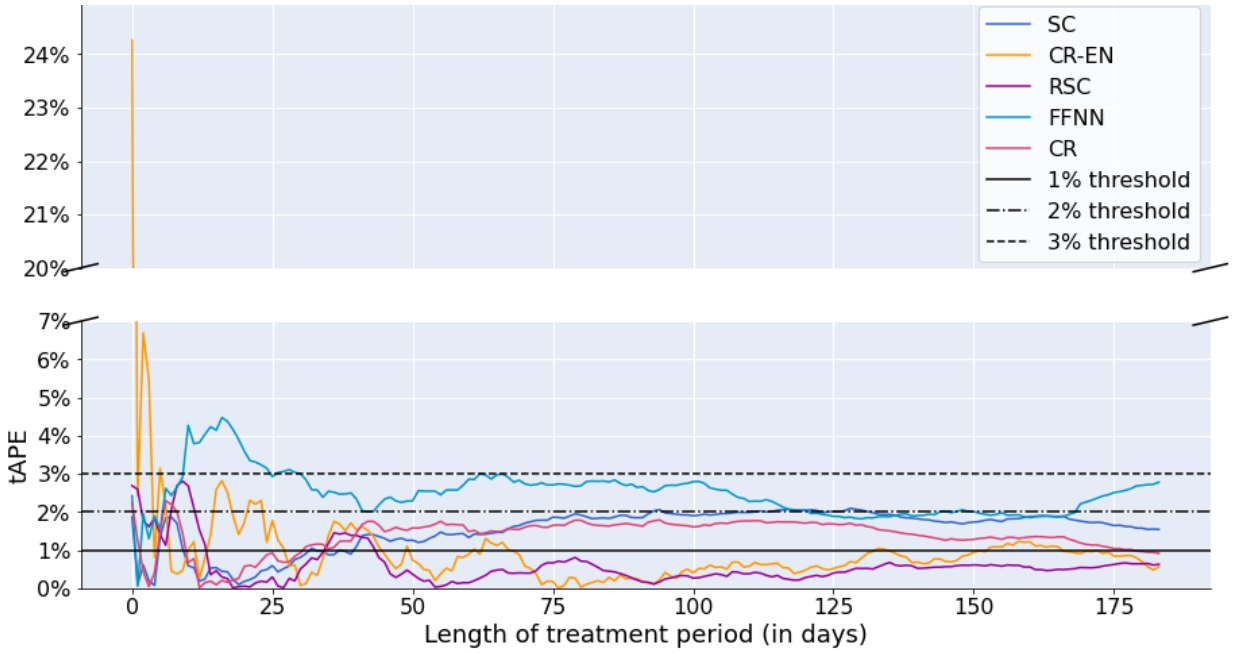
Figure 6.7 presents the variations of tAPE_2 against the length T_1 for the different models. The upper graph shows the results for $S1$ and the lower one the results for $S2$, respectively. The black lines (solid and dashed) represent the 1%, 2% and 3% thresholds. In $S1$, values of tAPE_2 for FFNN are below 3% from 30 days. After 30 and 39 days, respectively, tAPE_2 values for CR and SC are between 1% and 2%. Values of tAPE_2 are below 1% from 68 days for CR-EN and from 43 days for RSC. In $S2$, tAPE_2 for FFNN is below 2% from 52 days and below 1% from 84 days. For CR and CR-EN, it is below 2% from 10 days and 18 days, respectively. It is below 1% from 44 days for RSC. Hence, the results show that the length of the treatment period can be less than six months as models are accurate after only a few weeks.

The CR, RSC and FFNN models present high accuracy with errors less than 1.2% for the problem of counterfactual predictions on the total revenue. This is compelling since we are interested in detecting a small treatment impact. As anticipated in Section 6.4.1, we considered simpler approaches that are common practice. For example, comparing to year-over-year revenue. In this case, the counterfactual revenue is defined as the revenue generated during the same period but the year before. It had a poor performance, with a tAPE between 7% and 10% at each pseudo-treatment period. This approach is therefore not accurate enough to detect small impacts.

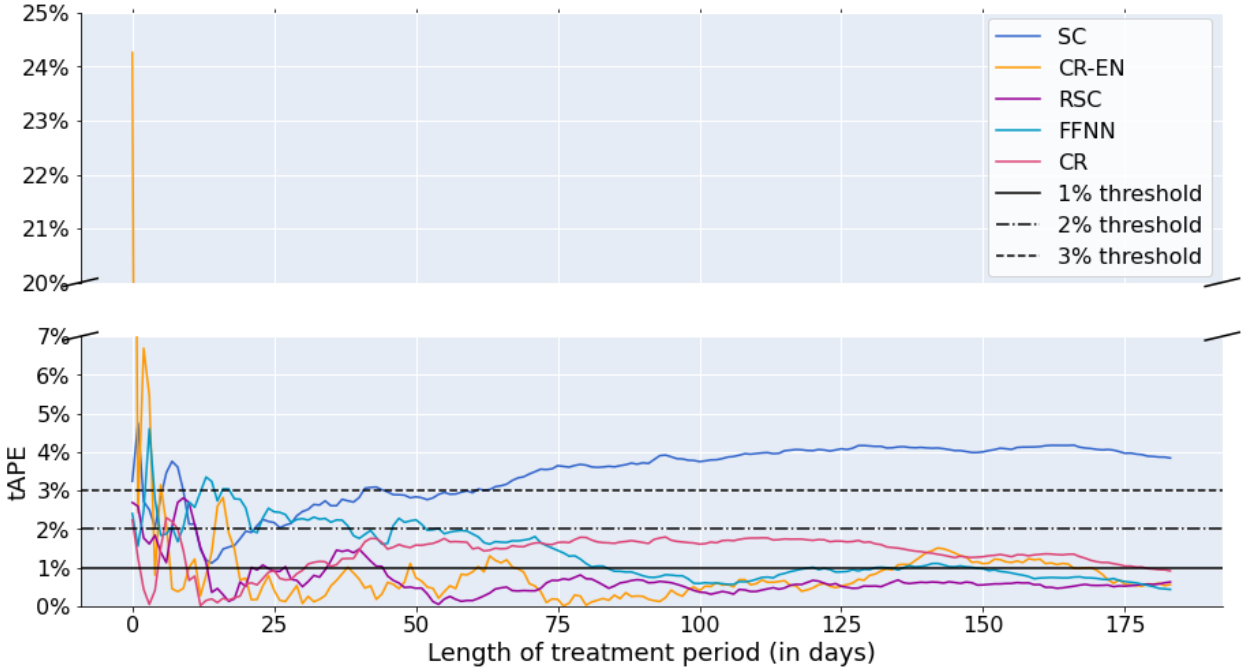
In the following section, we present a validation study where we simulate small impacts and assess our ability to estimate them with counterfactual prediction models.

