Doctoral theses at NTNU, 2022:302

Harald Støen Høyem

# Essays on the economic efficiency of car ferry crossings

Doctoral thesis

**NTNU**
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Engineering
Department of Civil and Environmental
Engineering

NTNU
Norwegian University of
Science and Technology

Harald Støen Høyem

# Essays on the economic efficiency of car ferry crossings

Thesis for the Degree of Philosophiae Doctor

Trondheim, October 2022

Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

NTNU
Norwegian University of
Science and Technology

**NTNU**
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Engineering
Department of Civil and Environmental Engineering



ISBN 978-82-326-6776-5 (printed ver.)
ISBN 978-82-326-6913-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:302

Printed by NTNU Grafisk senter

# Essays on the economic efficiency of car ferry crossings

Harald Støen Høyem

A dissertation submitted in partial fulfilment of the requirements for the degree of

Philosophiae Doctor at the

Norwegian University of Science and Technology

Faculty of Engineering

Department of Civil and Environmental Engineering

## Contents

## Abstract

Norway maintains a large network of ferries that provide connections across fjords and to islands without road access. Ferries provide accessibility but they also entail costs for the users and the operators involved, such as waiting time costs and labour costs, etc. In an economic sense, it is essential to provide an efficient production of services and select a level of service that yields the largest economic benefit, considering both the cost of users and operators of the service. Thus, both organizational reforms that may promote efficiency in the production and methods which enables assessment of economically optimal service levels are of interest.

The literature on transport economics is scant with respect to ferries. In particular, the optimization of service levels does not consider the specific costs of users and operators in the ferry sector and is limited to a subset of the relevant decision variables. Further, the literature on organizational reforms to promote efficiency in the production of services, is scant when it comes to ferry operations.

This thesis contains four papers that discuss the potential for improving the economic efficiency at car ferry crossings. Two main lines of enquiry are pursued.

**Firstly**, the potential for increased effectiveness by use of Competitive Tendering (CT) is examined. An econometric study based on cost and operational data from Norwegian car ferry crossings is performed. Results show that CT may increase the operational efficiency, by reducing costs. However, the long-term impacts are unknown and should be considered in future work.

**Secondly**, the potential for increased efficiency by selecting economically optimal service levels are investigated in three papers. Two of the papers develop methods to estimate user costs that are specific to ferry transportation. The aim is to develop tools that may be of use when selecting the optimal capacity and service levels of crossings. The methods are applied to case studies of Norwegian crossings. Results show, firstly, that too much capacity is offered and secondly, that the methods used to estimate user costs could be improved even further.

One paper considers wider costs of operating a ferry service through accident costs. It considers whether altering the frequency of a car ferry service could affect the speed of motorists traveling to it. A theoretical model is developed, and a simulation study is performed. Results show that the effect is generally complex. However, it is established that increasing frequency from low levels may increase risky behaviour of motorists. The thesis recommends that empirical studies are performed, possibly using the theoretical model as an interpretative framework.

The results from this thesis demonstrate that there is a potential for increasing the economic efficiency in the operation of car ferry crossings. It is argued that the largest benefit is most likely derived from optimizing service levels, and that the effects of organizational reforms (such as competitive tendering) may be limited. Further, it recommends that more effort is put into developing models of user costs associated with ferry operations, in particular those related to insufficient capacity and understanding the long-term effects of competitive tendering on cost efficiency.

## Acknowledgements

I would like to thank my two supervisors, James Odeck (main) and Svein Bråthen (co) for their invaluable support during the PhD-process, giving highly appreciated advice and guidance along the way.

Further, I would like to thank The Norwegian Public Roads Administration and Asplan Viak for financing my PhD-studies. In particular, I would like to thank Bård Norheim for facilitating support from my current employer, Asplan Viak; financially and for giving me time to work on the thesis.

I would also like to extend my gratitude towards, friends, family and colleagues who supported me along the way. Last, I would like to thank Dina for her patience and support throughout the process.

# 1. Introduction

*In this chapter, the background and motivation for studying the dissertation topic is given, followed by an introduction to the research questions and presentation of the dissertation structure.*

## 1.1 Background and motivation

Governments are often concerned about providing citizens with an efficient and attractive transportation system. In addition, countries differ with respect to geographical distribution and transportation system accessibility. Indeed, in some countries, areas exist where establishing a physical transportation infrastructure is relatively difficult. For example, Norway, Greece and Denmark have many islands on which a percentage of the population lives. In many cases, providing these areas with fixed links (i.e. roads, railway tunnels and/or bridges) is highly expensive. Moreover, these concerns do not necessarily limit themselves to islands; for instance, major roads in Norway cross deep fjords where bridge or tunnel construction is very costly.

Ferries may provide a certain level of accessibility for connecting islands or crossing fjords without having to invest in the physical infrastructure needed for regular road traffic. In fact, they constitute an important part of the transportation network in several countries around the world, including Norway, Denmark, Greece, Scotland and Canada (Baird & Wilmsmeier, 2011). Ferries are also part of many urban environments, such as the Staten Island Ferry in New York, the Washington State ferries or those to Juneau, Alaska (which is only reachable by ferry). The importance of these services in major cities in underlined by Bignon & Pojani (2018).

Next, when a ferry service is used instead of a fixed link, while initial investment costs are lower, final user costs may be higher. This is because users have to wait for their ferry service to arrive, may entail a ticket price and ferries often travel at lower speeds than what drivers may maintain on a road; in addition, ferries may be subjected to cancellations due to bad weather conditions. Consequently, although using a ferry will most likely limit investment costs in many cases, it will also introduce waiting times for road users, which they often perceive as a nuisance (Fosgerau, 2009). Moreover, when users are dependent upon a ferry service, they may also experience inconvenience costs such as having to plan their trip to match departure times as well as having reduced access to a larger labor market, safety services (police, firefighters, etc.) etc. (Díez-Gutiérrez & Tørset, 2019). Demand may also be greater than capacity to the extent that some travellers will be forced to wait for the next ferry departure (Findley et al., 2018). Thus, when ferries are operating normally, they often provide an important public service; at the same time, they also require maintaining a set of costs in contrast with cases where no ferries are required/present.

One way of mitigating the user costs associated with operating a ferry is to provide high-quality service. The constituting components of 'quality' may be subjective and vary according to individual opinion. However, there are a certain number of broad categories that may be quite important to a large set of travellers; these include waiting times to board, time on board the vessel, fares and opening hours in addition to ferries' regularity, punctuality and capacity (Mathisen & Solvoll, 2010), several of which are shared with users' attitudes towards public transport in general (Börjesson & Rubensson, 2019).

Looking at the situation from an economic context, the specific costs for users are normally compared to the broader economic costs of operating at a certain level of service (Asplund & Pyddoke, 2020; Börjesson et al., 2017; Mohring, 1972). Consequently, an important aspect of transportation planning is to provide recommendations regarding what (i) may be viewed as a reasonable level of service and capacity for a particular transportation system and (ii) the most efficient way of achieving this level. As the level of service provided by ferries may be adjusted (i.e., waiting time, speed, capacity, amenities on board etc.), this topic is highly relevant for planners. Indeed, having appropriate tools that enable an efficient use of public resources and a reasonable level of service is important in this respect.

A special feature of car ferries is that they may be considered to be a "hybrid" of public and private transportation. This is because while travelling to or from the vessel, although most users are private motorists, they still need to consider the ferry's schedule; consequently, their own speed is limited by that of the ferry's while on board, which means that ferry travellers are then using a public transport service during their trips. In the context of transportation economics, there have been numerous studies done on the appropriate level of service for different types of public transportation (Fielbaum et al., 2020; S. Jara-Díaz et al., 2017; S. R. Jara-Díaz & Gschwender, 2009; Tirachini et al., 2014; Tirachini & Antoniou, 2020). These studies should be considered relevant to the case of ferries. However, most of them are concerned with more "traditional" modes of public transport (for example, buses and trains); as a result, little attention has been paid to the specific case of ferries, resulting in models that are not directly transferable.

Further, project planners may face budget constraints when choosing a service level. Another way of increasing social efficiency by reducing user costs is to relax budget constraints. One way of doing this is to provide more public resources (if doing so is beneficial), while another is to improve operating efficiency so that although costs are lowered, the same level of service is provided. These savings may even be used to improve the service level. A third is to increase user fares, preferably by means of first or second-best criteria. Finally, competitive tendering has been suggested as one method of increasing efficiency (Sheng & Meng, 2020). However, the literature on competitive tendering is generally scant when it comes to evidence gathered from the ferry sector. The exception to this is Bråthen et al. (2004) who investigated the effects of tendering on six Norwegian car ferry crossings. Having information on the potential of regulatory reforms to increase social efficiency is also valuable to policymakers in the context of ferries.

The aim of this dissertation is to help fill certain gaps in the literature on transport service efficiency by examining the tools and knowledge available to enable a ferry-specific application. There is a special focus on factors that are relevant to the ferry industry, which have not been adequately addressed in the general literature on optimizing public transport services used in the ferry sector.

The following research questions are posed:

- RQ1: How are operational costs affected by Competitive Tendering (CT) in the context of ferries?

- RQ2: What are the optimal capacity and service levels for a ferry crossing?
    - RQ2a: What is the optimal departure frequency when taking into account excess waiting time costs caused by insufficient capacity levels?
    - RQ2b: What is the optimal capacity level and vessel size/frequency rate combination on a car ferry crossing?
    - RQ2c: How are accident rates affected by increasing departure frequency rates on car ferry crossings?

RQ1 presents the effects of subjecting ferry services to competitive tendering and derives estimates of the effects on efficiency from operational costs; effects on both costs and market concentration are calculated. The results provide new knowledge about how organizational reforms may promote increased social efficiency at ferry crossings. These results may be of interest to policymakers when they are considering if crossings should be subjected to tendering or not.

RQ2 concerns developing models to provide estimates of optimal service levels at car ferry crossings. The following three sub-topics are considered in this regard:

- RQ2a develops a method to estimate the cost of users being left behind and how the optimal frequency rate is affected when including this cost, which is likely more important in the ferry sector than, say, bus-related applications.
- RQ2b further develops the method of RQ2a, where only the frequency rate was changed, by estimating optimal capacity that allows the size of each vessel, price and frequency rate to be optimized simultaneously.
- RQ2c investigates whether the optimization of service levels may cause secondary effects, or, more specifically, effects which do not directly affect the service's user cost. It investigates whether altering a ferry service's frequency rate may provide an incentive to drivers to either increase or decrease their speed when driving their car to the quay. As some ferries have a low number of departures per hour, some users might engage in high-speed driving and overtaking other vehicles to reduce their own waiting time by reaching the ferry in time before it leaves. If there is such an effect, the optimal service levels may be affected.

The results of investigating these questions may provide policymakers with a better set of tools and an increased understanding of how social efficiency during ferry crossings can be increased. However, it is also important to consider which questions may be investigated further and how results might be limited.

## 1.2 Dissertation structure

The remainder of the dissertation is structured as follows: Chapter 2 presents a literature review to identify gaps that this dissertation aims to consider. Chapter 3 gives an overview of the research objectives and their relation to the papers in more detail. Chapter 4 provides a theoretical framework which illustrates how the questions fit into a broader scheme of optimizing transport services with respect to social efficiency. Chapter 5 presents the methodology and data used. Chapter 6 provides a discussion of the results and their implications for the research questions. Chapter 7 provides some implications of the research and suggests possibilities for further questions to investigate. Chapter 8 contains the references. Finally, Chapter 9 contains the scientific papers.

# 2 Literature review

*In this chapter, a literature review of the topics relevant to the research questions presented in the previous chapter is provided. Further, gaps in the literature which this dissertation aims to consider are discussed.*

The following three sub-topics are presented:

- **Tendering and costs:** This chapter reviews the scientific evidence of the potential for competitive tendering to reduce the costs of operating a transport service. This chapter is related to **RQ1**.
- **Optimization of public transport and car ferry service levels**: This chapter is mainly related to **RQ2a-b,** which considers the optimization of service levels from an economic point of view.
- **Traffic safety and public transport:** This chapter considers the literature on motorists' speeding behavior as well as economic models of this phenomenon. This chapter is related **to RQ2c**.

Each chapter reviews the relevant literature and identifies knowledge gaps relevant to the present dissertation.

## 2.1 Tendering and costs

Transport planners are often faced with budget constraints when planning a transport service. A great deal of the literature on public transport optimization has investigated the effect of financial constraints on the optimal service level chosen. Jara-Díaz & Gschwender (2009) noted that whenever a transport planner faced a financial constraint, the service levels chosen would be switched from the first-best solution to one similar to the solution chosen by a private operator. Consequently, tighter budget constraints would lead to a lower service level being provided to users. Relaxing the budget constraints was in turn supposed to improve the welfare of public transport users, as the planners would be able to choose a higher quality service level within the constraints faced.

One way of relaxing budget constraints is to increase the operational efficiency to the extent that the same level of quality is produced at a lower cost. The literature on transport economics has highlighted organizational reforms, for example competitive tendering (CT), as a possible method for increasing efficiency (Hensher & Wallis, 2005).

In the context of car ferries in Norway, competitive tendering entails that a contract for operating a service is publicly advertised. Normally, a list of requirements follows (for example, service levels, equipment, etc.). Potential contractors then submit bids, each bid being evaluated based on its price and fulfillment of the advertised requirements. Each contract is then scored, and the bidder who offers the best price and quality is rewarded with a contract for a given period of time. This procedure contrasts with an earlier one in which there is no competition; rather, one contractor negotiated their annual remuneration with the government based on "normative" cost numbers (Odeck & Høyem, 2021).

Theoretically, CT presumes that tendering provides incentives for private contractors to achieve higher cost efficiency, reducing the need for public subsidies or enabling an increase in service quality. Although initial savings are typically achieved, several authors have noted a tendency where costs decrease in the first round of tendering only to increase in subsequent ones (Aarhaug et al., 2018; Hensher & Wallis, 2005; Sheng & Meng, 2020).

Several factors have been proposed to explain this phenomenon, including reduced competition, a first-round winner's curse, increased operational knowledge and strategic bidding to enter the market initially (Hensher & Wallis 2005). When surveying the Norwegian bus industry, Aarhaug et al. (2018) found that the number of bidders significantly affects cost levels. Moreover, contract size was also influential, where larger contracts (in terms of vehicle kilometers) attracted more bidders. Vigren (2020) found that the distance from a bus operator's base facilities to the contract area significantly affected the probability of bidding. Thus, if CT reduces competition, one may expect higher bids on average and, consequently, lower gains over time. Moreover, as there may be a spatial component to submitting bids, competition may be further reduced by geographical factors.

Given the apparent reduction in gains from CTs over time, the literature has noted that policymakers should consider whether CT is the correct policy (Sheng & Meng, 2020). Some authors have proposed "Performance-Based Contracts" in which the contractor and a principal work together to achieve cost efficiency. For example, Hensher & Stanley (2008) noted that the contract's complexity reduces the auctions' efficiency (i.e., CTs). CTs often involve quite detailed requirements with respect to service levels, bus requirements, etc., and have a long-term perspective. These factors increase the probability of contracts being incomplete, which in turn reduces their efficiency, as the principals may be able to specify every contractual obligation in every contingency (Sheng & Meng, 2020). CTs also entail transaction costs, as the process requires specifying the tender and evaluating bids on the part of the principal, while the contractor needs to prepare and submit the bid, all of which may reduce the attractiveness of CT.

Consequently, several factors are of importance when a policymaker must decide which organizational setup to use. In simple economic terms, one may frame the problem as follows: if $C$ is the net cost reduction when using CT as compared to the current contract regime, and $A$ is the net difference in administrative costs, a policymaker will select CT if the difference, $\delta$, is positive:

$$\delta = C - A \tag{2-1}$$

Thus, information on the size of $C$ is important when deciding which contract setup to use. In their recent survey of the literature on CT. (Sheng & Meng, 2020) advise policymakers to not use CT as a default option without considering the case-specific factors. Hensher & Wallis (2005) address the "administrative costs" associated with CT as "typically 5 percentage points of the initial cost savings". Thus, they may not be negligible, and if $C$ is close to 5 percent (using this number for the sake of argument), then CT might not contribute to net savings at all. Thus, it is vital to have information on the size of $C$ for specific applications.

The literature on CT in the area of transportation has focused heavily on the bus sector; however, one notable exception to this is Bråthen et al. (2004) who assessed the change in efficiency rates of Norwegian ferry crossings after tendering. While they found no effect on technical efficiency, their sample size was quite small, having surveyed only six crossings, in connection with a testing scheme operated by the government.

Thus, there is a limited amount of information available regarding the effect on ferry markets. Moreover, some scholars have highlighted circumstances in the ferry sector which may be important to consider. Specifically, Bråthen et al. (2007) considered increasing returns to scale and the access to adequate ferries as two important factors that may create unique circumstances in the ferry sector. They noted that increasing returns to scale may increase market competition over time as large operators underbid smaller ones to gain a greater market share over time. Gaining access to repair and maintenance facilities is especially significant, which may in turn give rise to regional monopolies. Moreover, they argue that companies which have access to older ferries may in fact increase the tender bids. For instance, suppose there are two bidders, one with access to old and less costly ferries (bidder A) and one with only access to newer and more expensive ones (bidder B). If A knows B uses costlier ferries, they may increase their bid up to B's level; as a result, they gain a higher profit margin. Regarding the bus industry, Aarhaug et al. (2018) noted that allowing older vehicles reduces competition, which may indicate that operators which only have access to newer buses do not submit tenders as they know "bidder B", may undercut them. So in a certain sense, there is evidence of similar dynamics in the bus and ferry sector; however, the magnitude of this similarity may differ. Accordingly, it is of interest to increase our understanding of whether similar effects to the ones observed in the bus industry may also be present in car ferry operations.

In summation, it is evident from the literature that (i) empirical evidence with respect to ferry operations is scant (ii) some scholars have argued that ferry operations differ significantly from bus operations, a situation which may influence the outcome of reforms. Given that information on the cost savings gained from CT may be of importance to policymakers when selecting a contractual form, it would be useful to acquire knowledge about the effects of CT on the ferry sector. This is because this type of information will most likely enable policymakers to decide if a certain contractual arrangement is preferable (in an economic sense) to other ones. Several scholars have highlighted the transactional costs associated with CT (Hensher & Wallis, 2005); therefore, this information should be of interest. If CT increased cost efficiency, producing a net savings of transactional costs, the budget constraints in the planner's optimization problem could be loosened, as a result increasing the operation's societal efficiency. Nonetheless, this study does not consider transactional costs, focusing instead on the direct effects of CT. Subsequent studies, or policymakers themselves, may assess these costs at a later date.

### 2.2 Optimization of public transport and car ferry service levels

Car ferry transportation may be viewed as a hybrid of two different transportation modes: 1) private automobile, and 2) public transport. While users travel as regular car users on land, they convert to public transport users when entering the quay area. Consequently, car ferry services share several general features with public transport.

The literature has identified several relevant variables associated with optimizing a transit service, focusing on two specific ones: 1) service frequency, and 2) vehicle capacity, both of which form the basis for this review.

Next, it should be noted that Mohring (1972) may be considered as the founder of public transport economics. He developed the so-called *square root formula* by way of minimizing the sum of user and producer costs. This formula states that frequency should increase by the square root of demand (and not linearly). The intuition behind this result is that a positive externality is conferred upon users from new users entering the system; as these new users enter, more capacity is provided to accommodate them. However, when the capacity level is raised by increasing frequency, the cost of existing users is lowered as well since the waiting time is reduced. Mohring's analysis concerned a typical bus or rail service and assumed the characteristics of an individual period comprised of a homogenous route with passengers boarding at each stop and lasting only one time period. Another limitation was that he did not treat the question of optimal bus size explicitly.

Over time, several scholars have sought to expand on Mohring's work. For example, Jansson (1980) introduced a two-period model for peak and off-peak hours. However, Jansson did so in a restricted manner as he assumed equal frequencies in both periods. A major new contribution to the research was made by introducing bus size as an optimization variable. It was found that (Jansson, 1980, pp. 159-160) by letting bus size vary, the optimal average frequency throughout a given day increased as compared to the one-period case. Moreover, it was shown that smaller buses running more often should be used, as compared to a what he perceived as a typical service at that time.

For their part, Jara-Díaz & Gschwender (2009) sought to explain the apparent difference between optimal and current practices (as perceived by the authors as well as Jansson (1980) by introducing a financial constraint to the planner's optimization problem. Their finding was that tighter budget constraints moved the solution of a planner towards that of a private operator; hence, it was optimal to have a combination of larger vehicles and lower frequency levels. In practice, the problem was analogous to implicitly assuming that the planner has a lower valuation of user benefits than what is factual. Their findings may in some sense provide a provisional answer to the puzzle posed by Jansson (1980) (large vehicles, run seldomly as opposed to his recommendation). However, one conclusion may be that providing sufficient capacity makes it less costly to run a few large vehicles as compared to many smaller ones. Thus, if there is a lack of funding, the cheapest way of providing sufficient capacity is used.

Several other studies have taken place based on Jansson's work. For instance, Jara-Díaz et al. (2017) both modified the assumption of Jansson (1980) and extended the model to treat frequency rate and vehicle size in both periods. In short, their findings showed that frequency rates during peak periods should be higher and vehicles smaller (in contrast to the one-period case). An important assumption is that bus size is equal in both time periods. Thus, using larger buses during peak periods means that these must also run during off-peak periods, which increases the cost of having larger buses. In turn, frequencies increase to the extent that bus size could be lowered while still accommodating passengers during peak periods. Using a high-volume corridor in Stockholm, Sweden as an example, Börjesson et al. (2017) found the opposite result –

buses should be larger and frequency rates lower. Tirachini et al. (2014) found that frequency rates and bus size should increase when crowding costs were included in their analyses of a corridor in Sydney, Australia. Jara-Díaz & Gschwender (2003) suggested that frequency rates should be increased when crowding costs were present based on a theoretical model. Oldfield & Bly (1988) investigated optimal bus size and concluded that smaller buses should be used in the United Kingdom at that time.

Consequently, there are different findings in the literature with respect to optimal bus size and frequency rates and how these findings compare to 'current' practice.

With respect to the present dissertation topic, the main portion of the literature focuses on bus services and passenger transport. Research on the optimization of car ferry services is, however, scant. One could argue that the literature treating land-based transport retains some relevance for ferry services, which may be a valid point. However, several important aspects that are more characteristic of ferry services are not analysed in the literature on land-based operations.

In ferry operations, a service's capacity is important for users (Mathisen & Solvoll, 2010) , as there often exists no other (feasible) alternative route to their destination. If there is too low capacity, queues tend to develop (Findley et al., 2018) . Given this situation, providing sufficient capacity is important. At the same time, providing a high level of capacity also increases operational costs. Consequently, it is vital to consider optimal capacity by comparing user benefits with the cost of provision by using costs that are relevant to ferry crossings.

Consider first how capacity is traditionally measured in the public transport sector. Capacity may be measured in different ways. A textbook on public transport optimization provides several definitions, of which the simplest one implicitly defines capacity (Ceder, 2015):

$$F_j = \frac{\bar{P}_j}{\gamma_j * c} \tag{2-2}$$

Here, $F_j$ (departures/hour) is the frequency required to transport the average maximum number of users during period $j$ ($\bar{P}_j$), with a load factor $\gamma_j$ and capacity of the vehicle $c$. The total capacity is then $c * F_j$, and the relevant question becomes what is the economically optimal value of this variable ? There are two main gaps in the literature, which reduces the transferability of results to the ferry sector.

Firstly, the literature on bus optimization tends to include a capacity constraint in the analyses (Jansson, 1980; Jara-Díaz et al., 2017, 2020; Jara-Díaz & Gschwender, 2003; Jara-Díaz & Gschwender, 2009; Mohring, 1972), with some fixed level of capacity required (for example, to cover average peak load demand). However, if the optimal capacity level is to be estimated, a given level at the outset must not be assumed. This ties in with the notion of *denied boarding*, in which there is insufficient capacity at a given departure, so that users will have to wait. The mathematical treatment of bus optimization literature effectively considers demand to be a fixed size[1] but which in reality varies. Thus, there is a gap in the literature, one result of which is that demand should be treated as a random factor when estimating the optimal capacity. In turn, the relevant question becomes what *the optimal amount of denied boarding* is. If this problem is to be answered, one cannot assume a given level of capacity, as doing so would possibly skew the optimal level of denied boarding.

To highlight the practical importance of determining the optimal amount of denied boarding, the Norwegian Public Roads Administration, which is responsible for a major portion of the country's ferry operations, tracks capacity utilization using the percentage of users who cannot board the first arriving vessel at the quay. In addition, Jørgensen et al. (2007) highlight the importance of considering demand distribution when estimating the capacity required to meet a *specific target* in the ferry industry.

Secondly, the study of user costs related to capacity has mainly focused on onboard crowding in the bus sector (see, e.g., Börjesson et al., 2017; Hörcher & Graham, 2018; Tirachini et al., 2014). Some of these studies do not include a capacity constraint; however, the user costs associated with low capacity in these studies may not be relevant for the ferry sector. As there may be few alternative modes of transportation at ferry crossings, and each ferry has limited capacity, queues may form which are so long that users are not able to board the first arriving ferry (Findley et al., 2018). As regards bus or rail transportation, one may temporarily increase the practical capacity level by increasing onboard congestion. However, in ferry services, this option is not viable for two major reasons. Firstly, vehicles have a naturally different physical stature than humans, and it is therefore more difficult to pack vehicles tightly together. Secondly, ferries operate on water, where there are strict regulations concerning their maximum weight allowance to prevent them from sinking or losing their desired stability level. Having less flexible capacity and fewer alternatives than typical land-based transport means that users need to wait for more than one departure cycle, which is a relatively important element of the ferry transportation sector.

However, research on bus transportation generally treats onboard congestion/crowding as the most interesting variable (Börjesson et al., 2017; Jara-Díaz & Gschwender, 2003; Tirachini et al., 2014) to be studied. Yet including additional waiting time when being unable to board is a relevant element in ferry operations that has not been treated sufficiently in the transportation literature on land-based transport. For instance, although Škurić et al. (2021) investigated optimization of ferry services where the operating company's profit was optimized. However, from an economics perspective, it is the societal costs (including user costs) that are relevant. So, with the exception of these examples, the literature covers neither ferry service optimization nor any related questions to a sufficient degree.

In the specific context of ferry services, Jørgensen & Solvoll (2018) provide an adaption of the work of Mohring (1972) to ferry services where their main contribution is the separation of users' so-called "open" and "hidden" waiting times. However, they do not consider the question of vessel size. To estimate optimal capacity and service levels in

[1] Even though Oldfield & Bly (1988) did in fact discuss this problem, they only provided an explicit expression for bus size and not frequency rate. And while Jara-Díaz & Gschwender (2003) discussed the problem and sketched a model, they did not derive any solution for practical application.

the ferry sector, vessel size should also be included. Both variables are important, a fact which has been recognized in the literature (see, e.g., Jara-Díaz et al. (2017)). The same capacity level may be reached by either having a few large ferries run infrequently, or several small ones run often. The two ways of meeting capacity requirement present different costs to the users and operators, and a model is needed to decide what is optimal from an economic perspective. The current literature does not adequately consider the specific costs of ferry operations (such as denied boarding) in a framework where both frequency and vessel size are considered.

Addressing the gaps in the literature would enable policymakers to use tools that are better suited to analyse the specific costs of car ferry operations by (i) estimating the optimal capacity while (ii) taking ferry-specific costs, such as denied boarding, into account. A contribution of this thesis is to further develop these tools.

## 2.3 Traffic safety and public transport

Traffic safety is an important concern when designing transport systems as they carry high societal costs which policymakers must account for in their decisions on budget allocations (Odeck, 2010). For example, the road authorities in Sweden estimate the societal loss from a fatal accident to be 46 million 2014 SEK[2], and the Norwegian Public Roads Administration places this number at 30 million 2016 NOK[3], with significant figures for non-fatal but serious accidents as well. Thus, road accidents constitute an important cost to society.

In terms of economic accident costs, (Lindberg, 2005) defines three types: (a) private willingness to pay to avoid an accident (b) the cost to relatives and friends from losing a loved one or experiencing a decline in physical functioning, and (c) the cost to the rest of society, including production loss as affected individuals may no longer be able to work. These costs are of interest to policymakers, the question being if they are relevant to the ferry sector.

In the context of ferries, this topic may be of interest for a number of reasons:

- If users have few alternatives to car ferries on their journeys, they may take on risk to avoid having to wait for the next service, especially if the frequency rate is low. Users may select a higher than allowed speed to reach their connection in time. For example, Sadia et al. (2018) noted that road users who are ‘in a hurry’ may engage in speeding.
- If there is limited capacity aboard each ferry, road users may increase their speed to secure a place, for example, overtaking and passing other vehicles, etc.

The literature on traffic safety has in particular considered drivers’ risk estimation and speed limit adherence. A general finding suggests that drivers underestimate accident risks (DeJoy, 1989). Drivers are also found to drive above the speed limit (Haglund & Åberg, 2000). Further, speed is identified as an important factor in traffic safety (Aarts & van Schagen, 2006; Elvik et al., 2019). Thus, it is important to understand the possible mechanisms at play; it is also interesting to consider if altering a ferry’s frequency rate affects drivers’ incentive to speed.

[2] 1 SEK = 0.09 EUR/0.1 USD

[3] 1 NOK is approximately equal to 1 SEK in 08.02.2020.

The literature on optimization of public transport services is limited when it comes to traffic safety, and to the author's knowledge, no study has been completed on the relationship between traffic safety concerns and public transport optimization. Some authors have looked at the interplay between traffic congestion and public transport for bikes (Börjesson et al., 2018) or cars (Tirachini et al., 2014), which may have a connection with accidents. With more congestion and a more heterogenous set of traffic participants, more accidents may take place; however, this possible development is not explicitly treated in the literature.

The literature does contain a number of models showing optimal speed selections made by car drivers (see. e.g., Blomquist, 1986; Jørgensen & Hanssen, 2019; Tarko, 2009). These models typically study the trade-off between driving time and accident risk and consider only one decision variable: speed. This presents a limitation when analyzing the connection between car ferry frequency rates and speed, which reduces the transferability of these models.

In the context of scheduled transport (such as ferries), one could argue that users choose two different variables, (i) speed and (ii) which departure to (attempt to) reach. If one is to apply a model of optimal speed selection where the driver minimizes different cost elements influenced by speed, the current state of the literature is limited when describing drivers' speed selection to reach car ferry services. Bates et al. (2001), suggested this type of a two-stage approach when studying reliability in public transport (they did not consider speed selection).

Investigating how increased frequency rates may influence drivers' chosen speed may be done by extending the single-variable models found in the speed selection literature. Moreover, including this aspect in the optimization problem of transport planners would contribute to the literature by adding a relevant cost component.

This thesis contributes to the literature by extending one-stage speed selection models to treat both choice of departure and speed simultaneously, expanding on the work of Bates et al. (2001). We do not perform any direct empirical evaluation of the effect of changing frequency rates; rather, we use the model in a simulation and analytical study. Having a theoretical framework is beneficial to policymakers for several reasons. Firstly, it may help them interpret the results of possible future empirical studies. Secondly, it may provide guidance on how this type of study may be performed and which caveats to consider. Lastly, it may also give an indication of the actual effect and, more importantly, any parameters which may affect the results.

# 3 Research objectives

*This chapter details the connection between each research question and the papers included in this thesis.*

The main research goal of this thesis is the following:

> *To investigate possible measures enabling increased economic efficiency of car ferry crossing operations when considering organizational and service levels aspects as well as focussing on optimal capacity.*

The overall research question is divided into two main questions. In this chapter the different research questions are presented.

## 3.1 Research questions and their contributions

The aim of this thesis is to provide a better understanding of how optimal service levels at car ferry crossings may be designed. The questions are mainly answered by developing methods that may assist transportation planners in making better recommendations to decisionmakers. Figure 1 shows the different research questions and their individual contributions.

RQ1 investigates empirically how a specific organizational reform (CT) may indirectly cause improved service levels. RQ2 is mainly concerned with developing methods for optimizing the service level. Moreover, RQ2b builds partly on the results from RQ2a.

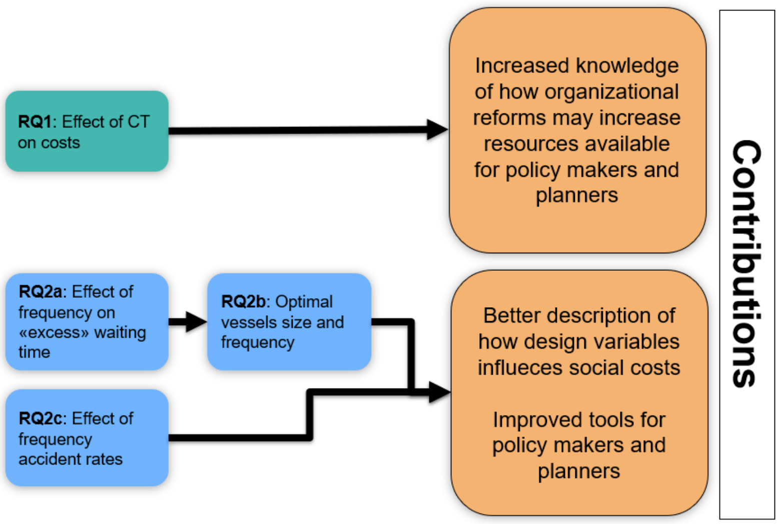



*Figure 1. Research questions, interrelationship and contributions*

## 3.2 Papers

The research questions and how they are related to each paper included in this thesis will now be presented.

### 3.2.1 RQ1 - Paper 1

The first research question asks: How are operational costs affected by Competitive Tendering (CT) in the context of ferries? The paper associated with this question has been published as follows:

*Odeck, J., & Høyem, H. (2020). The impact of competitive tendering on operational costs and market concentration in public transport: The Norwegian car ferry services. Research in Transportation Economics, 100883.*

The first question addresses how cost efficiency levels may be enhanced through a specific type of organizational reforms (CT) using the car ferry sector as a case study. We have gathered data from the NPRA[4] on operational variables and costs with regard to 53 crossings that were observed over 8 years from 2003-2010. Using regression-based panel data analysis, we have tested to see if crossings subjected to competitive tendering lowered their operational costs. We have found they did lower costs at an average rate of about 8 %. However, we have also observed an increase in market concentration and warned that the long-term effects of tendering may be quite different if a monopolistic situation is introduced. The paper includes a discussion of different elements that may or may not influence the long-term outcome.

In relation to research question RQ1, the paper shows that competitive tendering may reduce average operational costs. At the same time, one should be cautious regarding any market concentration increases.

### 3.2.2 RQ2a - Paper 2

Research question 2a asks: What is the optimal departure frequency rate when taking into account excess waiting time costs when capacity is too low? Along with RQ2b and 2c, it addresses the broader question of how the optimal service levels at a crossing may be estimated. The paper associated with this question has been published as follows:

*Høyem, H., & Odeck, J. (2020). Optimal public transit frequency under stochastic demand and fixed vehicle size: Application in the Norwegian car ferry sector. Research in Transportation Economics, 82, 100878.*

The purpose of this paper is to develop the methodological tradition started by Mohring (1972) which takes into account so-called *excess waiting time*. That is, waiting time users incur when capacity is too low to handle all demand at once, resulting in some users having to wait for the next departure before being allowed to board. This is relevant for ferry crossings, where demand fluctuates. We have developed an optimization model that minimizes the sum of user and producer costs. The model has been shown to regard the canonical model of Mohring (1972) as a special case; thus, it is consistent with the already existing methods found in the literature. Further, it has been applied to three separate crossings and shown to be relevant when there are limits on vessel size.

In relation to RQ2a, this paper shows a possible way of incorporating excess waiting time into optimization models where frequency rates are considered.

[4] Norwegian Public Roads Administration

### 3.2.3 RQ2b - Paper 3:

RQ2b expands upon RQ2a, by allowing more variables to be a part of optimization. It asks: What is the optimal combination of vessel size, price and frequency level at a car ferry crossing? The paper associated with this question has been published as follows:

*Høyem, H., & Odeck, J. (2021). Assessing the socially optimal capacity at a selection of Norwegian car ferry crossings. Case Studies on Transport Policy.*

This paper attempts to utilize the work done in paper #2 and expand upon it. A major assumption in paper #2 is that one optimizes frequency rates through given vessel sizes; therefore, its purpose is to develop a subsequent methodology to estimate both optimal capacity and service level. In paper #3, the framework has been carried forward in which frequency, price and vessel size are variables chosen by the planner. Further, in contrast to the current literature, we do not assume any capacity constraints to be binding, a factor which makes the number of vessels a variable, too. This framework enables estimation of optimal capacity when both vessel size and frequency are considered in the context of car ferries. Further, as no «optimal» capacity level is assumed, it is calculated solely on the basis of user costs of excess, open and hidden waiting times. Consequently, optimal capacity is determined endogenously, which contrasts with the current literature on car ferries. A model has been built using appropriate software and parameterized accordingly. We have used 3 ferry crossings as a case study. The results indicate that smaller ferries should be used along with slightly higher frequency rates. The differences between current and optimal levels are quite large and should be reviewed more closely. We perform sensitivity tests on several parameters to assess the model's results further, actions which do not alter the conclusions.

### 3.2.4 RQ 2c- Paper 4:

*Høyem, H. (2022). Public Transport Frequency and Risk-taking Behavior. Economics of Transportation.*

The last paper investigates how second-order effects (externalities) may affect optimal service levels. In comparison to RQ2a-b, this paper does not directly address the cost of using the service, but rather the broader costs associated with its operation, or, more specifically, how accident rates are affected by changing departure frequency rates at car ferry crossings.

This paper addresses the question of how departure frequency rates may influence road users' tendency to speed when driving to a ferry crossing. Some users of ferry services may engage in risky behaviour, e.g. speeding, to reach their ferry connection in time. If frequency rates influence speeding, it means policymakers should consider this effect when determining optimal service levels. We have established a simulation model showing how the users try to reach a transit connection in time. It is assumed that the users are uncertain of the travel time required, thereby selecting the departure that minimizes the sum of accident, time and scheduling costs in addition to the cost of arriving too late. A two-stage procedure has been implemented in which users estimate the optimal speed for each available departure and select the departure yielding the overall lowest costs. The results indicate that increased frequency rates might actually create a greater incentive for risky behaviour and that the effect may be circumstantial and difficult to predict.

In relation to RQ2c, this paper shows that increasing frequency rates may incentivize drivers to act less carefully and drive at higher speeds. However, the result is dependent upon specific parameter values and may differ between users and depending on their preferences, initial frequency and trip length. The paper contributes to the research by highlighting the point that while the effect may be present, it is expected to vary according to circumstance. Thus, researchers conducting empirical studies of the subject should be wary when interpreting and generalizing their conclusions, as these may change depending on the specific environment within which their studies are conducted.

### 3.2.5 Co-authorship

Several of the papers presented in this dissertation were written in cooperation with my supervisor James Odeck. Table 1 details how each paper has been co-authored and how I have contributed to each paper.

*Tabell 1. Summary of contribution for each paper and workload.*

| Paper | Estimated workload | Contribution |
|---|---|---|
| Paper 1 | 50 % | I gathered and structured the relevant data, performed the econometric work, produced the result tables and wrote the majority of the methodology, results and conclusion section |
| Paper 2 | 90 % | I developed the model, wrote a first- and final draft of the paper, collected the data and performed the simulations of optimal frequency |
| Paper 3 | 90 % | I developed the model, wrote a first- and final draft of the paper, collected the data and performed the simulations of optimal frequency and vessel size |
| Paper 4 | 100 % | Sole authorship |

# 4. Theoretical framework

*This chapter develops the theoretical framework used in this thesis. It has two main purposes. Firstly, it gives an overview of the theory used. Secondly, it attempts to unite the different papers together into a single framework, highlighting their connections. It starts with a review of general economic efficiency, then uses a basic model from the literature to illustrate how economic efficiency relates to providing public transport services. This model is further used to illustrate how the four research questions fit into this framework.*

## 4.1 Defining economic efficiency

The theoretical framework used in this thesis is an economic one: A central concept in economics is *efficiency*. Generally speaking, economists seek to find situations which are characterized by efficiency and ways of achieving it. Consequently, it is important to define what is meant by efficiency. In this section, concepts of economic efficiency used in this thesis are discussed.