6.4.3 Validation: Revenue Impact Estimate for Known Ground Truth

We consider a pseudo-treatment period of 6 months and the setting $S2$. In this case, models FFNN, CR and RSC provide accurate estimations of the counterfactual total revenue with respectively 1%, 1.1% and 1.2% of error on average over the pseudo-treatment periods. We restrict the analysis that follows to those models. We proceed in two steps: First, we simulate a treatment by adding a noise with positive mean to the revenue of the treated units at each day of each pseudo-treatment period. We denote $\tilde{Y}_t^{\text{obs}}$ the new treated value, $\tilde{Y}_t^{\text{obs}} = Y_t(0) \times \epsilon, \quad \epsilon \sim \text{Lognormal}(\mu_\epsilon, \sigma_\epsilon^2)$ and $\sigma_\epsilon^2 = 0.0005$. We simulate several treatment impacts



(a)



(b)

Figure 6.7 Values of tAPE_2 varying with the length of the treatment period T_1 (a) in Setting $S1$ with one model for a single aggregated unit and (b) in Setting $S2$ with one model per treated unit.

that differ by the value of μ_ϵ. Second, we compute the impact estimate with (6.5) from the counterfactual predictions and compare it to the actual treatment applied in the first step. We present the results for one pseudo-treatment period, $p = 2$.

Table 6.5 reports the values of the estimated impact for different values of μ_ϵ. The first row shows the values for the true counterfactuals. This is used as reference, as it is the exact simulated impact. Results show that RSC and CR models overestimate the impact while FFNN model underestimates it. This is because the former underestimates the counterfactual predictions while the latter overestimates them. Due to the high accuracy of counterfactual predictions, both the underestimation and overestimation are however small. We can detect impacts higher than the accuracy of the counterfactual prediction models. The simulation shows that we are close to the actual impact.

Table 6.5 Estimation of the revenue impact $\hat{\tau}$ of simulated treatment

Counterfactuals	$\mu_\epsilon = 0.01$	$\mu_\epsilon = 0.02$	$\mu_\epsilon = 0.03$	$\mu_\epsilon = 0.05$
Ground truth	1.0%	2.0%	3.0%	5.1%
RSC	1.7%	2.6%	3.7%	5.7%
CR	1.5%	2.5%	3.5%	5.6%
FFNN	0.6%	1.6%	2.6%	4.7%

Figure 6.8 presents the daily revenue on a subset of the treatment periods. The estimation of the daily revenue impact is the difference between the simulated revenue (solid and dashed black lines) and the counterfactual predictions (colored lines). This figure reveals that even though the accuracy of the daily predictions is not as good as on the complete treatment period, we can still detect even a small daily impact.

Prediction intervals. It is clear that prediction intervals for the estimated revenue impact are of high importance. However, it is far from trivial to compute them for most of the counterfactual prediction models in our setting. Under some assumptions, the CR model in setting $S1$ constitutes the exception. More precisely, if the residuals satisfy conditions (i) independent and identically distributed and (ii) normally distributed, then we can derive a prediction interval for the sum of the daily predicted revenue. For the simulated impacts reported in Table 6.5, we obtain 99% prediction intervals with widths of 2.2%. It means that we can detect an impact of 2% or more with high probability.

Cattaneo et al. (2020) develop prediction intervals for the SC model that account for two distinct sources of randomness: the construction of the weights ω and the unobservable stochastic error in the treatment period. Moreover, Zhu and Laptev (2017) build prediction

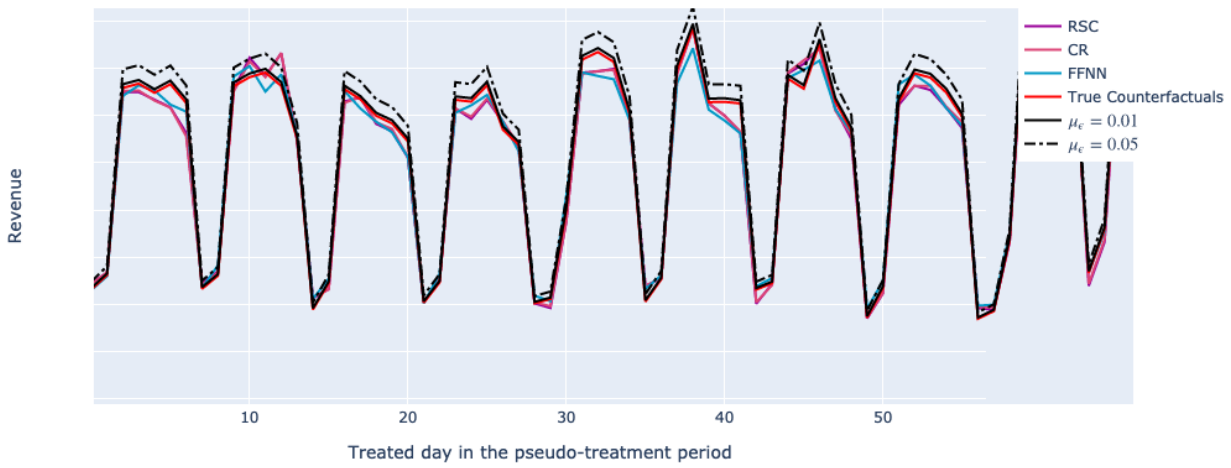


Figure 6.8 Daily revenue and predictions for a subset of the pseudo-treatment period 2. The labels in the y-axis are hidden for confidentiality reasons.

intervals for neural networks predictions that consider three sources of randomness: model uncertainty, model misspecification and data generation process uncertainty. Both studies focus on computing prediction intervals for *each* prediction. We face an additional issue as we need a prediction interval for the sum of the predictions. As evidenced by these two studies, computing accurate prediction intervals is a challenging topic on its own and we therefore leave it for future research.

6.5 Conclusion

Revenue management systems are crucial to the profitability of airlines and other industries. Due to their importance, solution providers and airlines invest in the improvement of the different system components. In this context, it is important to estimate the impact on an outcome such as revenue after a proof of concept. We addressed this problem using counterfactual prediction models.

In this paper, we assumed that an airline applies a treatment (a change to the system) on a set of ODs during a limited time period. We aimed to estimate the total impact over all of the treated units and over the treatment period. We proceeded in two steps. First we estimated the counterfactual predictions of the ODs' outcome, that is the outcome if no treatment were applied. Then, we estimated the impact as the difference between the observed revenue under treatment and the counterfactual predictions.

We compared the performance of several counterfactual prediction models and a deep-learning model in two different settings. In the first one, we predicted the aggregated outcome of the treated units while in the second one, we predicted the outcome of each treated unit and aggregated the predictions. We showed that synthetic control methods and the deep-learning model reached a competitive accuracy on the counterfactual predictions, which in turn allows to accurately estimate the revenue impact. The deep-learning model reaches the lowest error of 1% in the second setting, leveraging the dependency between treated units. The best counterfactual prediction model, which in the second setting assumes treated units are independent, reached 1.1% of error in both settings. We showed that we can reduce the length of a treatment period and preserve this level of accuracy. This can be useful as it potentially allows to reduce the cost of proofs of concepts.