### 4.1.1 Pareto efficiency

An economy, or society, is made up of individuals. Actions or governmental programs may benefit these individuals differently in terms of both magnitude and direction. Thus, it is important to have a method to characterize the societal impact of decisions that affect society.

A central efficiency concept in economics is called *Pareto efficiency*. According to Varian (1992), this is defined as follows: "A Pareto efficient allocation is one for which there is no way to make all agents better off. Said another way, a Pareto efficient allocation is one for which each agent is as well off as possible".

In general, there are two conditions that must be satisfied in order to achieve Pareto efficiency in a competitive economy (Cowell, 2005, pp. 238):

- Firstly, there needs to be no redistribution of different goods between the agents that would make everyone better off. If this condition is not met, a redistribution (or exchange) of goods between individuals may make someone better off, while not making anyone worse off.
- Secondly, given the profit maximization of firms, there cannot be any remaining possibility to alter their production so that consumers are made better off. If this condition is not met, the firms may reduce their production of one good while increasing the production of another, at which point consumers value the latter good's increase more than the reduction of the former – thereby making them better off.

These conditions will be satisfied in a competitive economy if there are no market failures like public goods and externalities (Cowell (2005), pp. 238; Strøm & Vislie, 2007), a situation known as the "first theorem of welfare economics". If market failures exist, then conditions will be modified.

### 4.1.2 Technical and cost efficiency

The following is a brief explanation of certain simple concepts relating to efficiency in service production. First, a definition of a production function is given, (Coelli et al., 2005, pp. 12)[5], $f$, yielding a certain amount of output $y$, using inputs $a$ and $b$:

$$y = f(a, b) \tag{4-1}$$

Of special interest is the marginal rate of technical substitution (Coelli et al., 2005, pp.16), which is defined as the ratio of each factor's marginal product $f'_x$:

$$MRTS = -\frac{f'_a}{f'_b} \tag{4-2}$$

This quantity measures the number of necessary units of input factor $b$ to add if input factor $a$ is reduced by one unit, for a constant output level. An underlying assumption is that all production is performed efficiently, so that $y = f(a, b)$ equals the maximum amount of output that is possible to produce, given the inputs.

**Cost efficiency**[6] is characterized by a state in which the marginal rate of technical substitution equals the ratio of input prices ($q_x$) (Cowell, 2005)

$$MRTS = -\frac{f'_a}{f'_b} = \frac{q_a}{q_b} \tag{4-3}$$

When this condition is fulfilled, the given level of output is produced at minimum cost. If factor $a$ is reduced by one unit, the amount of factor $b$ needed to maintain a certain level of production costs exactly the same as the savings from reducing $a$.

A central concept when defining **technical efficiency** is the production possibility curve. According to Coelli et al. (2005), the production possibility curve "[…] depicts the various output combinations that could be produced using a given input level". (ibid, pp. 44). Firms that operate on the production possibility curve are said to be technically efficient (Coelli et al., 2005, pp. 3). Thus, if any two "firms"[7] have the same input combination -- but firm A has a lower output than firm B – then B is more efficient than A.

[5] Coelli et al. (2005) used a slightly different notation with an *n*-dimensional vector of inputs, which I have simplified somewhat.

[6] This is sometimes called "allocative efficiency", see e.g., Coelli et al. (2005), pp. 5.

[7] In a more abstract sense, a "firm" could be any actor making decisions through using a production function, i.e., government agencies such as hospitals, etc.

### 4.1.3 Social welfare – Aggregating across users

Up to this point, we have only considered representative 'agents' in the economy (such as individuals and firms). However, to make decisions concerning a larger group of individuals, one needs to aggregate the utility of each one into a single metric.

The Pareto criterion is quite restrictive as no-one can be made worse off. In practice, most government programs have winners or losers. The (strict) Pareto criterion does not permit any such programs to be implemented. As a consequence, it does not provide any easy way to aggregate results across consumers, when there are both winners and losers in a program. To perform these evaluations, the concept of a social welfare function is important (Cowell, 2005, pp. 258), which I turn to next.

To judge how changing economic decision variables affects agents' efficiency in an economic setting, a *welfare function*[8]is often employed which aggregates the *utility*, $\boldsymbol{u_i(x)}$, of each agent $\boldsymbol{i}$ by using a set of weights $\boldsymbol{a_i}$:

$$W_i(\boldsymbol{x}) = \sum_{i=1}^{n} a_i\, u_i(\boldsymbol{x}) \qquad (4\text{-}4)$$

Where $\boldsymbol{x}$ is a vector of variables affecting the utility of each individual. The representation of welfare then depends on two factors: (i) the weights assigned to each individual and (ii) the assumed utility function. Consequently, the answers derived in this dissertation rely upon the specific choices of (i) and (ii). I will now briefly highlight some aspects of this assertion.

In the context of most applications in the transport sector, a so-called *cost-benefit analysis* is used (CBA). CBA uses aggregate Willingness-To-Pay (WTP) in the calculation of total welfare and is applied in the context of optimizing service levels in transportation (see, e.g., Asplund & Pyddoke, 2020; Jansson et al., 2015). Using this framework entails some "hidden" assumptions, all of which are important to highlight.

According to Nyborg (2014), a change in aggregate welfare from a given "reform" and willingness to pay is connected by applying the following equation[9]:

$$dW = \sum_i a_i * u_y' * (WTP_i - C_i) \qquad (4\text{-}5)$$

---

[8] Welfare in this sense is equal to the standard utilitarian approach in which some measure of well-being is aggregated across all of the economy's relevant agents. Other approaches may also be employed, including equal sacrifice (Mill, 1849). Moreover, the function is somewhat simplified as a specific form is assumed here. In a more abstract sense, the function would be written as welfare = W (u1, …, uN) where *u* is the utility of each individual in the economy, see, e.g., Cowell (2005), pp. 258.

[9] The equation is somewhat simplified by assuming a utilitarian welfare function compared to the formulation by Nyborg (2014). The weighting *a* is introduced at this stage in the derivation by the author of this thesis.

Here, $y$ is individual income, $C_i$ is the cost to individual $i$ from a reform[10] and $WTP_i$ is the individual's willingness to pay for this reform. The marginal utility of income is $u_y'$. If one assumes an equal weighting of all individuals, Nyborg (2014) demonstrates that $a_i$ must be equal to:

$$a_i = 1/u_y' \tag{4-6}$$

In essence, this equation implies that CBA weighs rich individuals more heavily than poor ones if the marginal utility of income is smaller for well-off individuals ($u_y'$ becomes smaller the higher $y$). This equation demonstrates that the change in total welfare is only equal to the undertaking's net benefit ($WTP_i - C_i$) when the marginal utility of income is equal for all individuals ($u_y'$). That is, one assumes a dollar to a very poor person matters equally to that of a very rich person. This type of assumption may be deemed unintuitive by some people (Farrow, 1998).

Secondly, there may be winners and losers of a specific reform to the extent that some are, in fact, made worse off. In transportation, these heterogenous effects may be important. For example, while constructing a new road may yield a time-saving benefit to users, residents living close by this road might consider it to be a nuisance because of the increased noise caused by it. One way of handling this obstacle in the context of CBA is to use the so-called *Kaldor-Hicks criterion*, which is an adjusted form of the Pareto-criterion (Hicks, 1939; Kaldor, 1939). The Kaldor-Hicks criterion states that a reform is an improvement if the winners can theoretically compensate the losers, so that a Pareto improvement could take place given appropriate transfers from the 'losers' to the 'winners'. Several authors have been critical of this approach, as some groups may be made worse off (Farrow, 1998); moreover, this criterion requires constant marginal utility of income (Martin, 2019). Others argue that distributional concerns should be resolved through the tax system (Zeckhauser & Hylland, 1979) and not considered in each project. In general, there are different opinions among economists concerning the Kaldor-Hicks criterion.

Further, changes in welfare are often estimated using the consumer's surplus, which is the aggregate difference between willingness to pay for a reform/project and its user cost (this is equal to $\sum_i (WTP_i - C_i)$). Consumer's surplus is only an exact measure of welfare change given a set of assumptions; the income elasticities for different goods (such as transportation) must be equal to one another, and consumer preferences must be homothetic[11] (Silbergberg, 1972). According to Slesnick (1998), these conditions are not likely to be fulfilled. Indeed, as argued by Varian (1992), using the consumer's surplus is often used "as an approximate measure of in consumer welfare in applied work" (Varian, 1992, p. 169). In this thesis, the consumer's surplus is used to make statements about welfare effects. Consequently, it does not represent an exact measure of welfare change, but rather an approximation of it.

[10] A reform in this sense is any project undertaken by the government, e.g., building a road or constructing a new recreational facility.

[11] This is a technical condition which requires that the following condition holds $u(t * x) = t * u(x)$, so gaining $t$ additional amounts of good $x$ is the same as having $t$ extra amount of utility at $x$.

Concerning the specific questions addressed in this thesis, the points made above have some concrete implications. Firstly, distributional aspects are not included explicitly, and based on some reasonable assumptions, may implicitly entail a specific value-judgement. Secondly, even if the reforms studied have a positive net social benefit, some people may be made worse off. The main implication is that the performed estimations are grounded in one specific way of defining and measuring social welfare/efficiency. As economic analysts, it is important to be open about which framework is used because using a different framework may produce different answers.

In the context of car ferries, users may have different WTP to reduce generalized costs, and/or they may have different WTP for specific elements of generalized costs. For example, some users may be very concerned with the cost of not boarding the first arriving vessel (having to wait), while others are less concerned with this cost. Some users may be driving a lorry on a strict time schedule, while others may be on vacation with a looser time schedule. Only the *aggregate change in WTP,* less the cost of using the service (i.e., the consumer's surplus), is considered in this thesis with regard to the *users* of the service.

## 4.2 Economic efficiency in public transportation

The last section reviewed conditions for economic efficiency in general economic terms. In this section, a definition that is more closely linked to transportation economics, in particular public transport, is exhibited. Moreover, the model is used to illustrate how each of the research questions is linked together.

Compared to section 4.1.1-2, there are no individuals, but rather a representative or "average" user, where the aggregation procedure from individual utility to social welfare outlined in 4.1.3 is used. In this setting, total welfare is measured by aggregating the willingness to pay for all users by a representative or "average" user less the production costs.

Moreover, instead of a profit-maximizing firm, there is a government planning agency which chooses the level and composition of outputs produced. That is, there are no private operators in the market, and therefore no competition among them. The transport service is managed by the government through a societal planning agency.

### 4.2.1 A basic model

A simple model of efficiency conditions in the provision of a public transport service will now be presented. This model builds on Jansson et al. (2015) with some adjustments and extensions made that are relevant to this dissertation. Jansson et al. (2015) solved the model for a general production quantity $X$, whereas I split this into different input factors (vessel size and number). This split is important when discussing the optimal capacity and service level at a ferry crossing. The same capacity can be achieved by operating either several small ferries or a few large ones. The model's ultimate purpose is to discuss the economic conditions that characterise efficiency and show how the different research questions are connected.

Let $x$ be the demand level, $\bar{h}(x)$ the inverse demand function[12], $V$ the number of vehicles, $k$ the capacity of each vehicle, $\theta(V,k)$ the generalized cost (excluding price), $\phi(V,k)$ the operating and capital cost and $t$ the time required to perform one round trip along the route. The societal optimization problem is thus the following:

$$\Pi^S = \int_0^x \bar{h}(x)dx - \theta(V,k) * x - \phi(V,k) \\ - \lambda_1[x - V/t * k] \quad (4\text{-}7)$$

The constraint with associated Lagrange multiplier, $\lambda_1$, ensures that there is sufficient capacity to carry $x$ passengers (where trip length is normalized to 1) as $V/t$ equals frequency with $t$ being the total round-trip time. For the purpose of simplicity, we set $t = 1$. We optimize frequency indirectly by selecting the number of vehicles $V$ (assuming they are all used in the operation). Note that in the absence of any marginal cost of public funds, what is paid in fares by users is simply a transfer to the producers and thereafter disappears from the equation (4-7).

In this model, total welfare is measured by aggregating the willingness to pay for the service across all users. There is only one period in the model. One could also assume that the demand level influences the costs directly, by e.g., reducing the average speed as more passengers need to alight and board (Jansson, 1979). However, for the sake of simplicity, this is left out.

The function is optimized with respect to $x, V$ and $k$. The first-order conditions are as follows:

$$\frac{\partial \Pi}{\partial V} = -\frac{\partial \phi}{\partial V} - \frac{\partial \theta}{\partial V} * x + \lambda_1 k = 0 \quad (4\text{-}8)$$

$$\frac{\partial \Pi}{\partial k} = -\frac{\partial \phi}{\partial k} - \frac{\partial \theta}{\partial k} * x + \lambda_1 V = 0 \quad (4\text{-}9)$$

$$\frac{\partial \Pi}{\partial x} = \bar{h}(x)_i - \theta(V,k) - \lambda_1 * t = 0 \quad (4\text{-}10)$$

Realizing that $\bar{h}(x)_i = p + \theta(V,k)$[13], the optimal price may be determined:

$$p^* = \lambda_1 \quad (4\text{-}11)$$

Solving (2) and (3) leads to the following:

$$\lambda_1 = \left(\frac{\partial \phi}{\partial V} + \frac{\partial \theta}{\partial V} * x\right)/k \quad (4\text{-}12)$$

[12] The inverse demand function shows the willingness to pay for the service for the x-th user entering the service.

[13] According to textbook economic theory, in equilibrium, the marginal willingness to pay for the last users (the $x$-th user) equals the (generalized) price.

$$\lambda_1 = \left(\frac{\partial \phi}{\partial k} + \frac{\partial \theta}{\partial k} * x\right)/V \tag{4-13}$$

These equations produce two results. First, the optimal price is equal to the marginal cost of providing an extra unit of capacity ($\frac{\partial \phi}{\partial V}$ or $\frac{\partial \phi}{\partial k}$), less the effect on average user costs $\frac{\partial \theta}{\partial V}$ or $\frac{\partial \theta}{\partial k}$). To attain an explicit solution for all variables, one must also use the constraint in order to find $\lambda_1$[14].

As first argued by Mohring, (1972), the two latter terms are likely to be negative. More users confer a positive externality upon the existing ones, which is called the *Mohring effect*. For example, if a bus service experiences increased demand, more buses may need to be run in order to increase capacity. However, adding another departure affects the costs for all users, not only the marginal ones, by reducing average waiting times. Thus, a positive externality exists. This logic is often used as an argument for subsidizing public transport (Basso & Jara-Díaz, 2010). In the absence of any positive externalities $\left(\frac{\partial \theta}{\partial V} = \frac{\partial \theta}{\partial k} = 0\right)$, optimal prices equal marginal cost – which is also a sufficient condition for full cost recovery, given that a capacity constraint is given. However, when positive externalities exist, optimal prices are less than marginal cost, which does not provide full cost recovery, and hence, a subsidy is required (Basso & Jara-Díaz, 2010).

Secondly, when using first order conditions (4-12) and (4-13), we may combine them to obtain a condition for optimality of input factors. When this condition is satisfied, it is no longer possible to obtain a more efficient use of resources in an economic sense. The condition reads as follows:

$$\frac{V^*}{k^*} = \frac{\left(\frac{\partial \phi}{\partial k} + \frac{\partial \theta}{\partial k} * x^*\right)}{\left(\frac{\partial \phi}{\partial V} + \frac{\partial \theta}{\partial V} * x^*\right)} = \frac{NSB_K}{NSB_V} \tag{4-14}$$

The condition relays the following economic intuition: The optimal ratio of vehicle size to number (i.e., frequency versus size) should be equal to the net social benefit ($NSB$) they provide when evaluated at the optimal level of demand, $x^*$. Consequently, if it is not possible to obtain a lower societal cost by changing the ratio of input factors, then economic efficiency has been achieved.

A simple example of how recommendations on socially optimal provision of public transport may be derived by this framework will now be presented. First, some basic

[14] This can be done by using the constraint to solve for either $V$ or $k$. Then inserting this into (4-14) and solve for equality. Then, finding $V^*$ or $k^*$ (from 4-14 and the constraint) and then $\lambda_1$ from either (4-12) or (4-13), and the solve for $p^*$ for (4-11).

assumptions on the functions are introduced. Subsequently, a shift in one variable and the effect this has on the optimal levels is considered.

No assumptions have been made concerning the functional form of each term in the equation. Some reasonable assumptions may be that $\frac{\partial \phi}{\partial k} > 0$ is concave (there is a falling marginal cost of adding capacity to the ferries), $\frac{\partial \theta}{\partial V} < 0$ convex (the time savings from adding another departure diminishes as frequency increases), $\frac{\partial \phi}{\partial V} > 0$ is concave (there is some return to scale by having a larger number of vessels, e.g., by reduced administrative costs and $\frac{\partial \theta}{\partial k} < 0$ is convex (more capacity is desirable, but with diminishing returns).

Now assume that for some reason, the net societal benefit of vehicle size is reduced, having been evaluated at a previous optimum. For example, $\frac{\partial \theta}{\partial k}$ may be shifted down (it is now more negative) as users become more averse to the presence of others (e.g. an increased aversion to congestion). This results in the following:

$$\frac{V}{k} > \frac{\left(\frac{\partial \phi}{\partial k} + \frac{\partial \theta}{\partial k} * x^*\right)}{\left(\frac{\partial \phi}{\partial V} + \frac{\partial \theta}{\partial V} * x^*\right)} \tag{4-15}$$

This situation means adjustments will have to be made for efficiency to be retained. What is optimal depends on the curvature of all functions, and the response of demand to changes made in service levels. One possible interpretation is to increase ferry size (reduce the absolute value of $\frac{\partial \theta}{\partial k}$) and/or increase the frequency/number of vessels (reduce the absolute value of $\frac{\partial \theta}{\partial V}$). The optimal values then depend on preferences and production technology.

Moreover, the price may also be changed. Recall that the optimal price was given by (insert (4.11) into (4.12) or (4.13) for):

$$p^* = \left(\frac{\partial \phi}{\partial k} + \frac{\partial \theta}{\partial k} * x\right)/V \tag{4-16}$$

If one uses larger vessels to the extent that $\left|\frac{\partial \theta}{\partial k}\right|$ becomes smaller, then prices may either increase as well to provide an increase in the capacity available or stay the same, depending on the new equilibrium value of $\frac{\partial \phi}{\partial k}$.

Consequently, finding optimal capacity and service levels depends on how the different functions are specified, e.g., which assumptions are made concerning the user and producer costs. Moreover, in the above framework, demand has been treated as a static factor; yet in real applications, demand will most likely vary if service levels are changed. Ideally, all of these factors should be accounted for. In order to compute optimal values in real applications, models and numerical procedures usually need to be applied, which is one of the aims of this dissertation.

I now use the framework to highlight the relationship between economic efficiency in public transport and each of the research questions.

## 4.3 The relationship between each research question and efficiency

Having defined the basic conditions for economic efficiency when providing a public transport service, I now move on to illustrating how the different research questions are related to the dissertation's main topic. The model presented thus far is highly abstract – there is little explicit meaning connected to each equation. However, having an abstract framework also allows flexibility where extensions can be made. I now introduce different aspects into the framework for the purpose of connecting each research question to the theoretical framework.

### 4.3.1 RQ1: Financial constraints

RQ1 concerns the effect of tendering on operational costs. The question is interesting, as having lower operational costs may yield more resources available for providing a high-quality public transit service. I now show in what way RQ1 affects economic efficiency.

If we revisit the definitions of technical and cost efficiency, we see that cost efficiency is the same as minimizing costs for a given level of production. If a service is not run in a cost-efficient manner, the cost level $\phi(V,k)$ may be lower for the same level of inputs as compared to what is possible. If tendering increases cost effectiveness, then the cost level pre-tendering, $\phi_{PRE}(\bar{V},\bar{k})$ will be larger than the one for post-tendering $\phi_{POST}(\bar{V},\bar{k})$ for the same level of inputs $\bar{V}$ and $\bar{k}$. Another way of stating this is that budget constraints are relaxed.

To illustrate the effect on economic efficiency, the basic model developed in the preceding chapter is expanded upon by adding a financial constraint stating that profits ($p * x - \phi(V,k)$) should equal an externally given level, $C$[15]. Typically, $C < 0$ in public transportation, which indicates that a subsidy is being provided.

Adding the constraint with Lagrange multiplier $\lambda_2$ produces the following:

$$\Pi^S = \int_0^x \bar{h}(x)dx - \theta(V,k) * x - \phi(V,k) - \lambda_1[x - V * k] - \lambda_2\big[\big(p * x - \phi(V,k)\big) - C\big] \tag{4-17}$$

First-order conditions for $x, V$ and $k$ are as follows[16]:

$$\frac{\partial \Pi}{\partial V} = -\frac{\partial \phi}{\partial V}(1+\lambda_2) - \frac{\partial \theta}{\partial V} * x + \lambda_1 k = 0 \tag{4-18}$$

[15] Strictly speaking, one often requires that subsidies do not exceed a certain level. However, for the sake of simplicity, I assume equality here to highlight the main points.

[16] Again, to find explicit values, one must also utilize the constraints to solve for $\lambda_1$ and $\lambda_2$. I abstract from this to highlight the principles.

$$\frac{\partial \Pi}{\partial k} = -\frac{\partial \phi}{\partial k}(1+\lambda_2) - \frac{\partial \theta}{\partial k} * x + \lambda_1 V = 0 \tag{4-19}$$

$$\frac{\partial \Pi}{\partial x} = \bar{h}(x)_i - \theta(V,k) - \lambda_1 - \lambda_2 * p = 0 \tag{4-20}$$

The optimal price is now given by the following equation:

$$p = \frac{\lambda_1}{1-\lambda_2} = p_{FB}^* * \frac{1}{1-\lambda_2} \tag{4-21}$$

Thus, if a financial constraint is present and binding, the optimal price will be equal to the first-best price in addition to a mark-up proportional to the Lagrange multiplier $1/(1-\lambda_2)$. In the end, the presence of a financial constraint moves the optimal price solution away from the first-best price, which in turn reduces the economic efficiency attainable.

The Lagrange multiplier has a distinct interpretation in an economic context, as the increase in welfare from attaining a higher budget constraint:

$$\frac{\partial \Pi}{\partial C} = \lambda_2 \tag{4-22}$$

Thus, the sign of the Lagrange multiplier is informative regarding the optimality of the budget constraint. If this constraint is binding and $\lambda_2 > 0$, societal welfare could be increased from relaxing the budget constraint. Consequently, if tendering increases efficiency to the extent that more funds are available, then social welfare can be increased by spending more on ferry services, if $\lambda_2 > 0$. Moreover, it should also be noted that if $\lambda_2 = 0$, there is no gain from relaxing the budget constraint. If $\lambda_2 < 0$, then too many resources have been allocated to the service.

In general, the presence of a financial constraint moves the solution away from a first-best situation and towards a second-best one, a movement that may also have consequences for the optimal service levels. The condition needed for achieving economic efficiency with a financial constraint present is given by:

$$\frac{V^*}{k^*} = \frac{\left(\frac{\partial \phi}{\partial k}(1+\lambda_2) + \frac{\partial \theta}{\partial k} * x^*(p_{FB}^*, \lambda_2)\right)}{\left(\frac{\partial \phi}{\partial V}(1+\lambda_2) + \frac{\partial \theta}{\partial V} * x^*(p_{FB}^*, \lambda_2)\right)} \tag{4-23}$$

$x^*$ is a function of $\lambda_2$ as it is dependent upon the price, and $\lambda_2$ is dependent on the "strictness" of the budget constraint $C$.

First, it should be noted that if $\lambda_2 = 0$, this produces the first-best solution given in the last section. Recall that $\lambda_2 > 0$ indicates a positive welfare effect of increasing the budget constraint, and further, that, $\lambda_2 = 0$ characterizes the first best. Next, increasing $|C|$ (given $\lambda_2 > 0$) moves $\lambda_2$ towards zero, and hence, the optimal allocation towards the first-best one. A similar point was made by Jara-Díaz & Gschwender (2009), who found that a societal planner facing a budget constraint would act more like a private operator, moving away from the first-best allocation. Hence, relaxing the budget constraint increases societal welfare, given that $\lambda_2 > 0$.

Thus, if one is able to relax the financial constraints by running a more efficient operation, both prices and production inputs will move closer to the first-best case. Hence, societal efficiency may increase. Consequently, RQ1 directly relates the societal efficiency through its effect on the shadow value of increased revenue/reduced cost, $\lambda_2$.

### 4.3.2 RQ2

RQ2 concerns the optimal service level at car ferry crossings. The efficiency of the service requires that all relevant costs and benefits are accounted for in the optimization. Specifically, the questions consider the following:

- RQ2a concerns the specification of user costs in the context of ferry operation. The variable excess waiting time, which arises when capacity is too low, is introduced. In the context of efficiency, it is important to include factors that may affect users' utility ($u$).
- RQ2b investigates how the optimal capacity may be estimated when vessel capacity, price and frequency vary. In this context, efficiency concerns whether the appropriate level of capacity has been chosen, so that the marginal effect upon welfare is equal to the marginal production cost.
- RQ2c looks at the presence of externalities when optimizing a car ferry service. More specifically, if car users maintain high speeds to reach the ferry in time while not internalizing the societal cost of accidents, an externality exists. The question asks if public transport services could help reduce the externality to attain a more efficient outcome.

Each question will now be examined in more detail.

#### *4.3.2.1 RQ2a Excess waiting time*

RQ2a, addressed in Odeck & Høyem (2021), regards how waiting time is affected when capacity is lower than demand. Thus, some users may have to sit out one departure. The contribution is to alter the function $\frac{\partial \theta}{\partial V}$ (the effect on generalized cost from increased frequency) by including the "excess" waiting time, which arise as users are not able to board their preferred departure, having to "sit back". Note that only $\frac{\partial \theta}{\partial V}$ and not $\frac{\partial \theta}{\partial k}$ is changed, which is addressed in RQ2b. Consequently, RQ2a investigates optimal frequency when excess waiting time is included and given a fixed vessel size, $k$.

Thus, $\theta(V, k)$ is expanded so that:

$$\theta(V;\bar{k}) = OW(V) + HW(V) + EW(V;\bar{k}) \tag{4-24}$$

Where $OW$ is open waiting time, $HW$ is hidden waiting time and $EW$ is excess waiting time, which has been added to the equation. Note that $k$ is kept constant at $\bar{k}$. It is assumed that $EW$ is convex in both $V$ and $k$; thus, excess waiting time decreases in both frequency and capacity.

RQ2a, only considers the effect of frequency on costs, as $k$ is kept constant. Consequently, RQ2a investigates how the function $EW(V;\bar{k})$ may be formulated, in turn applying it to a certain number of test cases.

#### *4.3.2.3 RQ2b Optimal capacity*

RQ2b concerns how the optimal capacity at a ferry crossing may be found; this starts with a modification of the model to enable theoretical conditions for optimal capacity to be derived.

When optimal capacity is to be found, two modifications are made to the problem. First, there is no capacity constraint, as I aim to find the optimal capacity exogenously. Secondly, the user cost is now dependent on the demand itself. This is to be interpreted as general crowding, or capacity effect, which is positive for the higher the level of demand for a fixed level of inputs. Producer cost is now directly affected by demand. For a given level of frequency and capacity, increased demand raises the producer cost as average speeds drops when more passengers need to board and disembark.

The revisited problem now reads:

$$\Pi^S = \int_0^x \bar{h}(x)dx - \theta(x,V,k) * x - \phi(V,k,x) \tag{4-25}$$

With first-order conditions as follows:

$$\frac{\partial \Pi}{\partial V} = -\frac{\partial \phi}{\partial V} - \frac{\partial \theta}{\partial V} * x = 0 \tag{4-26}$$

$$\frac{\partial \Pi}{\partial k} = -\frac{\partial \phi}{\partial k} - \frac{\partial \theta}{\partial k} * x = 0 \tag{4-27}$$

$$\frac{\partial \Pi}{\partial x} = \bar{h}(x)_i - \theta(x,V,k) - \frac{\partial \theta}{\partial x} * x = 0 \tag{4-28}$$

From the conditions, one may derive two optimality criteria. First, the optimal price is given as follows:

$$p^* = \frac{\partial \theta}{\partial x} * x + \frac{\partial \phi}{\partial x} \tag{4-29}$$

The left-hand term is the marginal cost on other users (i.e., crowding effect) one new user inflicts multiplied by the total number of current users, $x$. The right-hand term is the marginal cost that one additional user inflicts on the producer; that is, as the increased time spent waiting for boarding and disembarking. It is similar to the *short-run* price given by Jansson (1979) (equation 6), on the basis of Mohring (1972). In the above model, although capacity may vary, it does not enter the price equation explicitly as there is no capacity constraint. However, in the *long-run* price where capacity can be adjusted, the producer costs of capacity are also included by some authors (Holmgren, 2014; Jansson, 1979; Jansson et al., 2015). However, Börjesson et al. (2017) found a similar expression to the one presented above where only the externality conferred upon other users was included, and no capacity constraint was present.

The main difference between the different results lies in whether or not one (a) includes a capacity constraint and/or (b) assumes that users generate external costs. When a capacity constraint is included, the optimal price equals the shadow value of capacity (the Lagrange multiplier) (Jansson, 1979). However, when no such constraint is present, then only the externalities are pertinent. These externalities may either be crowding costs (Börjesson et al., 2017) or increases in production cost as more travellers mean lower operating speeds (Jansson, 1979), or other elements that may be argued to be relevant.

Optimal capacity is determined on the basis of marginal benefit versus marginal cost. The cost of capacity is implicitly included in the price by the term $\frac{\partial \theta}{\partial x}$, which is the external cost that is conferred upon each user when entering the transport system. That is, each user that enters the system reduces the probability of all other users being able to board. In the case of crowding, each user increases the cost of other users as crowding increases.

The condition for optimality in the input mix is now reduced to:

$$MC = \frac{\partial \phi}{\partial V} - \frac{\partial \phi}{\partial k} = -x\left(\frac{\partial \theta}{\partial V} - \frac{\partial \theta}{\partial k}\right) = MB \tag{4-30}$$

That is, the difference in marginal cost between the two input factors should equal the difference in marginal reduction in travel cost (less the price) totalled over all users, $x$.

It is interesting to find the optimal capacity, and by including a capacity constraint, the answer to the question may be effectively assumed outright. In this sense, the trade-off is not as much between the marginal cost of adding another user to the system on operating cost. It is more relevant to consider the trade-off between user cost from crowding $\frac{\partial \theta}{\partial x}$ versus a lower level of capacity, either through $V$ or $k$.

The question may now be asked: Why is this so? For a given level of demand, the planner may either accrue the cost to users by lowering their utility and keeping the capacity unchanged or accrue the cost to society at large by increasing capacity. To find the optimal capacity, we cannot assume a given level by including a constraint. Instead, we must include the costs from having a capacity level that is too low.

When a capacity constraint is included, one effectively assumes the planner guarantees a certain service quality. For example, that users will have a journey devoid of any user-generated externalities through $\frac{\partial\theta}{\partial x}$ – that no one must wait for the second departure or stand – or at least only an *externally* determined proportion must wait. However, using a constraint does not answer the question as to what the optimal proportion is, but rather merely assumes a value. The purpose of RQ2b is to find this value exogenously.

Naturally, the user needs to pay for the guaranteed service level somehow, which is reflected in models where a capacity constraint is included. However, when no quality of service (in terms of capacity) is guaranteed, this may be interpreted as the user not being obliged to pay for a certain level of quality, instead "paying" through two different channels:

Directly: The prices levied that reflect the external costs they impose on users $p^*$.

Indirectly: Crowding disutility included in the generalized cost increase whenever demand increases. With a service quality guarantee (capacity constraint), one pays directly for this guarantee as increased demand directly yields a higher capacity. Without a service quality guarantee (capacity constraint), one pays indirectly through increased costs as not everyone will be able to board at their preferred time.

RQ2b concerns how the optimal capacity may be found. Previously in RQ2a, a model was developed that kept the capacity of each vessel constant and only looked at how frequency affected the cost of not being able to board. In RQ2b, there are two additions that allow the estimation of the optimal capacity in full. First, no capacity constraint is assumed to be present. Secondly, both frequency and vessel size affect the cost related to being unable to board. Revisiting the definition of user cost, $k$ is no longer a fixed variable $\bar{k}$, but allowed to vary:

$$\theta(V,k) = OW(V) + HW(V) + EW(V;k) \tag{4-31}$$

Thus, $\frac{\partial\theta}{\partial V}$ both $\frac{\partial\theta}{\partial k}$ are different from zero. Further, as vessel capacity can vary, $\frac{\partial\phi}{\partial k}$ is also different from zero. Consequently, RQ2b investigates how optimal capacity may be achieved when optimizing size and frequency simultaneously. This is important as both variables are relevant for the total capacity; moreover, it is done by excluding the capacity constraint.

The functions $\frac{\partial\theta}{\partial V}$ and $\frac{\partial\theta}{\partial k}$ have been investigated in the literature, for example by (Börjesson et al., 2017, 2019; Jara-Díaz & Gschwender, 2003; Tirachini et al., 2014)). However, all these applications are related to the bus industry, where the crowding effect is important. The overall carrying capacity of the service offered is most likely more important in ferry operations. Thus, there is a gap in the literature where specification of $\frac{\partial\theta}{\partial V}$ and $\frac{\partial\theta}{\partial k}$ is necessary in order to use the principles outlined above in practice in ferry services.

#### *4.3.2.2 RQ2c Optimal frequency and traffic safety*

RQ2c concerns traffic safety's interplay with the service level at a car ferry crossing, or, more specifically, if some drivers speed to catch the ferry. Safety researchers have noted a tendency in drivers to select higher speeds when in a hurry to reach an appointment (Sadia et al., 2018). Reaching a ferry in time is a somewhat similar situation, as there is often no alternative other than waiting for the next departure. RQ2c is addressed by creating a speeding model and its relationship with a ferry service's departure frequency. A simple extension of the basic model to illustrate the connection with economic efficiency will now be presented.

An external accident cost term $\epsilon(V, k; \gamma)$ has been added to the optimization problem, so that the extended version becomes:

$$\Pi^S = \int_0^x \bar{h}(x)dx - \theta(V, k) * x - \phi(V, k) - \epsilon(V, k; \gamma) - \lambda_1[x - V/t * k] \tag{4-32}$$

The external accident cost term will now be explained in more detail. The term reflects accidents' cost inflicted upon society by drivers who engage in risky behaviour when trying to make a specific ferry departure. It is assumed that the externality is dependent on $V$ and $k$;however, I do not assume any specific direction in which the two variables influence the externality.

I start by defining the parameter $\gamma \in [0,1]$, which reflects the degree to which car drivers have internalized society's marginal cost of accidents. If $\gamma = 1$, the costs are fully internalized, meaning that accident risk is already optimal from society's point of view. When drivers fully internalize the social cost, their marginal gain from engaging in risky behaviour (by e.g., speeding) equals society's marginal cost. As such, there is no place for any planner's intervention to improve societal efficiency. However, if $\gamma < 1$, then it is interesting to consider the influence of $V$ and $k$ on $\epsilon$, as drivers do not fully internalize the societal cost of accidents. To derive conditions for $\gamma < 1$, I start by defining the expected accident cost ($EC$) as a function of speed as follows:

$$EC(s)_U = \alpha * \pi(s) * \beta * c \tag{4-33}$$

Here, $\pi(s)$is the objective probability a of having an accident, given a speed, $s$. Further, $\alpha$ is perceptive bias in estimating the probability by the driver. Thus, $\alpha * \pi(s)$, is the subjective probability the driver faces. The parameter $c$ is defined as the cost per accident (of 'abstract' severity), composed of private and external cost:

$$c_{Social} = c_{Private} + c_{External} \tag{4-34}$$

The private cost of accidents is the Willingness to Pay to avoid an accident (Lindberg, 2005). The external costs of accidents are loss of production output and the emotional loss to family and friends (Lindberg, 2005). $\beta \in [0, \infty]$ is defined as the subjective bias in societal cost estimation. If $\beta < 1$, when drivers underestimate the societal cost, and vice versa, if $\beta > 1$, costs are overestimated. The external costs are only relevant in an economic sense when they are not internalized by the driver (Lindberg, 2005).

The driver is assumed to select an optimal speed (in km/h) when trying to reach a ferry connection, defined by the function $s^*(V,k)$, which is assumed to be dependent upon frequency and capacity. That is, a higher frequency level may influence the speed by making it less detrimental having to wait for the next departure, as headway is lower. Moreover, drivers may want to overtake other drivers if there is a known shortage of capacity at the ferry in order to be “first in line”. These are just some motivating examples to illustrate the mechanisms behind selecting optimal speed.

It is assumed that $s^*(V,k)$ is the solution to a driver's utility maximization problem, where $V$ and $k$ are taken as the following:

$$s^*(V,k) = \text{argmax}\, u(s; V,k) \tag{4-35}$$

Typically, lowering speeds increases travel time costs, and increasing speed increases accident costs, both of which affect total utility $u$. I do not introduce any further assumptions on $u$ as this is the purpose of RQ2c (and done in paper 3).

Assume that total internalization of accident costs by the driver is now given by:

$$\gamma = \alpha * \beta \tag{4-36}$$

Thus, both the perception of accident risk and cost are relevant for the internalization of societal costs of accidents by drivers. We may now write the expected total societal cost of accidents as ($EC(V,k)_S$):

$$EC(V,k)_S = \pi\big(s^*(V,k)\big) * c \tag{4-37}$$

And the total user costs as ($EC(V,k)_U$)

$$EC(V,k)_U = \gamma * \pi\big(s^*(V,k)\big) * c \tag{4-38}$$

Where I have inserted for $s^*$ by using the fact that optimal speed is determined by $V$ and $k$. Thus, the external cost is given as:

$$\epsilon(V,k;\gamma) = EC(V,k)_S - EC(V,k)_U = \pi\big(s^*(V,k)\big) * c * (1-\gamma) \tag{4-39}$$

If $\gamma = 1$, the net external costs are zero and irrelevant for the optimization. This happens if drivers perfectly internalize the social costs of accidents; that is, they have a perfect estimation of accident probability and societal cost in mind when selecting their optimal speed. However, if $\gamma < 1$, then the external cost term is greater than zero. Consequently, it may be influenced by altering the service level variables $V$ and $k$.

Solving the optimization problem with the externality yields the following efficiency condition of inputs to the service level of a ferry crossing:

$$\frac{V^*}{k^*} = \frac{\left(\frac{\partial\phi}{\partial k} + \frac{\partial\theta}{\partial k} * x^* + \frac{\partial\epsilon}{\partial k}\right)}{\left(\frac{\partial\phi}{\partial V} + \frac{\partial\theta}{\partial V} * x^* + \frac{\partial\epsilon}{\partial V}\right)} \tag{4-40}$$

The optimality condition is now altered to the extent that the effect of the external cost is added for each input. The total effect is dependent upon the size and magnitude of each variable on the external cost. To illustrate, $\frac{\partial \epsilon}{\partial k}$ may be expanded to arrive at the following:

$$\frac{\partial \epsilon}{\partial k} = \frac{\partial \pi}{\partial s^*} * \frac{\partial s^*}{\partial k} * c * (1 - \gamma) \tag{4-41}$$

Thus, the external effects work through the effect on optimal speed (as perceived by the drivers), $\frac{\partial s^*}{\partial k}$, on accident probabilities and the extent to which this is internalized.

The purpose of RQ2c and paper 4 is simply to formulate and investigate the function $s^*(V, k)$ and its derivatives so that the influence of frequency ($V$) on optimal speed as viewed by the drivers may be found ($s^*$). That is, finding the size and magnitude of $\frac{\partial s^*}{\partial k}$. In the paper, I only consider the effect of increased waiting time from not reaching the departure in time. Thus, I assume that $u(s; V)$ is only a function of frequency and speed, thereby not taking capacity into account, which is a possible extension of the model. For example, some users may speed to secure a place aboard a ferry with capacity restrictions.