We believe that the methodology is broadly applicable to decision support systems, and not limited to revenue management (e.g., upgrade of a software, new marketing policy). It can assess the impact of a proof of concept under the following fairly mild assumptions: (i) the units under consideration (e.g., origin-destination pairs, markets, sites or products) can be divided into two subsets, one affected by the treatment and one that is unaffected (ii) time can be divided into two (not necessarily consecutive) periods, a pre-treatment period and a treatment period (iii) the outcome of interest (any objective function value, for example, revenue, cost or market share) can be measured for each unit.

Finally, we will dedicate future research to devise prediction intervals for the sum of the counterfactual predictions, which in turn will lead to a prediction interval for the estimated impact.

Acknowledgements

We are grateful for the invaluable support from the whole Crystal AI team who built the demand forecasting solution. The team included personnel from both Air Canada and IVADO Labs. In particular, we would like to thank Richard Cleaz-Savoyen and the Revenue Management team for careful reading and comments that have helped improving the manuscript. We also thank Florian Soudan from Ivado Labs and Pedro Garcia Fontova from Air Canada for their help and advice in training the neural network models. We would like to especially thank William Hamilton from IVADO Labs who has contributed with ideas and been involved in the results analysis. Maxime Cohen provided valuable comments that helped us improve the manuscript.We express our gratitude to Peter Wilson (Air Canada) who gave valuable business insights guiding the selection of control units. The project was partially funded by Scale AI. Finally, the first author would like to thank Louise Laage and the third author

would like to thank Luca Nunziata and Marco Musumeci for many stimulating discussions on counterfactual prediction and synthetic control.

CHAPTER 7 GENERAL DISCUSSION

The three works presented through Chapters 4 to 6 naturally influence and echo each other, both methodologically and in terms of content. They focus on decision-making systems, the information they require and the impact they have on the profitability of a transportation company.

The in-depth analysis of the importance of the periodic demand on the tactical planning costs (Chapters 4 and 5) belongs to the same thought of the impact of demand forecasts accuracy on the revenue. This is an important question for airlines, which make an extensive use of Revenue Management Systems (RMSs) that highly rely on demand forecasts. RMS, and more generally decision-making systems, are intricate and sophisticated, and estimating the impact of the performance of one of their components is challenging. This led to the second question we considered in this thesis: assessing the impact of improvements made to decision-making systems.

Most of the literature targets the improvements, on the modeling and solution methodology side, advancing the state-of-the-art. However, the models are in general only an approximation of reality and many other constraints should be taken into account in practice. There are numerous intermediate steps between the improvements made to the system, the impact on the transported demand, and therefore the profitability of the company. Being able to accurately assess the impact in a real system is crucial for investment decisions.

Overall, the three problems, forecasting, planning and impact assessment are well-studied problems separately, even though they are strongly related. Yet their integration for real-life large-scale applications has been overlooked in the literature in the context of our problems, while having high value in practice. Our work aimed at addressing this gap, while taking into account the constraints of real-life applications, namely the most used formulations (deterministic models for tactical planning for instance) and the need for interpretable solutions.

Finally, from the methodological standpoint, each work draws on several methodologies resulting in diverse methods studied in this research: statistics, econometrics, machine learning, mathematical programming and metaheuristics. Each one was originally developed to answer a specific problem, yet real-life applications gather multiple problems. In the past few years, many works have been focusing on combining methodologies, especially from an algorithmic perspective, for instance the use of machine learning for linear programming. We believe that it is also important to combine methodologies from an application perspective, as it gives access to broader research questions and allows to improve decision-making systems.

CHAPTER 8 CONCLUSION AND RECOMMENDATIONS

This thesis discussed and proposed methods to integrate demand forecasting and planning problems, and to estimate the impact on key performance indicators. After a brief summary of the presented contributions, we conclude this dissertation by highlighting their limitations and proposing some directions for related future research.

8.1 Summary of works

In Chapter 4, we focused on large-scale tactical planning problems, which often require deterministic formulations of the service network design problem. We formally introduced the *periodic demand estimation problem* which allows to estimate the periodic demand such that the planning costs are minimized. The latter include the fixed costs of the plan and the variable costs incurred by adapting the plan when demand changes. We proposed a methodology that proceeds in two steps. The first step consists in using a time series forecasting model to predict demand for each period in the tactical planning horizon. We developed and compared models from the statistics and machine learning literature containing observed demand as features. Statistical models provide good performances, and neural networks highlight the importance of considering external features such as the weather. The second step defines the periodic demand as a solution to a multilevel mathematical program that explicitly connects the estimation problem to the tactical planning problem. We introduced a new problem in this chapter and focused on the in-depth analysis of the importance of the periodic demand. Given the complexity of the problem, we limited the feasible set of variables to be small and discrete which allowed us to solve the problem by enumerating the solutions. We reported results for a real large-scale application at the Canadian National Railway Company. Even with the restrictions, we showed that using another estimate of periodic demand from the common practice that simply consists in averaging the time series forecasts over the tactical planning horizon lead to substantial reductions of costs.

Motivated by the results from Chapter 4, we developed further the periodic demand estimation problem in Chapter 5 by allowing a broad and continuous feasible set of periodic demands and propose a new solution approach to solve the problem. In fact, we defined the periodic demand variable as a deviation from the average of the demand forecasts. We hence proposed a new formulation where the decision variables are the deviation coefficients. We proposed two new local search metaheuristics to solve the problem and compare their performances to an off-the-shelf blackbox optimization solver. For large-scale applications,

the number of commodities, and in turn the number of variables is large, and the solution algorithms have limited performances. To address this challenge, we developed heuristic approaches creating clusters of commodities that have the same value of deviation, hence reducing the number of variables. They exploit the information contained in the network and the demand distribution of the commodities over the planning horizon. We reported results on the same application from Chapter 4, that showed that defining the periodic demand as a deviation from the average of the time series forecasts lead to substantial cost reductions. Moreover, the combined steps of clustering and search algorithms allowed to reach the best performance. By reducing the number of variables, the clustering step leverages the blackbox software even for large-scale applications with hundreds of variables.

In Chapter 6, we consider the problem of assessing the impact on the revenue of a carrier after an improvement made to its decision system. We cast the problem as a counterfactual prediction problem, and we aimed at estimating the counterfactual revenue. It is the revenue that would have been observed without improving the system. We focus on a setting where the improvements to the system concerned multiple Origin-Destination pairs (ODs), and the set of ODs in the carrier's network is divided into treated ODs, subject to the improvement, and control ODs, which are not affected. We formally introduced the problem and provided a comprehensive overview of existing counterfactual prediction models. We also presented a non-linear deep learning model taking as input the outcome of control units as well as time-specific features. We reported results for a real large-scale application at Air Canada. Results showed that the accuracy of the counterfactual predicted revenue of the treated ODs is high, allowing us to estimate a relatively small impact.