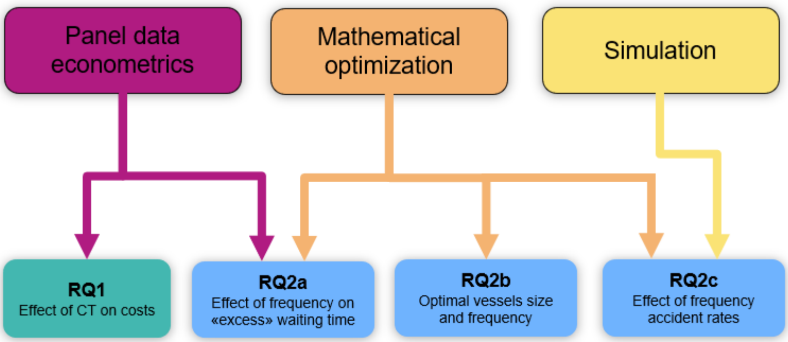
At the same time, it is important to mention some caveats. Firstly, measures other than changing the public transport frequency may be used to reduce drivers' incentive to speed. Moreover, according to standard economic theory, prices are usually employed to correct for externalities. Nonetheless it may be challenging to keep prices at a certain speed level directly and in real time. Moreover, other policies may be more efficient than changing the frequency, to influence speeding behaviour, which should be considered in practical applications.

# 5 Research methodology & data

*In this section, the research methodology is presented. Further, the reasons for choosing a particular methodology and its potential weaknesses are discussed. It starts by providing an overview of the different methods used before describing each one in greater detail. Lastly, the data is presented.*

## 5.1 Overview of methods

Figure 2 shows an overview of the methods used in this study and their connections to the different research questions. Panel data econometrics is used in answering RQ1 and RQ2a, where cost function parameters and the effect of competitive tendering are studied. Mathematical optimization is used in RQ2a-c to derive results regarding questions of how the service level of a ferry connection may be optimized from a societal planner's perspective and how individual travellers optimize their speed when connecting to a ferry service. Last, simulation is used in RQ2c to gauge the range of possible outcomes and shed light on the sensitivity of the results.




*Figur 2. Overview of methods used in the study*

Each method will now be presented and an attempt made to justify why they may be appropriate for each research question.

## 5.2 Panel data econometrics

In RQ1, the main question revolves around the effect of CT on costs and efficiency levels. This question has been investigated by several authors in the bus industry, but not, as far as I am aware of, in the ferry sector. Efficiency may be defined as "attaining the highest possible output with a given amount of input, or producing a given level of output to the lowest possible cost" (Holmgren, 2018). With respect to section 4.1.2, this is analogous to being cost efficient and on the production possibility frontier.

The literature has mainly focused on two methods: Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA) (Holmgren, 2018; Sheng & Meng, 2020). DEA is a non-parametric method which does not produce any measure of uncertainty. While certain methods do exist to approximate the uncertainty of DEA, (see, e.g., Simar & Wilson (1999)), it seems this method is seldom used over SFA in the transport sector. Moreover, in studies of ferry sector costs, simpler panel data models have been the preferred option (Jørgensen & Mathisen, 2010; Mathisen & Jørgensen, 2012). I have chosen to use simpler panel data methods in this study; I will now explain the rationale behind this choice over SFA and DEA. I start by briefly reviewing the concept of SFA before moving on to more specific arguments.

Efficiency in economic terms is dependent upon attributes of technology, most notably returns to scale. It is common to use either a cost or output-oriented framework in efficiency studies. In SFA, one usually employs some form of translog[17] function and specifies a parametric term which describes the effect on efficiency from different variables by establishing a production frontier[18]. The translog cost function has the property that the returns to scale may be derived from the parameter values (assuming a Cobb-Douglas production technology). For example, Coelli et al. (2005) introduce the SFA method as the following:

$$q_i = \exp(\boldsymbol{x}'\boldsymbol{\beta}) * \exp(v_i) * \exp(-u_i) \tag{5-1}$$

Where $q_i$ is some output or cost metric, $\boldsymbol{x}$ is a vector of explanatory variables (and possibly transformations of them), $\boldsymbol{\beta}$ is a parameter vector, $v_i$ is noise, and $u_i$ is efficiency. Thus, a higher $u_i$ reduces the efficiency level. The attractive feature of SFA is that $u_i$ can be specified parametrically and estimated by maximum likelihood so that measures of uncertainty may be derived. Moreover, it can vary over time in order to track temporal changes to efficiency. A major benefit of SFA is also that it allows for *comparing* the efficiency of different Decision-Making Units (DMUs); for example, different transit operators, banks, shops, etc. It is possible to produce a DMU-specific measure of cost efficiency (Coelli et al., 2005):

$$CE_i = \exp(-u_i) \tag{5-2}$$

In order to bind this comparison with economic theory, the framework is purposefully flexible in revealing the features of production technology (returns to scale), which in turn has a direct effect on the estimated efficiency scores and is thus important. As such, it seems to be an attractive method of assessing efficiency in the transportation sector.

However, there are two major reasons why SFA is not used to answer RQ1. The first is my lack of interest in comparing different DMUs to one another. The questions concern the effect of CT on specific crossings; consequently, the comparison of interest is intra-crossing-oriented and not inter-crossing-oriented as SFA. Panel data econometrics, in particular fixed effects estimation, have been purposefully made to study the effects of different units over time when compared to themselves. For example, (Bårdsen & Nymoen, 2014, pp. 135) show that the fixed-effects estimator of a parameter $\beta_{FE}$ of $x$ on $y = \alpha + \beta_{FE} * x$ may be decomposed into the following:

$$\beta_{FE} = \frac{\sum_{i=1}^{n} w_{xx}^{i} \beta_i}{\sum_{i=1}^{n} w_{xx}^{i}} \tag{5-3}$$

[17] The Translog function is a Taylor approximation to an underlying production or cost function.
[18] The possibility to produce given different inputs. The frontier's estimation is based on the most efficient units.

Where $w_{xx}^{i} = \sum_{t=1}^{T}(x_{it} - \overline{x_{it}})^2$ and $\beta_i = w_{xx}^{i}/w_{yx}^{i}$ is the least-squares estimator for each unit. Suppose $x_{it}$ is a dummy indicating whether crossing $i$ is under CT in period $t$ or not and $y_{it}$ the cost level. Thus, using the fixed-effects estimator produces a weighted average of the difference in cost when tendered and not for each crossing ($\beta_i$) and weights this to a total. It is notable that crossings that do not change tendering status has $w_{xx}^{i} = 0$ and do not affect the estimates. Consequently, the fixed-effects estimator and a panel approach are more suitable to answer the RQ1 than SFA. However, $u_i$ could be parametrically specified to get an indicator of tendering or not, even adding a time dimension to a panel data model (Coelli et al., 2005). The changes in efficiency could be decomposed at each crossing using Malmquist indices (Coelli et al., 2005, pp. 289). In an SFA, the efficiency terms are given as:

$$u_{it} = f(t) * u_i \tag{5-4}$$

With $f(t)$ a continuous function of time $t$. Using a panel approach would first require (i) specifying an appropriate form of $f(t)$ and, secondly, (ii) specifying a form of the inefficiency where $u_i$ depends on the CT status. As I am mainly concerned with the average effect of CT on costs, using a simpler fixed-effects framework seems preferable to employing a more involved approach that may require a significant degree of additional work with a level of benefit that is difficult to gauge, given that inter-crossing comparisons are not the study's focal point.

The second major reason panel data econometrics has been chosen over SFA is causal. If policymakers are to consider using CT, the causal effect, and not a statistical assertion, is the most interesting factor. As indicated by Hensher & Wallis (2005) the counterfactual case is important when addressing the consequences of CT; in other words, what would have happened if CT had not been used? The econometric literature has established a wide variety of tools which may identify or approximate causal questions. A specific branch is called the quasi-experimental approach, advocated among others by Angrist & Pischke (2009). As opposed to “proper” experiments, there is no random selection of ‘treatment’ (for example, being allocated a drug instead of a placebo during a trial) in using the quasi-experimental techniques. Using a traditional Randomized Control Trial (RCT) is often not feasible in practice (due to ethics, costs, etc.).

The benefit of RCT is the random assignment of treatment to the treatment and control group. In this case, the only differences between the groups are random, such that one may interpret the differences casually (Angrist & Pischke, 2009). In the quasi-experimental approach, different techniques are employed to approximate this ideal by adjusting for group differences. Among these techniques, panel data econometrics is one being used. Other methods have also been suggested, including Instrumental Variables (IV), Regression Discontinuity Design (RDD) or Differences-in-Differences (DiD) (Angrist & Pischke, 2009). I chose the fixed-effects estimator as this method (i) is able to adjust for other factors affecting costs through regression (for example, the number of kilometres sailed), (ii) adjusts for any crossings-specific differences in costs (iii) enables a difference in average costs within each crossing to be estimated before and after CT was implemented. Although other methods could also have been used, they require additional assumptions or data, which was not available. Perhaps DiD could have been used instead; however, doing so would have required assessing if the trends of the treated and untreated crossings to be equal before and after CT was introduced. As CT was introduced at different times and at different crossings, it is difficult to assess the validity of this assumption, and hence, the validity of the results.

In RQ2a, I use a fixed-effects regression to estimate each ferry's cost per kilometre sailed. This method was chosen because it is most commonly used and accepted by other researchers studying the cost structures of ferry operations (see, e.g., Jørgensen & Mathisen, 2010; Mathisen & Jørgensen, 2012).

## 5.3 Mathematical optimization

Mathematical optimization techniques are used in RQ2a-c. Apart from the fact that this technique is a main tool in economics and transport economics, it carries some benefits that are essential when discussing optimal service levels. Economic theory concerns itself to a large extent with the simple equation below:

$$MB(A) = MC(A) \tag{5-5}$$

This equation states that the extent of an "activity" ($A$), should be such that the marginal benefit of it ($MB$) equals the marginal cost ($MC$). A convenient way to assess the appropriate level of $A$ is therefore to specify mathematical representations of $MB(A)$ and $MC(A)$ by way of functions and solve for the value of $A$ that satisfies the equation. These techniques have been ubiquitous in transportation economics on public transport, or at least since Mohring (1972) and onward. Thus, using an economic framework, the above condition seems necessary in order to discuss *optimality* – which is the goal of RQ2a-c; optimal service levels and optimal speeds.

The following section contains a discussion of optimization's use when estimating both optimal service levels and optimal speeds.

### 5.3.1 Optimal service levels

To estimate optimal service levels, researchers have used different approaches. Some of the early research and research performed at a more conceptual level have only used analytical methods, as did Mohring (1972) and Jara-Díaz et al. (2017), among many others. However, due to the complexity of the problems involved, I was unable to derive any applicable and analytical solution. When this occurs, numerical techniques may be used to derive the optimum (Nocedal & Wright, 2000).

In the literature, there have been two major numerical method approaches used to estimate optimal service levels. The first approach is to specify the objective function to be minimized and use numerical algorithms to find the optimum (see, e.g., Asplund & Pyddoke, 2020; Tirachini et al., 2014). The second approach is to derive the conditions for optimality and then solve for the relevant optimization variables in order for conditions to be satisfied (see, e.g. Börjesson et al. (2017). In the latter case, it is important to check that an actual minimum is found through the second-order conditions.

In paper 2, an approach is used where the objective function is directly minimized. In this case, there is a cost function $C(x)$, which is minimized with respect to the argument $\boldsymbol{x}$. I use a simple method called *gradient-descent* (Nocedal & Wright, 2000) to find the minimum, at the same time verifying that the function is convex.

In paper 3, an approach is used where optimality conditions are derived and then solved. I chose this method for two primary reasons. Firstly, it is beneficial to derive optimality conditions as this enables a more fruitful discussion concerning the elements that influence the optimal service level. Secondly, paper 2 includes a measure of the consumer's surplus. If the objective function were to be optimized directly by using a numerical scheme, a mathematical form of the consumer's surplus would have to be specified. Several alternatives are available, and it is no easy matter to decide which one to choose. While several alternatives were tried, it was quite difficult to find one that both made sense theoretically and was stable numerically. Finally, I decided to take a simpler approach by deriving the optimality conditions and solving for them. To give a brief overview of the difference, consider the objective function $C\big(x, h(x)\big)$, where $x$ is the service level and $D = h(x)$ is the demand level, given the service level. This function could either be optimized by numerically finding the maximum of $C\big(x, h(x)\big)$ (approach 1) or deriving the optimality conditions and solving for these (approach 2).

Solving the optimality conditions produces the following equation:

$$f(x, D) = \frac{\partial C(x, D)}{\partial x} = 0 \tag{5-6}$$

However, this equation depends on demand, which again depends on the service level. This means that if the second approach is taken, there will be a need to iterate in order to find a solution to the following system of equations:

$$x^* = \{x: f(x, D) = 0\} \tag{5-7}$$

$$x^* = h(x^*) \tag{5-8}$$

where $h(x)$ is a demand function. Interestingly, Börjesson et al. (2017) solve this problem by simultaneously solving for all service and demand levels. However, Li (2002) uses a slightly easier approach when estimating optimal congestion tolls for cars (his problem is analogous to the current one). In paper 2, the externality is based upon each ferry's limited capacity, whereas in congestion pricing, it is the limited capacity of a road, a fact that makes the problem analogous in nature.

Li (2002) uses a simple root-finding algorithm (called *the Bisection method*), realizing that $x = h(x)$ is a fixed point on the demand curve. Thus, by sequentially finding $x^*$, and inserting this into $h(x^*)$, there is a convergence towards the economic optimum. I use an approach called *fixed point iteration* (Wood, 1999), which is just a slightly more convenient way to approximate the fixed point (the equilibrium). To solve the model, the optimal service levels have been solved, given a level of demand. Next, the service level is inserted into the demand function to obtain a new demand estimate. The process then continues until there is a miniscule change in the demand level between two iterations, resulting in the optimal service and demand levels.

It should be noted that when using this approach, setting the derivatives equal to zero for the objective function finds a minimum. The function's maximization is performed by the iteration, where one also finds the optimal demand. Effectively, optimal demand and service levels are found by finding optimal service levels for a **fixed** level of demand in each iteration (hence, a minimization), and the solving for optimal demand **by the iteration process** and the condition that established the economic optimum.

### 5.3.2 Optimal speed

Optimal speed calculation is based upon two different optimization variables, one being discrete (which departure to choose) and the other continuous (speed for a given departure). It is quite difficult to combine discrete and continuous optimization in a non-linear setting. However, as there are some natural boundaries to the problem, it may be solved by a clear enumeration of each alternative. The methodology used is very straightforward: For each departure, the optimal speed is estimated and then compared with the cost of each departure, given the optimal speed. The option with the overall lowest cost is then chosen.

The problem of determining the optimal speed for a given departure is solved by finding a first-order condition and then solving it numerically. Next, after verifying that it is a minimum, the cost of all departures is compared and the best alternative chosen.

## 5.4 Simulation

Simulation is used in RQ2c when considering the effect on speed caused by changing the departure frequency of a public transit service. Gilbert & Troitzsch (2005) highlight that simulation may be beneficial in cases of complex models, in particular non-linear ones (Gilbert & Troitzsch, 2005, pp. 16); in addition, they note that simulation may be useful when performing sensitivity analyses (Gilbert & Troitzsch, 2005, pp. 24). A Monte Carlo simulation (Kennedy, 2008, pp. 22) has been used here in which different parameter values are drawn several times over, the model being solved each time and yielding a large set of estimated effects.

Regarding the specific research question, the potential effects on the speed of individual users after changing the departure frequency of a public transport service has been considered. As these parameters may vary, simulation is used to assess how different values affect the results. Different parameter values are drawn from a statistical distribution and the corresponding results given those values are found. The process is then repeated many times over to assess what if any overall trends are shown by the results.

## 5.5 Data

In this section, the different data sources used in this thesis are presented, and there are three main datasets that are discussed separately. Because the papers contain more information regarding the data, the descriptions below are to be viewed as summaries.

### 5.5.1 Cost data for ferries

To investigate RQ1, I use a dataset provided by the Norwegian Public Roads Administration (NPRA) covering different financial measurements for a collection of Norwegian car ferry crossings.

The data covers a total of 53 links out of a total of 130 operating in Norway, which comprises 41 % of all existing links in 2009 (Odeck & Høyem, 2021) Measured in terms of demand, the dataset covers 69 % of all PCE[19] units transported in Norway in 2010. Moreover, the majority of these are two-port crossings, meaning that 3- or 4-port crossings are covered to a lesser extent. Consequently, the data does not cover all crossings, and the effects observed may differ in the total population, which is a weakness in the study.

The data covers the years 2003-2010; after 2010, the companies no longer reported any specific data to the NPRA. As the data has a panel structure where the same units are observed over time, the total number of observations is 8 * 53 = 424.

Data on costs are divided into the following categories: fuel, maintenance, lubricant and labor costs. There were also some operational data on the number of PCE transported, kilometres sailed, operating company and type of contract (CT or otherwise). Further, data on the number, size, and age of each ferry was added.

In RQ1, the data was also used to assess the effect of CT on market concentration, while in the RQ2a, the data was used to obtain a cost per kilometre sailed.

### 5.5.2 Demand data for car ferry trips

Demand data for car ferry trips was provided by the NPRA in a specialized excerpt from their database on ferry demand (Ferjedatabanken). The datasets covered demand for the case study crossings examined in papers 2 and 3.

Demand is delivered in aggregated form to hourly figures for each crossing and direction on a separate basis. Moreover, the NPRA provided a table which enabled conversion of the simple demand data into PCE units.

Demand is registered by two different systems. In the "Autopass" system, an electronic chip is used that registers all cars entering the area. In the "Riksregulativ" system, demand is counted manually by the ferry staff. During periods of very high demand, the staff may not be able to count all vehicles. However, this is likely not a major issue as (i) the periods of excessive demand are quite short, and (ii) it is mostly relevant for very short crossings, where the staff have limited time to count and register all vehicles. The Autopass system may have a theoretical weakness if it is unable to register cars or if it goes offline due to system errors, etc. (these errors are not widely reported).

### 5.5.3 National travel survey

Data from the national travel survey was used in the RQ2c simulation process. This data from 2018 covers a representative national sample of individuals aged 13 years or older. It was gathered by survey company Opinion on behalf of the Ministry of Transportation.

[19] Passenger Car Equivalent. A standard "personal" car is 1.025 PCE, while a truck that is 19 meters or above in length is 10.682 PCE.

The survey was used in a highly restricted manner as only the length of car trips was actually employed in RQ2c.

The national travel survey retains certain weaknesses with respect to representativeness. For example, it is weighted for sample biases in gender, age, season, geographical location and day of travel. Moreover, it covers all trips, not just ferry trips, and as a result is first and foremost only relevant for the average car trip in Norway. Trips to a ferry crossing may be longer or shorter than an average trip; consequently, an ideal dataset would only contain car trips to a ferry crossing. However, the main purpose of using the NTS data is to obtain variation in trip lengths in order to simulate uncertainty surrounding the effect of frequency on speed. As such, the primary focus is not on estimating a “final” effect but rather on estimating the range of possible outcomes. In short, the average trip lengths are not themselves the primary focus of study. This point is also discussed in paper #4.

# 6 Discussion of results

*In this chapter, the results from the papers are discussed in relation to the research questions posed. In addition, important limitations and opportunities for further research are also discussed.*

## 6.1 RQ1

In paper 1, the first question addresses how cost efficiency may be enhanced through organizational reforms using the car ferry sector as a case study and asking how operational costs are affected by Competitive Tendering (CT) in the context of ferries.

### 6.1.1 Results

The general results from Odeck & Høyem (2021) concerning RQ1 is that CT reduces the operational costs of car ferry crossings by an average of 8 %. These reductions are at the lower end of the figures reported in the literature ranging from 20-30 % (Sheng & Meng, 2020). Consequently, while there is an effect caused by cost reductions, it is smaller than the average found in the bus sector. Hensher & Wallis (2005) noted that in the case of New Zealand, the cost savings were about 5 % when a private operator ran the service prior to tendering and 40 % when a public one did the same. As several crossings were operated by private companies[20], this may be a factor in explaining the difference.

Another result was increased market concentration as some larger firms increased their share of the crossings. It is important to underline that only a sample of the total market for ferry services has been observed, while in the sample, a higher concertation was observed. Increased market concentration is proposed by Hensher & Wallis (2005) as a possible reason for CT's reduced effectiveness in subsequent rounds of tendering. Aarhaug et al. (2018) noted an association between market share and bid prices in the Norwegian bus industry; fewer bidders resulted in higher bids. If market concentration and bids increase over the long run, the effectiveness of CT may diminish over time.