8.2 Limitations and Future research

The works presented in this thesis belong to the body of improving decision-making systems for large-scale transportation networks. We focused on well-studied problems for which we identified important issues that have not been recognized as such, and have high value for carriers in practice. Our contributions propose new methodologies, perspectives and managerial insights for those issues, and expose avenues for future research.

The methodology of the periodic demand estimation problem relies on the demand forecasts. Our *a posteriori* analysis in Chapter 4 showed that the best tactical costs are obtained by solving the PDE problem when the periodic demand is defined as a mapping from the *actual demand* instead of the *demand forecasts*. However, only the latter are available at the time of planning. Therefore, the first avenue for future research would be the improvement of forecasting models. Additional data are collected every day, which could be useful to develop

combined statistics and machine learning models.

The proposed approach defined the periodic demand as a mapping from the point estimates of demand forecasts at each period of the tactical planning horizon. While we focused on deterministic formulations of SND problems, our approach and stochastic programming are not mutually exclusive. Indeed, recognizing the uncertainty around the point estimates of demand forecasts at each period of the tactical planning horizon is an interesting avenue for future research. Introducing robustness in the PDE problem is a potential approach to handle demand uncertainty, and developing revenue management strategies to better control the demand spikes is another.

In the same spirit of combining methodologies within a model, another direction for future research is the development of machine learning models learning the clusters to create in the heuristic developed in Chapter 5 with the objective of minimizing the tactical costs. Such direction could combine machine learning models with optimization models to embed the network structure and the periodic demand estimation problem. The challenges would reside in first the computational tractability, but also the interpretability, a crucial question for practitioners.

Finally, regarding the work on impact assessment, an important future research avenue is the development of prediction intervals for the impact estimation, another crucial point for practitioners.

In addition to the research avenues, it is undeniable that the implementation of the proposed methods in transport companies fully integrating the business challenges (computing times, robustness) constitutes an avenue to be privileged. The integration of demand forecasting and planning for rail freight carriers could generate revenue management opportunities that are still in their early stage of development. Then, being able to accurately estimate the impact would encourage companies to frequently test new implementations or strategies.

REFERENCES

Abadie, A. Using synthetic controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

Abadie, A. and Gardeazabal, J. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003.

Abadie, A., Diamond, A., and Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Abadie, A., Diamond, A., and Hainmueller, J. Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510, 2015.

Abramson, M. A., Audet, C., and Dennis Jr, J. Filter Pattern Search Algorithms for Mixed Variable Constrained Optimization Problems. *SIAM Journal on Optimization*, 11:573–594, 2004.

Abramson, M. A., Audet, C., Chrissis, J. W., and Walston, J. G. Mesh adaptive direct search algorithms for mixed variable optimization. *Optimization Letters*, 3(1):35–47, 2009.

Agarwal, R. and Ergun, Ö. Ship Scheduling and Network Design for Cargo Routing in Liner Shipping. *Transportation Science*, 42(2):175–196, 2008.

Alarie, S., Audet, C., Garnier, V., Le Digabel, S., and Leclaire, L.-A. Snow water equivalent estimation using blackbox optimization. *Pac. J. Optim.*, 9(1):1–21, 2013.

Amjad, M., Shah, D., and Shen, D. Robust Synthetic Control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

Angrist, J. D. and Krueger, A. B. Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, volume 3, 1277–1366. 1999.

Angrist, J. D. and Pischke, J.-S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Ashenfelter, O. and Card, D. Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *Review of Economics and Statistics*, 67(4):648–660, 1985.

Athey, S. and Imbens, G. W. Identification and Inference in Nonlinear Difference-In-Differences Models. *Econometrica*, 74(2):431–497, 2006.

Athey, S. and Imbens, G. W. The state of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, 1–41, 2021.

Audet, C. and Dennis Jr, J. E. Mesh Adaptive Direct Search Algorithms for Constrained Optimization. *SIAM Journal on optimization*, 17(1):188–217, 2006.

Audet, C. and Hare, W. *Derivative-Free and Blackbox Optimization*. Springer, 2017.

Azadeh, S. S., Marcotte, P., and Savard, G. A taxonomy of demand uncensoring methods in revenue management. *Journal of Revenue and Pricing Management*, 13(6):440–456, 2014.

Bai, J. Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1): 135–171, 2003.

Bai, J. and Ng, S. Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221, 2002.

Bai, R., Wallace, S. W., Li, J., and Chong, A. Y.-L. Stochastic service network design with rerouting. *Transportation Research Part B: Methodological*, 60:50–65, 2014.

Barron, A. R. Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, 14(1):115–133, 1994.

Bergstra, J., Yamins, D., and Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *ICML'13: Proceedings of the 30th International Conference on Machine Learning*, 28:115–123, 2013.

Bertrand, M., Duflo, E., and Mullainathan, S. How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Breiman, L. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231, 2001.

Candès, E. J. and Plan, Y. Matrix Completion with Noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

Candès, E. J. and Recht, B. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

Card, D. The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relation*, 43(2):245–257, 1990.

Card, D. and Krueger, A. B. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4): 772–793, 1994.

Cattaneo, M. D., Feng, Y., and Titiunik, R. Prediction intervals for Synthetic Control Methods. *arXiv:1912.07120*, 2020.

Černỳ, V. Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.

Chatterjee, S. et al. Matrix Estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Chouman, M., Crainic, T. G., and Gendron, B. Commodity Representations and Cut-Set-Based Inequalities for Multicommodity Capacitated Fixed-Charge Network Design. *Transportation Science*, 51(2):650–667, 2017.

Cohen, M., Jacquillat, A., and Serpa, J. A Field Experiment on Airline Lead-in Fares. Technical report, Working Paper, 2019.

Colson, B., Marcotte, P., and Savard, G. Bilevel programming: A survey. *4OR*, 3:87–107, 2005.

Crainic, T. G. Service network design in freight transportation. *European Journal of Operational Research*, 122(2):272–288, 2000.

Crainic, T. G. and Kim, K. H. Chapter 8 Intermodal Transportation. *Handbooks in Operations Research and Management Science*, 14:467 – 537, 12 2007.

Crainic, T. G., Hewitt, M., Maggioni, F., and Rei, W. Partial Benders Decomposition: General Methodology and Application to Stochastic Network Design. *Transportation Science*, 55(2):275–552, 2020.

Doudchenko, N. and Imbens, G. W. Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis. Technical report, National Bureau of Economic Research, 2016.

Feo, T. A. and Resende, M. G. A Probabilistic Heuristic for a Computationally Difficult Set Covering Problem. *Operations Research Letters*, 8(2):67–71, 1989.

Feo, T. A. and Resende, M. G. Greedy Randomized Adaptive Search Procedures. *Journal of Global Optimization*, 6(2):109–133, 1995.

Fiig, T., Weatherford, L. R., and Wittman, M. D. Can demand forecast accuracy be linked to airline revenue? *Journal of Revenue and Pricing Management*, 18(4):291–305, 2019.

Glover, F. Future Paths for Integer Programming and Links to Artificial Intelligence. *Computers & Operations Research*, 13(5):533–549, 1986.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Hertz, A. Notes de Cours, MTH 6311: Optimisation Combinatoire, 2016.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Holt, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

Hornik, K., Stinchcombe, M., and White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989.

Imbens, G. W. and Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning*. Springer, 2013.

Karlaftis, M. G. and Vlahogianni, E. I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399, 2011.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

Koushik, D., Higbie, J. A., and Eister, C. Retail Price Optimization at InterContinental Hotels Group. *INFORMS Journal on Applied Analytics*, 42(1):45–57, 2012.

Laage, G., Frejinger, E., Lodi, A., and Rabusseau, G. Assessing the Impact: Does an Improvement to a Revenue Management System Lead to an Improved Revenue? *arXiv preprint arXiv:2101.10249*, 2021a.

Laage, G., Frejinger, E., and Savard, G. Periodic Freight Demand Forecasting for Large-scale Tactical Planning. *arXiv preprint arXiv:2105.09136*, 2021b.

Längkvist, M., Karlsson, L., and Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.

Le Digabel, S. Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 37(4):1–15, 2011.

Lium, A.-G., Crainic, T. G., and Wallace, S. W. A Study of Demand Stochasticity in Service Network Design. *Transportation Science*, 43(2):144–157, 2009.

Lopez Mateos, D., Cohen, M. C., and Pyron, N. Field Experiments for Testing Revenue Strategies in the Hospitality Industry. *Cornell Hospitality Quarterly*, 1–10, 2021.

Magnanti, T. L. and Wong, R. T. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science*, 18(1):1–55, 1984.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):1–26, 2018.

Mantovani, S., Morganti, G., Umang, N., Crainic, T. G., Frejinger, E., and Larsen, E. The load planning problem for double-stack intermodal trains. *European Journal of Operational Research*, 267(1):107 – 119, 2018.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *The American Economic Review*, 85(3): 322–340, 1995.

Milenković, M., Milosavljevic, N., Bojović, N., and Val, S. Container flow forecasting through neural networks based on metaheuristics. *Operational Research*, 1–33, 2019.

Mladenović, N. and Hansen, P. Variable Neighborhood Search. *Computers & Operations Research*, 24(11):1097–1100, 1997.

Morganti, G., Crainic, T. G., Frejinger, E., and Ricciardi, N. Block planning for intermodal rail: Methodology and case study. *Transportation Research Procedia*, 47:19 – 26, 2020.

Nguyen, H., Kieu, L.-M., Wen, T., and Cai, C. Deep learning methods in transportation domain: A review. *IET Intelligent Transport Systems*, 12(9):998–1004, 2018.

Park, J. W., Genton, M. G., and Ghosh, S. K. Censored time series analysis with autoregressive moving average models. *Canadian Journal of Statistics*, 35(1):151–168, 2007.

Pekgün, P., Menich, R. P., Acharya, S., Finch, P. G., Deschamps, F., Mallery, K., Sistine, J. V., Christianson, K., and Fuller, J. Carlson Rezidor Hotel Group Maximizes Revenue Through Improved Demand Management and Price Optimization. *INFORMS Journal on Applied Analytics*, 43(1):21–36, 2013.

Poulos, J. RNN-based counterfactual time-series prediction. *arXiv preprint arXiv:1712.03553*, 2017.

Ritchie, H. and Roser, M. CO2 and Greenhouse Gas Emissions. *Our world in data*, 2020.

Rosenbaum, P. R. and Rubin, D. B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

Sagi, O. and Rokach, L. Ensemble Learning: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), 2018.

Schulze, P. M. and Prinz, A. Forecasting container transshipment in Germany. *Applied Economics*, 41(22):2809–2815, 2009.

Shorten, C. and Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, 2019.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27, 3104–3112, 2014.

Talbi, E.-G. *Metaheuristics: from design to implementation*. John Wiley & Sons, 2009.

Talluri, K. T. and Van Ryzin, G. J. *The Theory and Practice of Revenue Management*. Springer Science & Business Media, 2005.

Torres, R., Chaptal, J., Bès, C., and Hiriart-Urruty, J.-B. Optimal, Environmentally Friendly Departure Procedures for Civil Aircraft. *Journal of Aircraft*, 48(1):11–22, 2011.

Tsai, F.-M. and Huang, L. J. Using artificial neural networks to predict container flows between the major ports of Asia. *International Journal of Production Research*, 55(17): 5001–5010, 2017.

Weatherford, L. R. and Pölt, S. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management*, 1(3):234–254, 2002.

Weatherford, L. and Belobaba, P. Revenue impacts of fare input and demand forecast accuracy in airline yield management. *Journal of the Operational Research Society*, 53(8): 811–821, 2002.

Wieberneit, N. Service network design for freight transportation: a review. *OR spectrum*, 30(1):77–112, 2008.

Winters, P. R. Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, 6(3):324–342, 1960.

Yang, T.-H. Stochastic air freight hub location and flight routes planning. *Applied Mathematical Modelling*, 33(12):4424–4430, 2009.

Zhu, L. and Laptev, N. Deep and Confident Prediction for Time Series at Uber. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 103–110, 2017.

APPENDIX A ARTICLE 1: APPENDIX

Input features of neural networks

Input Layer 1 It is dedicated to external data features such as weather data and temporal context, for all models. We use real observed weather data described in Section 4.5.1, to assess their potential in an optimistic setting in regards to their accuracy. At prediction time, we would then rely on weather forecasts.

All models contain two features as temporal context, that is the week number of the forecasted week and the month number of the Monday of the forecasted week. Weather features, for models that include them, consist in the average daily temperature, the accumulated snow (cm) and the accumulated precipitations (mm) of the forecasted week for the main 17 terminals in the network.

Input Layer 2 Features in Input Layer 2 are lagged observed or forecasted (when doing inferences) demand of commodities predicted by the model. The number of lags was found through hyperparameters optimization, and is given in Table 6.1 below.