All in all, it seems that the effect of CT on the car ferry sector is one of reducing costs; however, this occurs at the lower end of the spectrum noted in the literature. Moreover, it seems that market concentration increased after CT was introduced. If one accepts the typical administrative costs noted by Hensher & Wallis (2005) at five percentage points of savings at face value, this leaves only a real-cost reduction of 3 percentage points. Further, if the increased market concentration results in increased bid prices (as has been noted in the bus industry), the net gain may fall even more over time. As such, policymakers should carefully consider whether CT is appropriate or not and weigh any cost reductions against the transaction costs that may arise when performing a tender. This point also applies to the long-term implications for market structure and bids as the net gains appear to be relatively minor on average in the Norwegian case.

In the long run, using CT may lead to reduced competition. However, it is important to emphasize that there was only a correlation observed between tendering and market concentration, which is not a causal result. Moreover, several factors were not included in the analysis that may be relevant for policymakers' decision-making processes. For

[20] They were operated by private companies that had long-term government contracts.

example, moving from negotiations to CT may affect employees' working conditions, involve transaction costs when performing the tendering, etc. These concerns may also be relevant for policymakers' decisions.

Considering its theoretical framework, the paper finds that planners' budget constraints may be somewhat relaxed. However, the results should not be interpreted as there are 8 % more resources since the estimate has needed to be adjusted to reflect the difference in transaction costs.

Moreover, the cost structure of the transaction costs themselves may be relevant to consider, which will likely include the tender's preparation. It seems reasonable that there are some fixed costs involved in this work, for example preparing documents so they meet legal standards, etc. If there are economies of scale present, it may be that the attractiveness of using CT may vary according to crossing size. In the literature review, the following function has been introduced:

$$\delta = C - A \tag{6-1}$$

If we assume that cost savings are a fixed proportion of operating cost, $\alpha$, given by the number of PCE and a cost per unit $c_O$, $C = \alpha * PCE * c_O$, and that the cost of preparing the tender is given by $A = c_T * PCE + F_T$, with $F_T$ a fixed cost, we may solve for the $\alpha$ so that $\delta = 0$:

$$\alpha = \frac{1}{c_O}\left(c_T + \frac{F_T}{PCE}\right) \tag{6-2}$$

The function exhibits economies of scale as the average cost falls in accordance with the amount produced (PCE). It is therefore reasonable that the unit cost of preparing the tender is much smaller than the actual cost of operating it, such that $c_T/c_O \approx 0$. In this case, the larger the crossing, the smaller the cost reduction (in percent) required. Consequently, if the estimated effect of CT on costs is minor, it is less likely that it is advisable to use it. Policymakers should take the size of the crossing into account when considering putting a crossing up for tender if economics of scale are present when preparing the tender.

### 6.1.2 Limitations and possibilities for further research

There are several weaknesses and possibilities for further research that are important to mention.

First of all, having a more complete dataset would have enhanced the analysis. There are several dimensions in which the data set could have been extended:

- Time: We retained data for 8 years, a length of time which permits evaluating the first round of tendering. However, even though subsequent rounds were not included in the analysis, they are interesting as they enable a more thorough long-term analysis of market concentration. Several authors have observed that gains tend to shrink in subsequent rounds (see, e.g., Sheng & Meng, 2020), and longer time sequences would enable studies of these effects.

- Scope: The dataset covered 41 % of all crossings, which indicates that over half were not analysed. Even though there were significant variations in the sample with respect to crossing size, including further crossings may have altered the effects. The paper did not consider crossings' regional spread in the sample; neither did it consider if the pre-tendered crossings were operated by public or private companies. The latter has been noted to yield lower estimates of increased efficiency when private companies operated the service prior to tendering (Hensher & Wallis, 2005). Investigating the representativeness of the sample along these dimensions may have enhanced the discussion on generalizing the results.
- Variables: Given that only cost efficiency and market structure were considered in the analysis, it would have been interesting to include other variables, for instance the number of bidders. Vigren (2018) investigated factors that influenced the number of bidders for tenders in the Swedish bus sector, and identified the local competitive environment as important. Aarhaug et al. (2018) summarized developments in the number of bidders per tender in Norwegian bus contracts. A reduction in the number of bids may increase the price, which would in turn reduce the gains over time.

Gathering more data along the dimensions listed above would enable a more complete analysis of how tendering impacts the sector's long-term efficiency.

Secondly, the methodology used to assess the effect of tendering upon costs was linear regression with fixed effects. As the research question is causal in nature, it is important to evaluate if the estimated effect may be attributed to tendering or other factors through omitted variable bias (Angrist & Pischke, 2009).

In the study, numerous operational variables that are expected to influence costs were included to mitigate the possibility of having other factors explain the observed effect, including PCE transported, ferry size, age, number of kilometres sailed, etc. We included a trend term to adjust for trend changes in the overall cost levels. Such factors could be wages, prices of input factors like fuel and maintenance, etc. A weakness of the study is that these variables were not observed directly, but rather by proxy using a trend term. Including data on both factor use and factor prices would have reduced the possibility of omitted variable bias and enhancing the robustness of the results further. Including data on factor prices directly may therefore be a valuable next step to take in the research.

Alternative methods that are less sensitive to omitted variable bias could also be explored. One particular example is *Instrumental Variables (IV) estimation* (Kennedy, 2008). This method requires an "instrument" which is connected to the independent variable (tendering status) but not the variable of interest (costs). I was unable to find a valid instrument based on the available dataset. More detailed knowledge of *how* the crossings were subjected to tendering would have been useful in this regard.

For example, if crossings were put up for tendering simply because the contract was to be renewed, then contract start-time could be a potential instrument. However, this may also be a risky strategy. In order to understand this point, suppose the crossings with the highest cost levels were put up first (irrespective of contract renewal time), this could not have been used as an instrument (as the instrument would be correlated with the cost level). More specifically, if the contract start-up time was somehow correlated with higher cost levels, the instrument's validity would be called into question because start-up time would then be correlated with cost levels. An even longer time series of cost would have been required to assess this question, an option which was not available.

Having more data and finding potential instruments would have enhanced the analysis, and further research may consider these questions during the design and data acquisition phase.

Lastly, although the effect on market concentration was assessed in Odeck & Høyem (2021), this was only done on a correlative level. That is, the Herfindahl-Hirschman index (HHI) was tracked along with the proportion of tendered services. This approach has two weaknesses that stand out:

- Firstly, the market concentration was based on the same sample as the cost estimation, covering 41 % of the market. Thus, the HHI is only valid for 41 % of the market; as a result, it has been effectively treated as a submarket in the calculations. It would have been preferable to have had more data in order to attain a more precise measurement of the effect.
- Secondly, the analysis was restricted to a correlative discussion of the (possible) co-dependence of market concentration and share of tendered services. Analysing the same material using a causal method, which restricts the possibility of the correlation being spurious, would have strengthened the analysis. Moreover, the Herfindahl-Hirschman index only indicates market concentration and not market power.

In summation, having more data with respect to time, scope and variables (particularly input factor prices) would have enhanced the analysis. So it would be advisable to conduct more research using higher quality data. Secondly, the long-term impacts of the gains from tendering in the ferry market are not yet fully understood. More research is needed on the evolution of competition and bid prices over time. For example, studies that document how the number of bidders, market structure and bid prices have evolved in subsequent rounds of tendering would be of great interest.

## 6.2 RQ2

In this section, the different components of RQ2 are discussed separately and subdivided by each paper. Lastly, a synthesis is provided.

### 6.2.1 RQ2a

Research question 2a asks what the optimal departure frequency is when taking into account excess waiting time costs when capacity is too low. It addresses, along with RQ2b and 2c, the broader question of how the optimal service level at a crossing may be estimated. The question is addressed in paper 2.

#### *6.2.1.1 Results*

Paper 2 developed a methodology to estimate the number of users left behind in cases of insufficient capacity. It introduced excess waiting time for ferries due to the fact that waiting time arises when users need to sit out at least one departure in order to secure a spot. Two methods are tested: one in which there is FIFO[21] queuing at the quays, and one in which there is no organized queue (modified newsvendor model). The revised methods, including excess waiting time, are compared to the Mohring (1972) model, in which excess time is not included. The paper comprises an important part of the dissertation as it tries to develop a way of estimating costs in situations of insufficient capacity.

The paper's main result shows that the revised method yields higher frequencies compared to the approach of Mohring (1972) when ferry sizes are limited. That is, the method yields optimal higher frequencies in periods of high demand when there are capacity restrictions. Secondly, the approach in which a FIFO queue is modelled and an "unstructured" queue formed yields similar results. Given that most ferry services have a structured FIFO queue, it is interesting to note that these two approaches yield similar results.

The main purpose of the paper is to develop a methodology that can be used later in the dissertation. One of the paper's findings is that frequencies should be higher during periods of high demand (given the ferries' capacity ). However, there are some important caveats that are investigated further in RQ2b, having been left out of RQ2a. For example, the ferry size is fixed. In reality, it may be changed by either acquiring a new ferry or rerouting an existing one. Thus, capacity could also be increased by using a larger ferry and not having a higher number of departures. Further, there is only an average cost per sailed kilometre, which does not vary according to ferry size. A shortcoming of the paper is the fact that ferry size is always fixed, which is the main reason for extending the framework in RQ2b.

Moreover, it is shown that the method used is consistent with the approach of Mohring (1972); when an infinite capacity is assumed, the optimal frequency is equal under both Mohring's method and the revised one.

#### *6.2.1.2 Limitations and possibilities for further research*

In relation to RQ2a, a model designed to assess the excess waiting time cost has been developed. However, due to lack of proper data, there has been no direct validation of the predicted levels of users left behind. At the time of writing, data on the number of users left behind were not available on an hourly basis, which is the unit of prediction. This is not only a limitation and but also the reason why the specific model called "newsvendor model" was chosen; this is an established method in the supply chain literature (Thonemann, 2005) to limit excess capacity.

The newsvendor model uses the average demand within a specific time period. If demand varies within this time period, the average will tend to underestimate the true probability of being left behind because the total capacity per hour is used to estimate the probability. Therefore, if 90 % of demand is allocated to 50 % of departures in reality, the model assumes that demand is uniformly distributed, which will yield estimates of the percentage left behind that are too low.

[21] First-in-First-Out. The first to arrive at the queue is also the first to depart it. A typical example is a check-in counter at an airport.

Moreover, the method assumes that users must wait at the most one additional departure. In practice, queues might become so long that they end up having to sit out several departures. This is a drawback of the method, which may yield estimates of excess waiting time that are too low.

Developing a better representation of the excess waiting time cost could be an interesting topic for further research. Two specific areas here stand out: (i) developing methods that do not average capacity utilization over several departures (ii) enabling the possibility of having to wait for more than one departure.

This could be done by developing better analytical or statistical models. Analytical methods would most likely require some increase in mathematical complexity, and statistical methods would require more data (or, alternatively, a combination could be used). Last, simulation could also be an option.

A brief outline of a method which considers both individual departures (to reduce the aggregation error) and the possibility of having to wait for more than one departure will be presented. It is not meant to be in any way exhaustive, but to function as an example of how to proceed.

Consider a demand level, $y$, for a given period, which has a distribution function over time $t$, $f(t)$. The expected number of users arriving between two consecutive departures at times is then $t_h$ and $t_{h-1}$, where $h$ denotes the departure number:

$$e_h = y \int_{t_{h-1}}^{t_h} g(t) * dt \tag{6-3}$$

Further, let the total number of users waiting to board departure $h$ be defined as:

$$x_h = e_{h-1} + z_h(x_{h-1}) \tag{6-4}$$

Here, $z_h(x_{h-1})$ is the number of users that were unable to board the preceding departure, $h-1$. This is defined as the number of users in the last period times the probability of not being able to board:

$$z_h(x_{h-1}) = x_{h-1} * \left(1 - p_h(x_{h-1})\right) \tag{6-5}$$

In this framework, users that are unable to board departure $h$ simply move forward in time to the next departure. The probability of being unable to board is the inverse of the probability of being able to board, which is given by the demand waiting to board and the vessel capacity, $k$:

$$p_h(x_h) = \min\left(\frac{k}{x_h}, 1\right) \tag{6-6}$$

This function is tied to each individual departure $h$, instead of an hourly, average capacity. Increasing the capacity results in a higher probability of being able to board. If $\lambda$ is equal to frequency, we can then compute the total number of users that have to wait and scale it by the waiting time for each missed departure:

$$C = \frac{1}{\lambda}\sum_{h=1}^{\lambda-1} z_h(x_{h-1}) \tag{6-7}$$

This equation shows the total amount of time spent after having missed a departure due to capacity restrictions. The departure times are given by $t_h = t_{h-1} + 1/\lambda$. Applying a framework similar either to this one or another that addresses the same shortcomings in a better way is of interest in further research.

### 6.2.2 RQ2b

RQ2b expands upon RQ2a by allowing more variables to be a part of the optimization, asking about the optimal combination of vessel size, price and frequency level at a car ferry crossing. The question is addressed in paper 3.

#### *6.2.2.1 Results*

RQ2b concerns the optimal service level and capacity at car ferry crossings. In RQ2a, only frequency was investigated when optimizing service levels, whereas RQ2b includes the number and size of ferries as well as ticket price. This is done while incorporating the cost of being left behind, which is perhaps more important in the ferry sector as compared to the bus sector. The model is run on three different crossings in Norway as case studies, comparing optimal to current service levels. Further, several sensitivity tests to assess the results' robustness are performed.

The main result is that the current capacity is too large at all surveyed crossings. Both vessel size and frequency are investigated. Of these two, frequency should be slightly higher or equal, and ferry size should be reduced by quite a large margin. Moreover, prices should be lower and optimal demand higher. Further, it has been found that the current capacity is set to cover a very high demand level, so that quite a few departures are expected to provide limited capacity. Consequently, it seems capacity is set in a way that ensures the highest periods of peak demand will be covered. Typically, this happens during the summer, but there remain large time periods during which a significantly lower capacity level would cover demand.

When capacity is reduced, the number of users left behind increases. However, the model's base scenario recommends a very small vessel size and high-capacity utilization. It may be that the method used to estimate the number of users being left behind underestimates the true number. To assess the robustness of the results with respect to this specific assumption, several sensitivity tests have been run in which the number of users being left behind has increased. Generally, the results have remained the same; capacity that is too high has been offered. However, this dissertation recommends that more effort should be put into better modelling of the number of users left behind in order to get more accurate results.

It has been concluded that policymakers should consider revising the service levels of specific crossings. Moreover, policymakers should consider if improvements are in order at other crossings. The crossings surveyed are among the largest in Norway. As such, if there is overcapacity at the largest crossings, it may also be interesting to consider smaller ones as well.

Jansson (1980) observed that there was a general trend where policymakers used buses that were too large and too few in number, and the same situation has been observed in the case studies. However, a major assumption is that these are the first-best service levels in which only technical constraints are present, and not financial ones. Jara-Díaz & Gschwender (2009) noted that a financial constraint 'forced' the planner to act as a private profit maximiser, thereby limiting costs. In this case, running a few large vessels is optimal; increasing the number of departures reduces the cost to users but increases the cost to operators. Waters et al. (1996) made a similar observation, noting that large ferries would be used if the only requirements to be met were related to capacity.

However, the general recommendation is that too much capacity is being offered as compared to the aggregate willingness to pay for the service. Further, more research is needed to develop methods to estimate the number of users being left behind.

#### *6.2.2.2 Limitations and possibilities for further research*

The study represents a first step in recommending optimal service levels at ferry crossings; therefore, several limitations are important to discuss in the context of RQ2b:

- If the planners face budgetary restrictions, the first-best conditions no longer apply, and in this case, running a few large ferries may be beneficial. This may also have implications for the optimal price levels, which are likely to be significantly higher.
- Planners may be required to make decisions with respect to several different crossings at once, and not treat each in isolation (as is done in this thesis). Thus, having "flexible" ferries that are capable of handling a variety of crossing types may be beneficial.
- If budget restrictions apply on an aggregate level to the entire sector, optimization should ideally be performed simultaneously for all crossings. It follows that this is an important extension that should be seriously considered in further research.
- Moreover, the method used to estimate the number of users left behind could be refined, as was indicated in RQ2a.
- Demand varies in accordance with the time of year, with higher levels during the summer for many crossings. Adding a separation of seasons into the model could be a possible avenue for further research.
- There is no time horizon in the model as it is run for a static period. As ferries tend to have long service times (up to 30 years) and traffic tends to grow, allowing the costs and benefits of increased capacity to be extended temporally may impact the model's conclusions. For example, it might underestimate the

current optimal capacity as ferries need to be larger to handle future demand as compared to current levels.

- We did not include the marginal cost of public funds, which may affect the results, most likely by increasing ticket prices.

### 6.2.3 RQ2c

RQ2c investigates how second-order effects (externalities) may affect optimal service levels. In comparison to RQ2a-b, it does not directly address the cost of using the service, but rather the broader costs associated with its operation, or, more specifically, how accident rates are affected by increasing departure frequency at car ferry crossings. The question is addressed in paper 4.

#### *6.2.3.1 Results*

The last research question concerns an indirect effect of optimizing the service level of a public transport service. Paper 4 develops a model of a representative ferry passenger who drives their car to the quay. The purpose here is to gain theoretical insight into which mechanisms may be at play when altering the frequency and their effect on speed selection.

In the model, the driver experiences uncertainty if they will arrive too late to make the departure. The paper explores how this uncertainty affects drivers' choice of speed and whether changing public transport service frequency will affect the outcome. An economic model has been developed and several simulations performed in order to learn more about the possible mechanisms at play. Each driver makes two distinct choices: (i) which departure to try to reach, and (ii) which speed to drive at. By choosing a later departure, although scheduling[22] costs are increased, a lower probability of being "too late" is experienced. By choosing an earlier departure, although scheduling time is saved, there may be an increase in speed and/or probability of arriving too late.

As a result, there are two possible effects: If frequency increases, drivers may save scheduling time by aiming for an earlier departure, which may increase the speed required to reach it. At the same time, increasing frequency means that the cost of arriving too late is lower for a given departure, which may reduce speeds.

The net result is that frequency's effect upon chosen speeds is both difficult to predict and quite sensitive to several assumptions:

- The relative size of the scheduling and cost of arriving too late are important factors. This is because having a higher scheduling cost makes it more costly to choose a later departure. In turn, when a later departure is not chosen, a higher speed is maintained. It would therefore be interesting to obtain information on how "cumbersome" users view scheduling times versus the cost of arriving too late.
- Trip distance is also an important factor. During longer trips, maintaining a higher speed is necessary to attain a low probability of arriving too late.

Using a representative sample of trip lengths in Norway, it has been found that increasing frequency may induce higher speeds when departure frequency is very low

[22] These are costs that arise when the traveller is not able to arrive/depart at their preferred time. For example, having to leave earlier for work to avoid being too late if there is a change in travel times (due to, e.g. queues, etc.).

to begin with. Moreover, if there are many users choosing a high speed, increasing frequency may lower their risk-taking behavior. However, the effect decreases in accordance with the number of departures as the cost of arriving too late is low for higher frequencies.

Thus, planners should be aware that increasing frequency rates at a ferry crossing may induce higher speeds when this rate is low at the outset. Further, the cost of arriving too late relative to the scheduling cost can be reduced for the purpose of incentivizing users to choose a later departure, which would in turn limit speeding.

#### *6.2.3.2 Limitations and possibilities for further research*

The results are dependent upon several parameter assumptions; consequently, they only provide a limited number of theoretical insights into what might happen if frequency is changed at a ferry crossing. Given this fact, the study is not suitable for definitively answering what the likely effect is in reality. To answer these questions, empirical studies need to be performed.

The NPRA retains data on road speeds that are in proximity to ferry quays. This data would be of great interest to use in an empirical study given that frequency is altered at the ferry service.

Additionally, the study points out possible mechanisms which may be in play. This type of information may be of interest when designing empirical studies or interpreting results from such studies. Local, empirical estimates of the effect may vary; thus, it may be salient to assess the results from several different situations if the predictions are to be tested empirically.

Further, RQ2c has only considered the case in which a driver is delayed until the next scheduled departure, effectively assuming that there will always be a "next departure". It would also be interesting to consider whether changing the opening hours of a service would impact speeding behaviour. For example, if you live on an island where the ferry is your only mainland connection, you may drive at an excessive speed to make the last ferry departure of the day. If empirical studies were to be performed, it would be interesting to assess if such behaviour could be observed.