Hyper-parameters of neural networks

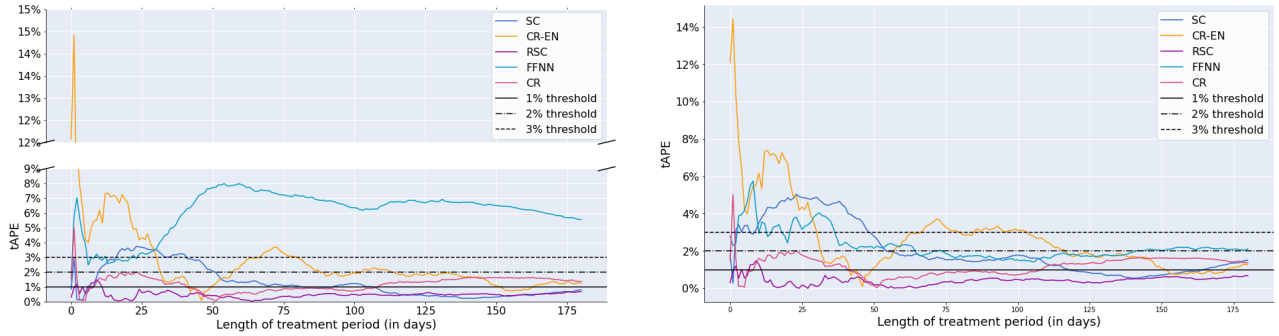
Table A.1 Table of hyperparameters of the neural networks

Model	Lags	Weather Features	Hidden layers	Size Hidden Layers	Dropout	Learning Rate
RNN	3	NO	1	700	0.25	0.1
RNN-W	4	YES	1	260	0.12	0.1
RNN-W-SPLIT1	3	YES	3	460	0.10	0.1
RNN-W-SPLIT2	8	YES	2	220	0.23	0.1
FFNN	3	NO	3	540	0.14	0.01
FFNN-W	4	YES	3	600	0.11	0.1
FFNN-W-SPLIT1	3	YES	2	700	0.06	0.01
FFNN-W-SPLIT2	5	YES	3	580	0.15	0.1

APPENDIX B ARTICLE 3: APPENDIX

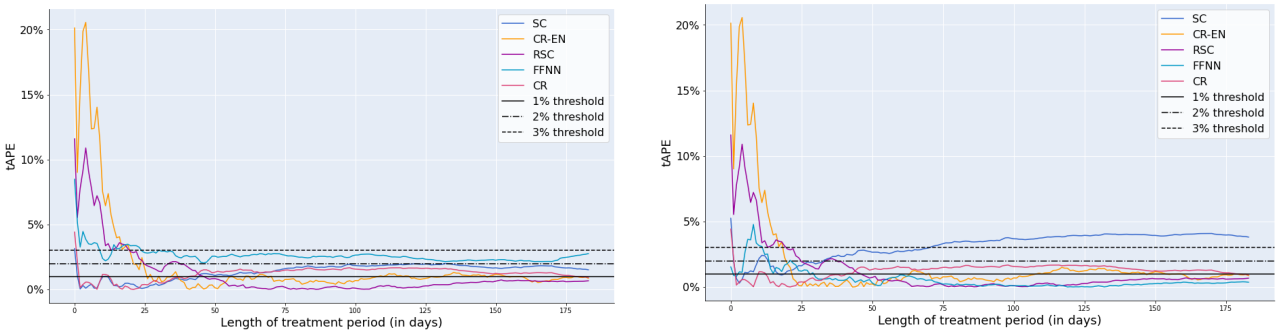
Length of Treatment Period

We present here the results on the analysis of the length of the treatment-period for all pseudo-treatment periods.



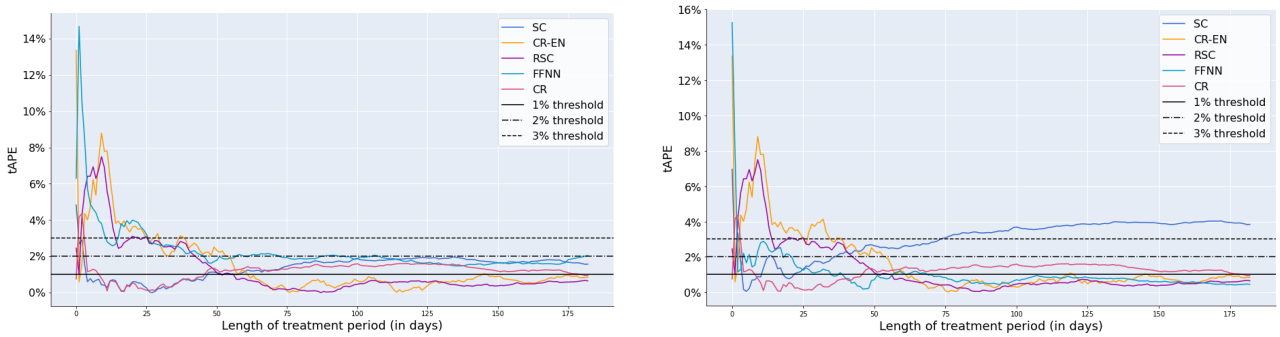
(a) (b)

Figure B.1 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 1 (a) in Setting $S1$ and (b) in Setting $S2$



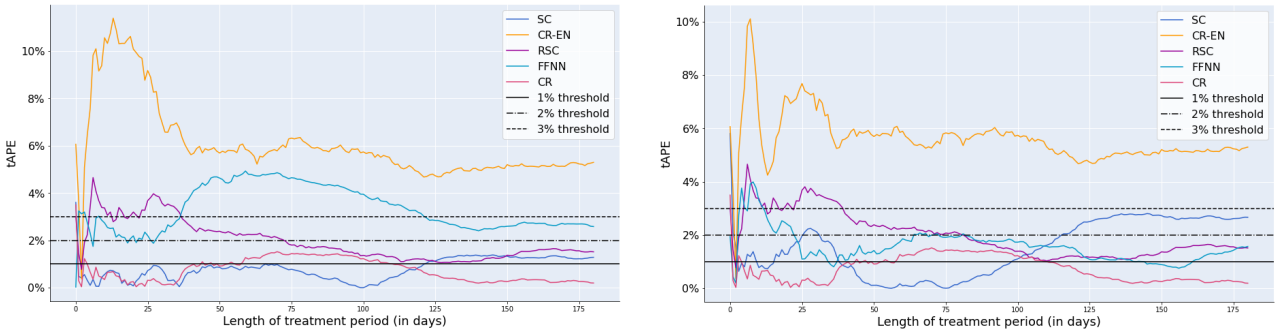
(a) (b)

Figure B.2 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 3 (a) in Setting $S1$ and (b) in Setting $S2$



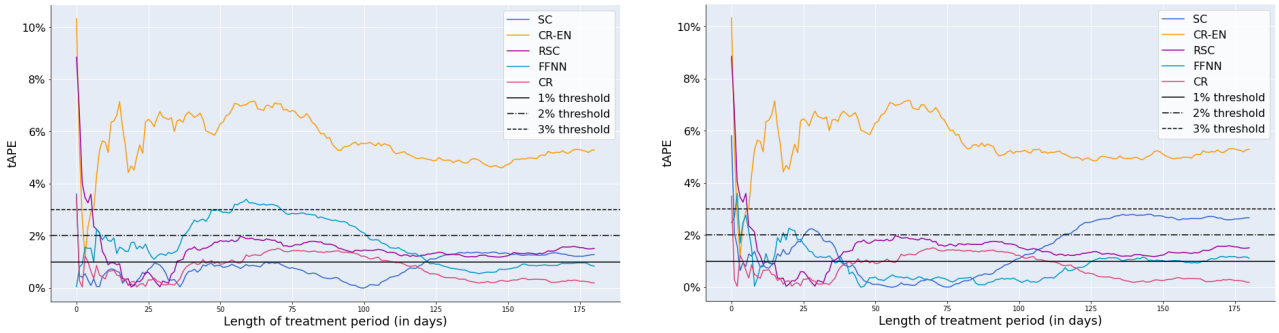
(a) (b)

Figure B.3 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 4 (a) in Setting $S1$ and (b) in Setting $S2$



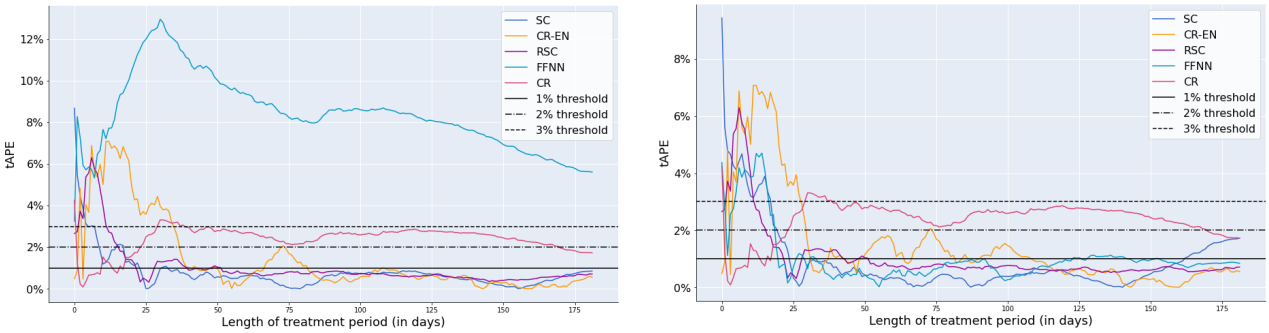
(a) (b)

Figure B.4 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 5 (a) in Setting $S1$ and (b) in Setting $S2$



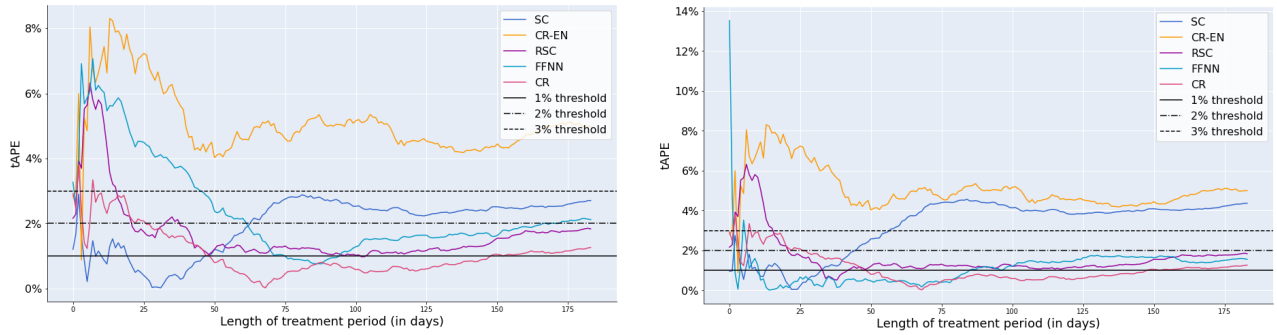
(a) (b)

Figure B.5 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 6 (a) in Setting $S1$ and (b) in Setting $S2$



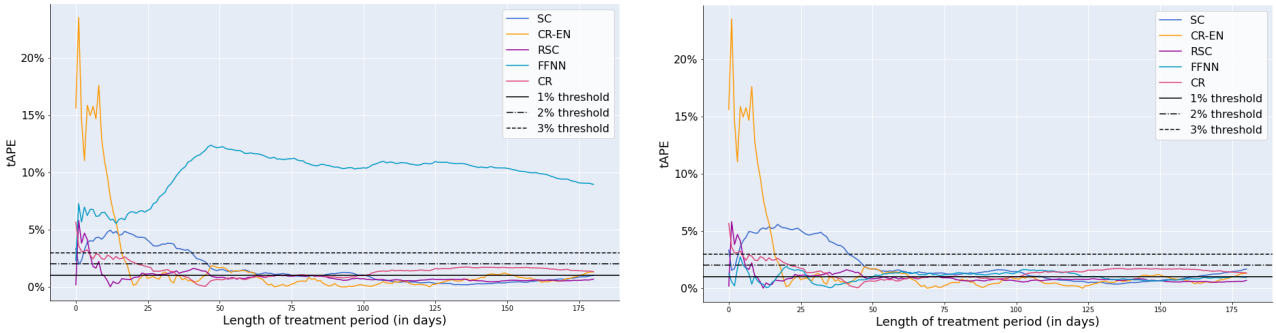
(a) (b)

Figure B.6 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 7 (a) in Setting $S1$ and (b) in Setting $S2$



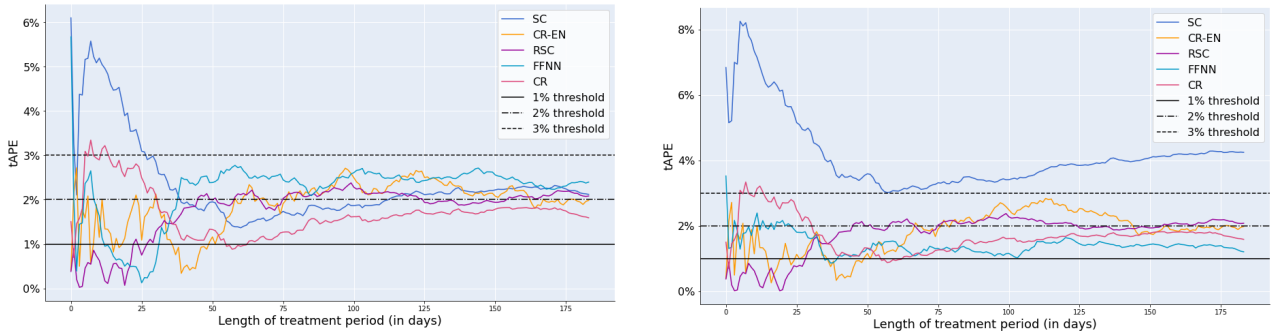
(a) (b)

Figure B.7 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 8 (a) in Setting $S1$ and (b) in Setting $S2$



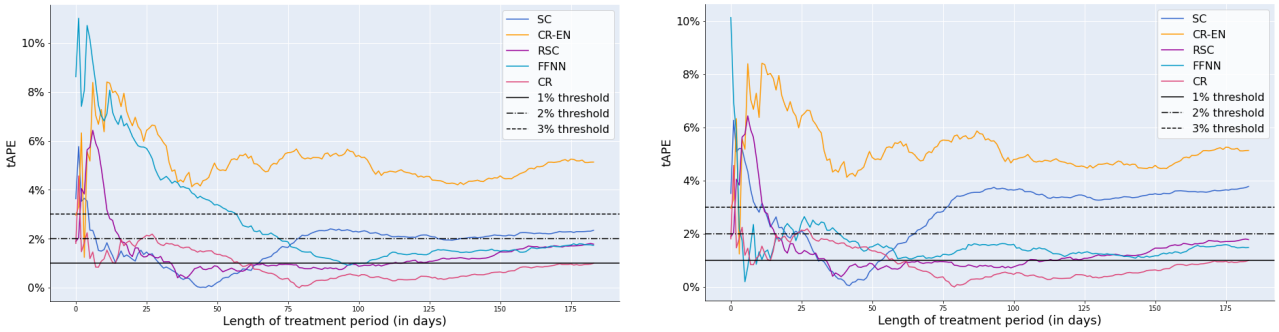
(a) (b)