The study is a stylized example of using parameter values that are uncertain. While these parameters have been gathered from the literature, the specific values relevant to ferry applications may differ. It would therefore be interesting to conduct studies that enable an estimation of case-specific parameters instead of relying on analogous ones from the literature. Further, the study has indicated that the effect may vary, including that among users who have "different parameter values". Yet it is interesting to consider assessing not just the mean values, but also their variations. For example, it may be that only a subset of drivers has preferences that lead to excessive speeding.

Finally, there may be other ways of reducing drivers' risky behaviour than changing a ferry service's frequency. Alternative methods have not been considered, which could be an interesting topic for further research; the cost effectiveness of alternative measures is of particular interest by for example investigating if speeding could be reduced at the same magnitude but at a lower cost.

# 7 Conclusions & further research

The main research goal of this dissertation is the following:

> *To investigate possible measures enabling increased economic efficiency of car ferry crossing operations when considering organizational and service levels aspects as well as focussing on optimal capacity.*

This question has been investigated through several research sub-questions and addressed in a total of four papers.

Principally, two different routes have been taken to answer this question. RQ1 concerns organizational issues, and RQ2 considers the planning of the operation itself. The theoretical framework was used to show how the different questions are connected to each other. In this section, I will briefly summarize the implications derived from the study and point towards possible areas for further research.

## 7.1 Contributions of the thesis and discussion

In general terms, the aim of organizational reforms like CT is to relax the budget constraints encountered by the government planner. The aim of optimizing service levels through quantitative models is to derive a number of recommendations as to how a service should be optimally designed.

Based on the evidence in this dissertation alone, it seems that the potential for increasing social efficiency in car ferry operations may be greatest when it comes to choosing the appropriate service level rather than organizational reforms. The latter is of smaller size compared to average results found in the literature. Moreover, the long-term effect is uncertain, while the former exhibits a stark contrast between current and recommended levels. To increase efficiency in the car ferry sector, planners are advised to place more emphasis on designing proper service levels rather than pursuing organizational reforms. Papers 2 and 3 have indicated that it is important to consider the point that the cost of providing a capacity level which is too low; and yet, having capacity which is too high is currently a possible feature of a number of the largest crossings in Norway.

As an alternative to CT, some researchers have proposed so-called *performance-based contracts* (PBC) (Hensher & Stanley, 2003, 2008). When following the terms of these contracts, the operator would be required to meet a pre-specified level of quality, and subsequent costs would be on a level comparable to the efficiency in CT. The transaction costs of CT would be lower (which is a benefit), but at the same time the planner would need skills in optimizing service levels – which may require organizational and methodological investment (which is a cost). Further, the government should be able to penalize operators for not meeting the required standards.

Using PBC requires operators to have a good knowledge of their users' preferences. More knowledge relevant to the ferry sector is available (Díez-Gutiérrez & Tørset, 2019; Mathisen & Solvoll, 2010), and utilizing this information in an efficient manner seems to be advisable. Moreover, doing so ties in with the results from papers 2 and 3, which highlight the point that proper design of service levels is important to increasing societal efficiency. For example, Börjesson et al. (2017) studied a high-volume bus corridor in Stockholm and found that optimizing service levels was more important than prices[23].

The results derived in this dissertation partly support the idea of PBC; the greatest potential for increasing societal efficiency may come from choosing the right service level. Accordingly, planners should perhaps ensure they have an economically sound product over the specific operational responsibilities. If there is too little or too much capacity, this is still the case even if one uses negotiations or CT.

An issue mentioned by Sheng & Meng (2020) is that CT reduces the planner's options for changing the service levels within the contract period, when contracts are incomplete, as they often will be, since it is difficult to foresee every contingency. This indicates, especially with longer contract terms, that it is highly important to make sound decisions regarding service levels if CT is to be used.

Moreover, the results suggest that planners may have been focussing too much on limiting the number of users being left behind by operating large ferries. This finding is similar to that of Jansson (1980) concerning the bus sector (oversized ferries are being used). It would be interesting to consider factors that underline planners' decision to choose a particular service level. Assuming they are rational actors, it may be another objective function is present (or other constraints are present) than the ones shown here.

Further, the dissertation has investigated whether altering service levels affects accident costs, which may be relevant from an optimization perspective. It has been shown that frequency can theoretically affect motorists' speeding behaviour while they are enroute to a ferry quay. The study has pointed to two separate effects: One that drives the intensive margin; that is, which departure to choose. Secondly, an intensive margin; that is, the speed considered optimal for each departure. The collective magnitude of these two effects constitutes the final results; hence, the effect is expected to vary.

More specifically, it has been shown that the effect is expected to depend on many factors, including preferences, trip length and initial frequency. For this reason, more research is needed to better understand the effects that might occur when the frequency of a public transport service (such as ferries) is altered. However, the study has contributed to the literature by pointing at possible mechanisms which may be in play, a factor which could be beneficial when designing empirical studies of the effect.

## 7.2 Possibilities for further research

Some possibilities for further research have been identified:

- **Optimization under budget constraints is important to consider:** If budget constraints are present, the optimal service level may be different from the one derived here (Jara-Díaz & Gschwender, 2009). In this sense, there may be little room for implementing the first-best solution. Measures to relax budget constraints may then become more important, resulting in the results derived in RQ1 becoming of greater interest.

[23] Their overall result was that the frequency rate was too high.

- **More research is needed on the transaction costs of CT to make final recommendations:** Moreover, the question of transaction costs is important, and has not been investigated thoroughly in the literature, opening new research possibilities. If the effectiveness of CT is reduced— but the transaction costs are lowered – CT may still be attractive from a purely cost-saving perspective.

- **More research is needed on the long-term effects of CT:** More research is needed on the long-term effects of CT, including how market concentration is affected and what implications this entails for bidding. Mathisen (2016) discusses the presence of cross-ownership and market concentration following the introduction of CT and warns that increases in the minimum bid may follow and give rise to local monopolies. This suggests that policymakers should not adopt CT unconditionally, but rather consider the potential long-term impacts.

- **The models of excess waiting time may be improved:** The method used to estimate the number of users left behind may be improved. While a mathematical model was used in this thesis, a purely statistical one could also be viable. Developing a statistical model requires that there is sufficient variance in the capacity to assess its effect on the number of users being left behind.

  Currently, the NPRA maintains an extensive database on demand at each crossing. However, data on service levels, such as vessel size and frequency, is not contained within one centralized database, instead being maintained by each operator. Further, there is no historical database present which enables statistical modelling. Consequently, developing records of past service levels would be of great interest and benefit in future research.

# 8 References


Aarhaug, J., Fearnley, N., Gregersen, F. A., & Norseng, R. B. (2018). 20 years of competitive tendering in the Norwegian bus industry – An analysis of bidders and winning bids. *Research in Transportation Economics*, *69*, 97–105. https://doi.org/10.1016/j.retrec.2018.05.012

Aarts, L., & van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, *38*(2), 215–224. https://doi.org/10.1016/j.aap.2005.07.004

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press. https://doi.org/10.1515/9781400829828

Asplund, D., & Pyddoke, R. (2020). Optimal fares and frequencies for bus services in a small city. *Research in Transportation Economics*, *80*, 100796. https://doi.org/10.1016/j.retrec.2019.100796

Baird, A., & Wilmsmeier, G. (2011). Public tendering of ferry services in Europe . *European Transport*, *49*, 90–111.

Bårdsen, G., & Nymoen, R. (2014). *Videregående emner i økonometri*. Fagbokforlaget Vigmostad & Bjørke.

Basso, L. J., & Jara-Díaz, S. R. (2010). The Case for Subsidisation of Urban Public Transport and the Mohring Effect. *Journal of Transport Economics and Policy*, *44*(3), 365–372.

Bates, J., Polak, J., Jones, P., & Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, *37*(2–3), 191–229. https://doi.org/10.1016/S1366-5545(00)00011-9

Bignon, E., & Pojani, D. (2018). River-based public transport: Why won't Paris jump on board? *Case Studies on Transport Policy*, *6*(2), 200–205.

Blomquist, G. (1986). A utility maximization model of driver traffic safety behavior. *Accident Analysis & Prevention*, *18*(5), 371–375. https://doi.org/10.1016/0001-4575(86)90011-4

Börjesson, M., Fung, C. M., & Proost, S. (2017). Optimal prices and frequencies for buses in Stockholm. *Economics of Transportation*, *9*, 20–36. https://doi.org/10.1016/j.ecotra.2016.12.001

Börjesson, M., Fung, C. M., Proost, S., & Yan, Z. (2018). Do buses hinder cyclists or is it the other way around? Optimal bus fares, bus stops and cycling tolls. *Transportation Research Part A: Policy and Practice*, *111*, 326–346. https://doi.org/10.1016/j.tra.2018.03.023

Börjesson, M., Fung, C. M., Proost, S., & Yan, Z. (2019). Do Small Cities Need More Public Transport Subsidies Than Big Cities? *Journal of Transport Economics and Policy*, *53*(4), 275–298.

Börjesson, M., & Rubensson, I. (2019). Satisfaction with crowding and other attributes in public transport. *Transport Policy*, *79*, 213–222. https://doi.org/10.1016/j.tranpol.2019.05.010

Bråthen, S., Hervik, A., Odeck, J., & Sunde, O. (2004). Competitive tendering versus subsidy schemes in public transport services: comparing efficiency provision of ferry services. *Proceedings of the European Transport Conference (ETC) 2004, Held 4-6 October 2004, Strasbourg France*.

Bråthen, S., Hervik, A., & Sunde, Ø. (2007). Optimalt tidsforløp ved ferjeanbud. *Møreforskning Rapport 0710*.

Ceder, A. (2015). *Public transit planning and operation: Modeling, practice and behavior*. Taylor & Francis Group CRC Press .

Coelli, T., Prasada, D., O'Donnell, C., & Battese, G. (2005). *An Introduction to Efficiency and Productivity Analysis* (Second Edition). Springer.

Cowell, F. (2005). *Microeconomics: principles and analysis*.

DeJoy, D. M. (1989). The optimism bias and traffic accident risk perception. *Accident Analysis & Prevention*, *21*(4), 333–340. https://doi.org/10.1016/0001-4575(89)90024-9

Díez-Gutiérrez, M., & Tørset, T. (2019). Perception of inconvenience costs: Evidence from seven ferry services in Norway. *Transport Policy*, *77*, 56–67.

Elvik, R., Vadeby, A., Hels, T., & van Schagen, I. (2019). Updated estimates of the relationship between speed and road safety at the aggregate and individual levels. *Accident Analysis & Prevention*, *123*, 114–122. https://doi.org/10.1016/j.aap.2018.11.014

Farrow, S. (1998). Environmental equity and sustainability: rejecting the Kaldor-Hicks criteria. *Ecological Economics*, *27*(2), 183–188. https://doi.org/10.1016/S0921-8009(97)00135-3

Fielbaum, A., Jara-Diaz, S., & Gschwender, A. (2020). Beyond the Mohring effect: Scale economies induced by transit lines structures design. *Economics of Transportation*, *22*, 100163. https://doi.org/10.1016/j.ecotra.2020.100163

Findley, D. J., Anderson, T. J., Bert, S. A., Nye, T., & Letchworth, W. (2018). Evaluation of wait times and queue lengths at ferry terminals. *Research in Transportation Economics*, *71*, 27–33. https://doi.org/10.1016/j.retrec.2018.06.009

Fosgerau, M. (2009). The marginal social cost of headway for a scheduled service. *Transportation Reserach Part B: Methodological*, *43*(8–9), 813–820.

Gilbert, N., & Troitzsch, K. (2005). *Simulation For The Social Scientist* (Second Edition). Open University Press.

Haglund, M., & Åberg, L. (2000). Speed choice in relation to speed limit and influences from other drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *3*(1), 39–51. https://doi.org/10.1016/S1369-8478(00)00014-0

Hensher, D. A., & Stanley, J. (2003). Performance-based quality contracts in bus service provision. *Transportation Research Part A: Policy and Practice*, *37*(6), 519–538. https://doi.org/10.1016/S0965-8564(03)00006-5

Hensher, D. A., & Stanley, J. (2008). Transacting under a performance-based contract: The role of negotiation and competitive tendering. *Transportation Research Part A: Policy and Practice*, *42*(9), 1143–1151. https://doi.org/10.1016/j.tra.2008.05.004

Hensher, D., & Wallis, I. (2005). Competitive tendering as a contracting mechanism for subsidising transportation: The bus experience. *Working Paper, Institute of Transport and Logistics Studies*.

Hicks, J. R. (1939). The Foundations of Welfare Economics. *The Economic Journal*, *49*(196), 696. https://doi.org/10.2307/2225023

Holmgren, J. (2014). A strategy for increased public transport usage – The effects of implementing a welfare maximizing policy. *Research in Transportation Economics*, *48*, 221–226. https://doi.org/10.1016/j.retrec.2014.09.046

Holmgren, J. (2018). The effects of using different output measures in efficiency analysis of public transport operations. *Research in Transportation Business & Management*, *28*, 12–22. https://doi.org/10.1016/j.rtbm.2018.02.006

Hörcher, D., & Graham, D. J. (2018). Demand imbalances and multi-period public transport supply. *Transportation Research Part B: Methodological*, *108*, 106–126. https://doi.org/10.1016/j.trb.2017.12.009

Jansson, J. O. (1979). Marginal Cost Pricing of Scheduled Transport Services: A Development and Generalisation of Turvey and Mohring's Theory of Optimal Bus Fares. *Journal of Transport Economics and Policy*, *13*(3), 268–294.

Jansson, J. O. (1980). A Simple Bus Line Model for Optimisation of Service Frequency and Bus Size. *Journal of Transport Economics and Policy*, *14*(1), 53–80.

Jansson, J. O., Holmgren, J., & Ljungberg, A. (2015). Pricing public transport services. In *Handbook of Research Methods and Applications in Transport Economics and Policy* (pp. 260–308). Edward Elgar Publishing. https://doi.org/10.4337/9780857937933.00022

Jara-Díaz, S., Fielbaum, A., & Gschwender, A. (2017). Optimal fleet size, frequencies and vehicle capacities considering peak and off-peak periods in public transport. *Transportation Research Part A: Policy and Practice*, *106*, 65–74. https://doi.org/10.1016/j.tra.2017.09.005

Jara-Díaz, S., Fielbaum, A., & Gschwender, A. (2020). Strategies for transit fleet design considering peak and off-peak periods using the single-line model. *Transportation Research Part B: Methodological*, *142*, 1–18. https://doi.org/10.1016/j.trb.2020.09.012

Jara-Díaz, S., & Gschwender, A. (2003). Towards a general microeconomic model for the operation of public transport. *Transport Reviews*, *23*(4), 453–469. https://doi.org/10.1080/0144164032000048922

Jara-Díaz, S. R., & Gschwender, A. (2009). The effect of financial constraints on the optimal design of public transport services. *Transportation*, *36*(1), 65–75. https://doi.org/10.1007/s11116-008-9182-8

Jørgensen, F., & Hanssen, T.-E. S. (2019). Implications for traffic safety from car drivers' secondary task engagement – An economist's view. *Economics of Transportation*, *20*, 100136. https://doi.org/10.1016/j.ecotra.2019.100136

Jørgensen, F., & Mathisen, T. A. (2010). Using Standardized Revenue and Cost Norm Analyses to Reveal Subsidy Fraud in Contracted Public Transport Services. *Transport Reviews*, *30*(3), 299–313. https://doi.org/10.1080/01441640902986866

Jørgensen, F., Mathisen, T., & Solvoll, G. (2007). Kapasitet i fergesektoren. *SIB-Report 1/2007*.

Jørgensen, F., & Solvoll, G. (2018). Determining optimal frequency at ferry crossings. *Transport Policy*, *63*, 200–208. https://doi.org/10.1016/j.tranpol.2017.12.017

Kaldor, N. (1939). Welfare Propositions of Economics and Interpersonal Comparisons of Utility. *The Economic Journal*, *49*(195), 549. https://doi.org/10.2307/2224835

Kennedy, P. (2008). *A Guide to Econometrics* (Sixth edition). Blackwell Publishing.

Li, M. Z. F. (2002). The role of speed–flow relationship in congestion pricing implementation with an application to Singapore. *Transportation Research Part B: Methodological*, *36*(8), 731–754. https://doi.org/10.1016/S0191-2615(01)00026-1

Lindberg, G. (2005). Accidents. In *Handbook of Research Methods and Applications in Transport Economics and Policy* (pp. 154–170). Edward Elgar Publishing. https://doi.org/10.4337/9780857937933.00015

Martin, S. (2019). The Kaldor–Hicks Potential Compensation Principle and the Constant Marginal Utility of Income. *Review of Industrial Organization*, *55*(3), 493–513. https://doi.org/10.1007/s11151-019-09716-3

Mathisen, T. A. (2016). Competitive tendering and cross-shareholding in public passenger transport. *Transport Policy*, *48*, 45–48. https://doi.org/10.1016/j.tranpol.2016.02.011

Mathisen, T. A., & Jørgensen, F. (2012). Panel data analysis of operating costs in the Norwegian car ferry industry. *Maritime Economics & Logistics*, *14*(2), 249–263. https://doi.org/10.1057/mel.2012.2

Mathisen, T., & Solvoll, G. (2010). Service Quality Aspects in Ferry Passenger Transport - Examples from Norway. *European Journal of Transport and Infrastructure Research*, *10*(2), 142–157.

Mill, J. S. (1849). *Principles of Political Economy*. John W. Parker.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *The American Economic Review*, *62*(4), 591–604.

Nocedal, J., & Wright, S. (2000). *Numerical optimization*. Springer.

Nyborg, K. (2014). Project evaluation with democratic decision-making: What does cost–benefit analysis really measure? *Ecological Economics*, *106*, 124–131. https://doi.org/10.1016/j.ecolecon.2014.07.009

Odeck, J. (2010). What Determines Decision-Makers' Preferences for Road Investments? Evidence from the Norwegian Road Sector. *Transport Reviews*, *30*(4), 473–494. https://doi.org/10.1080/01441640903138640

Odeck, J., & Høyem, H. (2021). The impact of competitive tendering on operational costs and market concentration in public transport: The Norwegian car ferry services. *Research in Transportation Economics*, *90*, 100883. https://doi.org/10.1016/j.retrec.2020.100883

Oldfield, R. H., & Bly, P. H. (1988). An analytic investigation of optimal bus size. *Transportation Research Part B: Methodological*, *22*(5), 319–337. https://doi.org/10.1016/0191-2615(88)90038-0

Sadia, R., Bekhor, S., & Polus, A. (2018). Structural equations modelling of drivers' speed selection using environmental, driver, and risk factors. *Accident Analysis & Prevention*, *116*, 21–29. https://doi.org/10.1016/j.aap.2017.08.034

Sheng, D., & Meng, Q. (2020). Public bus service contracting: A critical review and future research opportunities. *Research in Transportation Economics*, *83*, 100938. https://doi.org/10.1016/j.retrec.2020.100938

Silbergberg, E. (1972). Duality and the Many Consumer's Surpluses. *The American Economic Review*, *62*(5), 942–952.

Simar, L., & Wilson, P. W. (1999). Estimating and bootstrapping Malmquist indices. *European Journal of Operational Research*, *115*(3), 459–471. https://doi.org/10.1016/S0377-2217(97)00450-5

Škurić, M., Maraš, V., Davidović, T., & Radonjić, A. (2021). Optimal allocating and sizing of passenger ferry fleet in maritime transport. *Research in Transportation Economics*, *90*, 100868. https://doi.org/10.1016/j.retrec.2020.100868

Slesnick, D. (1998). Empirical Approaches to the Measurement of Welfare. *Journal of Economic Literature*, *36*(4), 2108–2165.

Strøm, S., & Vislie, J. (2007). *Effektivitet, fordeling og økonomisk politikk*. Universitetsforlaget.

Tarko, A. P. (2009). Modeling drivers' speed selection as a trade-off behavior. *Accident Analysis & Prevention*, *41*(3), 608–616. https://doi.org/10.1016/j.aap.2009.02.008

Thonemann, U. (2005). *Operations Management. Konzepte, Methoden und Anwendungen*. Pearson Education Deutschland.