Figure B.8 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 9 (a) in Setting $S1$ and (b) in Setting $S2$



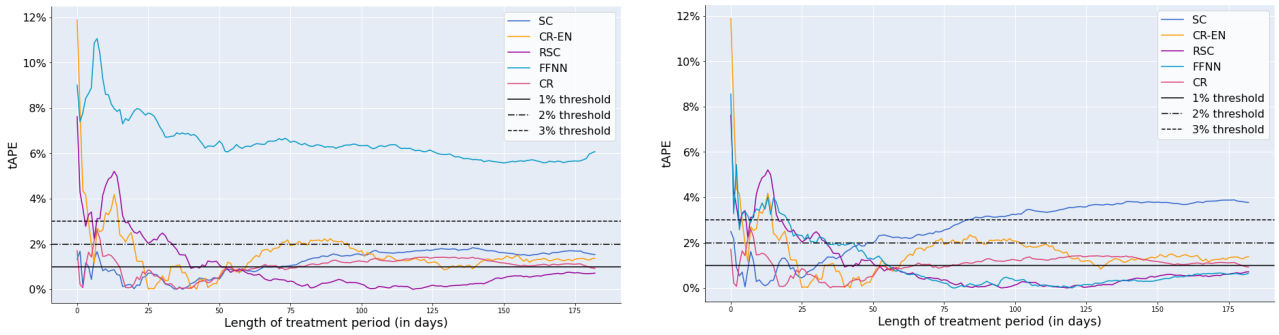
(a) (b)

Figure B.9 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 10 (a) in Setting $S1$ and (b) in Setting $S2$



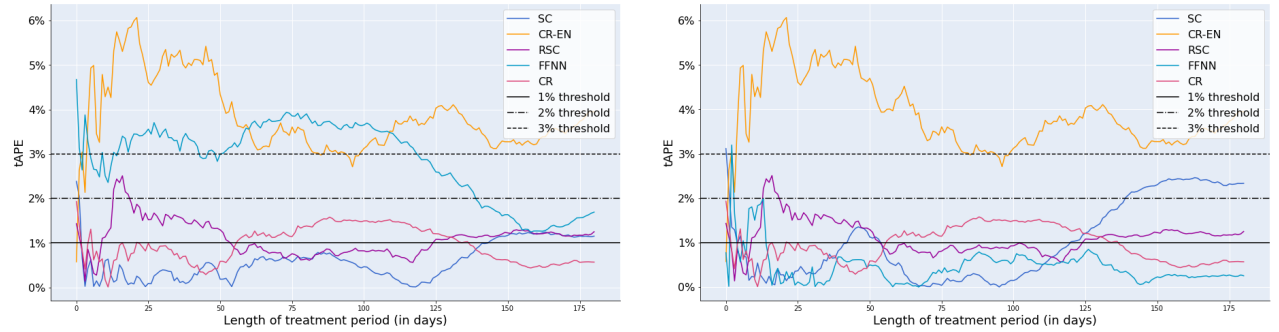
(a) (b)

Figure B.10 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 11 (a) in Setting $S1$ and (b) in Setting $S2$



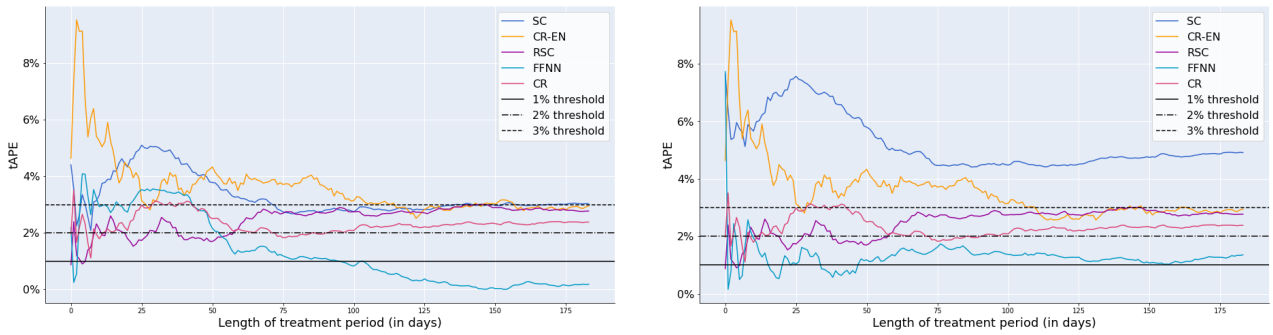
(a) (b)

Figure B.11 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 12 (a) in Setting $S1$ and (b) in Setting $S2$



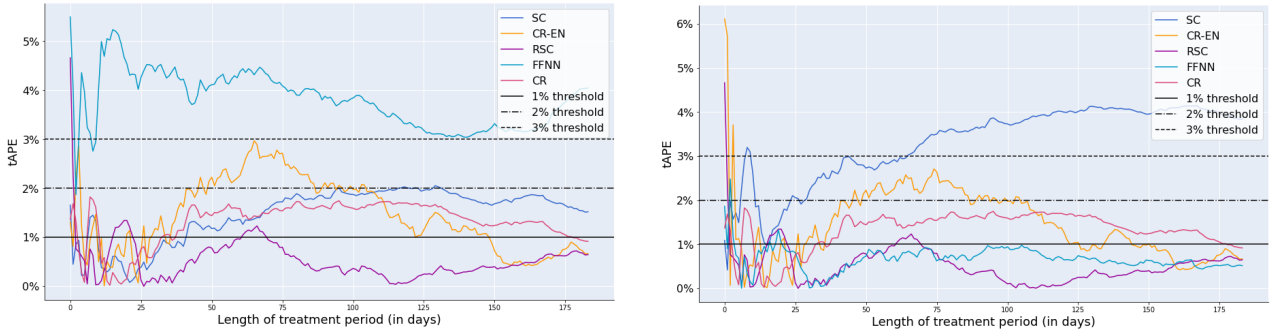
(a) (b)

Figure B.12 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 13 (a) in Setting $S1$ and (b) in Setting $S2$



(a) (b)

Figure B.13 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 14 (a) in Setting $S1$ and (b) in Setting $S2$



(a) (b)

Figure B.14 Values of tAPE varying with the length of the treatment period for pseudo-treatment period 15 (a) in Setting $S1$ and (b) in Setting $S2$