Tirachini, A., & Antoniou, C. (2020). The economics of automated public transport: Effects on operator cost, travel time, fare and subsidy. *Economics of Transportation*, *21*, 100151. https://doi.org/10.1016/j.ecotra.2019.100151

Tirachini, A., Hensher, D. A., & Rose, J. M. (2014). Multimodal pricing and optimal design of urban public transport: The interplay between traffic congestion and bus crowding. *Transportation Research Part B: Methodological*, *61*, 33–54. https://doi.org/10.1016/j.trb.2014.01.003

Varian, H. (1992). *Microeconomic Analysis. Third Edition.* . W. W. Norton & Company, Inc.

Vigren, A. (2018). How many want to drive the bus? Analyzing the number of bids for public transport bus contracts. *Transport Policy*, *72*, 138–147. https://doi.org/10.1016/j.tranpol.2018.08.008

Vigren, A. (2020). The Distance Factor in Swedish Bus Contracts How far are operators willing to go? *Transportation Research Part A: Policy and Practice*, *132*, 188–204. https://doi.org/10.1016/j.tra.2019.11.010

Waters, W. G., Evans, J., & Caravan, J. (1996). Subsidy policy on low volume ferry routes. *Transport Policy*, *3*(3), 111–121. https://doi.org/10.1016/0967-070X(96)00013-3

Wood, A. (1999). *Introduction to numerical analysis*. Addison Wesley.

# 9. Papers

# Paper I

Research in Transportation Economics 90 (2021) 100883

Contents lists available at ScienceDirect

Research in Transportation Economics

journal homepage: http://www.elsevier.com/locate/retrec

# The impact of competitive tendering on operational costs and market concentration in public transport: The Norwegian car ferry services

James Odeck [a,b,c,*], Harald Høyem [c,d]

[a] The Norwegian University of Science and Technology, 7491, Trondheim, Norway
[b] Molde University College, PO Box 6405, Molde, Norway
[c] The Norwegian Public Roads Administration, PO box 8142, 0033, Oslo, Norway
[d] Urbanet Analyses, PO Box 337, 0101, Oslo, Norway

A R T I C L E I N F O

*JEL classification:*
R40
R41
R42
R48

*Keywords:*
Competitive tendering (CT)
Transport sector
Car ferry services
Cost-efficiency
Panel data

A B S T R A C T

The literature on how competitive tendering (CT) affects the operational costs of transportation services has been inconclusive; some concluded that it reduces operational costs, while others maintained that it increases or does not affect operational costs. We add to the literature by assessing the impact of CT on the operational costs of car ferry services in the case of Norway. Note the similarity with public transport: car ferries transport vehicles from one road point to another, like public transport does with passengers. Furthermore, we study the impact of CT on market concentration. The data comprises 53 ferry links across 8 years, yielding 424 observations. The results reveal that: (i) implementation of CT significantly lowered operational costs; and (ii) market concentration increased. These findings suggest that CT might not lead to free competition in the long run but rather to monopolistic/duopolistic tendencies, contrary to its intentions. We urge policymakers to reconsider CT carefully because its results could be counter to intentions.

## 1. Introduction

Over the last three to four decades, governments worldwide have been concerned with the way in which public resources, including the transportation sector, are used. The result of this concern has been that the transportation sector has been a subject of reforms to promote efficiency and/or to reduce costs in the provision of public services; see, for instance, Hensher and Stanley (2010). Such reforms have included the privatization of government-owned transportation enterprises and/or the subjection of the delivery of public transportation services to competition. The latter type of reform is called competitive tendering (CT). It has been most promoted in the provision of public transportation services, with the major objective being the containment of costs to governments; see, for instance, Hensher and Wallis (2005). In the transportation sector in general, CT implies that a contract to operate or provide a transportation service is subjected to competition, in which the bidder with the lowest cost, albeit the bidder offering the most attractive cost/quality combination, is awarded the contract to operate for a period of time, after which the operation is again subjected to competition. CT is a contrast to the traditional way of providing public transportation services, which has been that one incumbent is awarded a contract year after year based on negotiations using some standard cost norms as the base negotiation and that the services to be delivered are pre-specified by the authorities. Typically, as in the Norwegian case for car ferry transport, for example, it means that the incumbents were reimbursed the difference between the actual operational costs and their ticket earnings, whereas the transportation authorities determined the ticket earnings, which were not sufficient to cover the operational costs.

The attractiveness of CT over the traditional ways of providing transportation services is appealing from an economics (or a cost saving) point of view for two particular reasons, both of which have been advocated by economists and transportation planners for a long time; see, for instance, Savas (2000) and Hensher and Wallis (2005). The first reason is that competitors for tenders will most likely strive to deliver the most efficient or the least costly way of providing the pre-specified services because it will both maximize the company's probability of winning the tender competition, and it will maximize the company's profit if it is chosen as the winner of the contract. Second, the incumbent could be replaced by others offering lower costs for the provision of the same services, hence lowering government costs for the provision of transportation services. The end results, as believed by economists and transportation planners, are that the government expenditures on

* Corresponding author. The Norwegian University of Science and Technology, 7491, Trondheim, Norway.
*E-mail address:* james.odeck@vegvesen.no (J. Odeck).

https://doi.org/10.1016/j.retrec.2020.100883
Received 24 July 2019; Received in revised form 10 April 2020; Accepted 22 May 2020
Available online 18 July 2020

transportation services are reduced, and/or the quality of services provided is increased.

Although CT has been promoted as a mechanism for providing transportation services at the least expensive level for governments, there is still no conclusive evidence on whether it reduces government expenditures; see, for instance, Pollitt and Bouckaert (2004) in the case of the public sector in general and Hensher and Wallis (2005) in the case of the transportation sector. Hensher and Wallis (2005), for instance, observed that, when savings do occur, they typically result from a first-time tendering process and can retract significantly when retendering. Moreover, recent empirical and theoretical work has suggested that CT affects the ownership structure – see Aarhaug and Fearnley (2016) – and that such changes can reduce the efficiency of CT; see, for instance, Mathisen (2016). An overall observation from the literature is that the working of CT is circumstantial, depending on the specific situation considered. This fact was evidenced in Hensher and Wallis (2005), ceteris paribus, attesting that the literature has been inconclusive with respect to how CT affects the operational costs/efficiency of transportation services.

In this paper, we assess the impact of CT on the operational costs and the impact on market concentrations in the case of the Norwegian car ferry sector. We thus contribute to the debate on the workings of CT in the transportation sector, with the Norwegian car ferry sector as a case study, understanding that car ferries are an integral part of the road systems in many countries, e.g., Norway, Denmark, Sweden, Greece, Scotland, Canada, the United States and Turkey. For a review on tendered ferry services in Europe, see e.g. Baird & Wilmsmeier (2011). Car ferries transport vehicles from one road end to the other and are thus very similar to public transport in operations. Another similarity to public transport is that car ferry link services, like bus or rail routes, are operated by private companies and are largely subsidized by central or regional governments. Therefore, this study is comparable to studies that have examined the impact of CT on costs in public transport.

The scientific contributions of this paper to the transportation literature are twofold. First, we bring experience on how CT has influenced the operational costs of the car ferry sector, which has been considered less often in the literature, in terms of the impacts of CTs, although it is an important sector in the provision of transportation services worldwide. The operations of the car ferry sector are very similar to the provision of public transport services in general; thus, the conclusions drawn in the present study are value added to the overall literature on how CT influences the operational costs of public transport services. Second, we address the impact of CT on market concentration. If CT leads to market concentration, its long-term effects could be monopolistic/duopolistic tendencies rather than free competition as intended by CT, contrary to the intentions of implementing CT. The literature on transportation has not addressed this issue adequately.

We proceed as follows in the rest of the paper. In section 2, we offer a brief overview of the developments in subsidies/procurement regimes in the Norwegian car ferry sector. Section 3 is a short literature review on how CT has worked in the transportation sector. Sector 4 describes the methodological framework used to assess the impact of CT on operational costs and the impacts of CT on market concentration. Section 5 describes the data used. Section 6 presents the results, including the impacts of CT on market concentration. Section 7 provides concluding remarks.

## 2. Procurement developments in the Norwegian car ferry sector

Norway, like, e.g., Scotland and Canada, has many fjords and islands, making fast links by roads, tunnels or bridges costly to build. To connect the road transportation network across fjords and to islands, car ferries are traditionally used. Each ferry link is therefore an integral part of the Norwegian road network in the sense that the ferry link primarily transports vehicles from one road end to the other. For example, Jørgensen et al. (2011) estimated the total social surplus of the Norwegian ferry sector to be 4.3 billion NOK,[1] which underlines the importance of the service. Moreover, the Norwegian authorities continue to retain an interest in reducing costs and increasing the efficiency in the ferry sector, as stated in a press release from the Ministry of Transportation on October 04, 2014 (Oslo Economics, 2016). The statement expressed a need to "[…] reduce costs in the ferry sector". The government also survey the markets status on a yearly basis (see e.g. Oslo Economics, 2019), and has issued projects to further develop strategies to increase efficiency in the sector (Oslo Economics, 2016).

As of 2009, there were 130 ferry links in the national Norwegian trunk road system and approximately 70 in the regional road system. As the Norwegian government reformed the road system and downgraded approximately 40 percent of the trunk road system to regional roads, the number of ferry links serving the trunk road system was reduced to only approximately 17 links by 2014. However, ferry links remain an important part of the total road system, irrespective of whether they are trunk or regional road systems.

Private companies operate all of the Norwegian car ferry links, subsidized by links operated by the central government in the case of trunk roads and by regional governments in the case of regional roads. A given company can operate several links. The simple reason for subsidization, very similar to what has been observed in the public transport sector, is that there are high costs associated with running ferries, and given the relatively low traffic volumes, ferry links are run at a loss. Governments dictate the fares such that the system users are not penalized much more for being dependent on ferries, compared to other road network users in general. A traditional system of subsidization has been that the ferry link operators are remunerated for the difference between the income from fares and their actual operational costs.

The system of remunerating ferry link operators has evolved over the years. Until 1990, subsidies to ferry link operators were awarded ex post, i.e., after operations were accomplished, based on some historical costnorms that could be observed from operational data. This system is the well-known cost-plus system in the literature on contracting, and it entailed that the companies were remunerated the costs expected to be incurred, as well as some markup to ensure that the operations would be conducted. The markup was 10 percent of the total operational costs. It is important to remember that cost-plus contracts were awarded to the same incumbent who has been there since time immemorial, with negotiations every sixth year. There was no competition for contracts to provide ferry link services; hence, operators were virtually monopolists. It is therefore clear that the incentives for cost efficiency in this period were weak; ferry link operators had no incentives to reduce costs and/or operate efficiently because the remuneration was awarded beforehand and without competition, and there was no threat of losing the contract in the subsequent period.

To enhance cost efficiency while still using the cost-plus subsidization regime, the Norwegian government decided in 1991 to award the contracts ex ante, i.e., before the operations were performed. The intention was that the operators would be encouraged to find more efficient ways of providing services and hence increase their profits. Despite this change in subsidization regime, there was no noticeable change in either the efficiency or the costs of providing ferry services.

In the meantime, between the two systems mentioned above, competition in the supply of public services e.g., transportation services, was gaining momentum in Europe. The Norwegian government seemed to be aware of the developments and wanted to exploit the potential for competition. The transport act was amended in 1991, allowing for CT from 1994. The Norwegian Public Roads Administration (NPRA), charged with overseeing road transport, took this opportunity, and over the years that followed, it exposed six ferry links to CT on a trial basis (Bråthen, Hervik, Odeck, & Sunde, 2004). The first were CT exposed from 1997 and the last two from 1999. From 2003, full CT was

[1] 1 NOK is currently equal to 0.09 EURO (March 2020).

implemented on four ferry links (Bråthen et al., 2004, p. 408). Since then, an increasing number of ferry links have been exposed to CT: seven in 2007, eight in 2008 and 32 in 2010.

The regional offices of NPRA design the CT documents. NPRA officials and lawyers undertake the evaluation using a pre-specified template. The administrative management of the NPRA makes the final choice of operator. In the calls for the first six tenders in the period of 1997–1999, six to nine bids were received for each ferry link. Significant differences in the required subsidy levels were found in the tenders for each link. In five of six cases, the incumbent operators won the tender (Hervik & Sunde, 2001). The duration of CT contracts varies between 3 and 12 years, depending on the need for asset-specific investments. In recent years, trunk road ferry CT has resulted in investment by private operators in newer and larger ferries. It has also resulted in mergers, in which operators have strengthened their market positions. It should be noted here that CTs are per ferry links but in a few circumstances are per a group of links, with the bidder offering the lowest costs in combination with highest level of services winning the contract. However, the cost and services offered for each link are known; thus, an analysis of each link is the most relevant to consider in the subsequent analyses.

The selection criteria in the tendering process have evolved over the years. Prior to 2012 (part of the period for which we have data), the government specified the criteria a bidder should fulfill with respect to service levels, size of vessels etc., and the bidder having the lowest price won the tender.[2] After 2014, environmental factors have been introduced as a criterion for the trunk road ferries. Regional ferries also tend to include environmental factors, but more often as minimum requirements or a bonus. For the trunk road ferries, environmental factors are related to total energy consumption, along with carbon and NOx emissions. Most trunk links weigh price at 70% and emissions at 30% after 2014.

## 3. Literature review

Critics of CT have contended that CT has not consistently delivered on the promised high quality and low-cost services; see, for instance, Krugman (2003, p. 17). Others have contended that CT could even lead to a reduction in the quality of services provided; see, for instance, Hart, Schleifer, & Vishny (1997). Nonetheless, there are also proponents for CT; see, for instance, Savas (2000) for an overview. A general argument by proponents of CT is that CT can and will most likely lead to companies striving to deliver the most efficient or least costly way of providing the pre-specified services because doing so will both maximize the company's probability of winning the tender competition and maximize its profit if chosen as the winner of contracts. This outcome, it is argued, will lead to a reduction in costs to the government. Ceteris paribus, this last claim implies that CT leads to a reduction in operational costs.

There is a growing literature that has addressed CT in relation to public transport, and some studies have specifically addressed the Norwegian ferry sector. This literature includes and is not limited to the following: Norheim (1999), Bråthen et al. (2004, p. 408), Hensher and Wallis (2005), Houghton and Hensher (2005), Kain (2006), Economic and Policy Services Pty Ltd. (2007), Ian Wallis Associates Ltd. (2007), Myers and Ashmore (2007), Bray and Wallis (2008), Hensher and Stanley (2008), Stanley and van de Velde (2008), Walters and Cloete (2008), Nash and Wolański (2010), Wallis et al. (2010) and Mathisen (2016). The key findings of this literature on the performance of CT in transportation, as summarized in Wallis et al. (2010) but with certain modifications, are as follows.

i). Competitive tendering can lead to "adverse selection", i.e., selection of the bidder who is most successful at identifying and exploiting the flaws in the tender appraisal and contract documents, rather than selection of who offers the best value for the money. The competitive tendering process might encourage a potential operator to bid an unreasonably low price to win a contract and then lead to a 'winner's curse' as the operator attempts to provide services at the agreed price or to renegotiate a higher price. Contracts that are financially unsustainable can result in an operator becoming unduly focused on cost minimization rather than on the quality and added value of the services and could eventually lead to the operator surrendering the contract. Competitive tendering thus requires a sound tender assessment system that can identify trade-offs between price and quality and can determine whether a bid price is likely to be sustainable over the term of the contract.

ii). Transaction costs will be significantly higher with competitive tendering, both for the public authority and for the operators involved (including unsuccessful bidders). Other transaction costs that will also be higher with competitive tendering include: (a) public authority costs associated with any operator establishment/winding-up and with adjustment to a new operator; (b) operator costs for an incoming operator that must establish a new operation and winding-up costs for the unsuccessful outgoing operator; (c) user costs, through deterioration of the quality of services provided by an unsuccessful outgoing operator due to reduced motivations following the announcement of the results of tendering, and service disruptions at the commencement of operations by a new operator due to the unfamiliarity of a new operator with issues such as routes and operations for an initial period of operation; and (d) user and operator costs during the period until a new operator establishes links with the community and develops the necessary skills and knowledge.

iii). Competitive tendering is more appropriate when a strong market of keen suppliers is anticipated. Competitive tendering is likely to yield better results in the early stages of any outsourcing strategy, when market price and quality offerings are untested, and efficient market prices are not well established.

iv). Competitive tendering has advantages in terms of its ready accountability and transparency in the use of public funds. It involves a transparent process of operator selection, open to all parties with the interest and ability to provide the required services, and such a process ensures good accountability for the use of public funds. Negotiation is generally a less transparent process, although the process and its results can be open to scrutiny.

v). Competitive tendering must attract strong market interest if it is to be effective in securing an efficient price for the desired service, which in turn requires that the tender conditions be clear and attractive to potential bidders and that bidders consider the tender evaluation process to be unambiguous and impartial. The potential for competitive tendering leading to excessive market consolidation can be addressed through the use of other strategies, for example, market share limits.

In short, the international literature, according to Wallis et al. (2010), has not provided clear evidence of when competitive tendering for transport services is more appropriate, compared to other forms of procurements, such as negotiations.

In their review of the literature on benchmarking the outcome of CTs in the transportation sector, Nash and Wolański (2010) found that competitive tendering had generally been successful in terms of quality and costs, but problems had occurred in a number of cases. They recommended that careful attention be paid to the design of tendering exercises, details of the contract, risk-sharing arrangements and the approaches to any renegotiations found to be necessary. Overall, their

[2] There existed a small exception; if any tender delivered "significantly better" service with respect to environmental factors, it could be selected, provided it was less than 5% more expensive than the cheapest one.

conclusions concur with those of Wallis et al. (2009) in many respects. To this conclusion, we add that the result of CT in transportation is not conclusive but rather is circumstantial; CT might work in some cases and not in others, which is where the contribution of this paper comes in: we add yet another different circumstance, namely, the Norwegian ferry sector. Although several studies have addressed the efficiency and productivity of the Norwegian ferry sector, some have only addressed the efficiency performance in general, e.g., Førsund (1992), Odeck (2008), Jørgensen et al. (2011), Mathisen and Jørgensen (2012) and Jørgensen and Mathisen (2010). Other studies have addressed the impact of CT on the technical efficiency performance of the Norwegian ferry sector; see, e.g., Odeck and Bråthen (2009). However, the impact of CT on operational costs, which is the subject of this paper, has not been addressed intensively in the case of the Norwegian car ferries in the manner in which we do in this paper.

## 4. The statistical/econometric framework

Our stated objectives are to infer two different issues, which are the impact of CT on the operational costs of ferry link services and the impact of CT on market. Addressing these issues require different frameworks. In the following, we explain the econometric/statistical approaches that we used to ascertain each of these issues.

### 4.1. The impact of CT on the operational costs of ferry link services

Other researchers investigating the cost structure of ferry operations have used panel-data techniques as the preferred method (Jørgensen & Mathisen, 2010; Mathisen & Jørgensen, 2012). The model of Mathisen and Jørgensen (2012) was formulated in cost levels, whereas Jørgensen and Mathisen (2010) used a translog-function. In this paper, we have adopted the method of Mathisen and Jørgensen (2012) for the sake of model parsimony, as we are mainly concerned with the average effect of tendering and not the overall cost structure, although a non-linear specification might have been used as well, which is discussed below and in appendix 1. Our basic model is defined as follows:

$$y_{it} = \alpha + \varphi_i + \omega D_{it} + \beta_1 SKM_{it} + \beta_2 PCE_{it} + \beta_3 PCE_{it} \times L_i + \beta_4 PCE_{it} \times SW_i + \beta_5 FSIZE_{it} + \beta_6 NFERRIES_{it} + \beta_7 AGE_{it} + \beta_8 FCRIS_t + \lambda \times t + e_{it}, \quad (1)$$

where $y_{it}$ is the total operational costs for ferry link $i$ in year $t$. The cost variable includes fuel consumed, lubricant, maintenance and labor costs of operating the ferry link, and is adjusted for inflation and measured in 2010 NOK.[3] For inflation adjustments, we used the general consumer price index according to Statistics Norway (2020). As we aim to assess the impact of tendering both in absolute and relative terms, we use two different formulations of model 1. In the first formulation (model 1A), we use costs in levels which gives us the average reduction in costs in NOK per crossing. In the second formulation (model 1B), we transform the cost using the logarithm, substituting $\ln y_{it}$ for $y_{it}$. This enables interpretation of the coefficient as the average percentage change in costs from tendering the link.[4] Further, α is a constant to be estimated; $\varphi_i$ is a panel-specific constant; $D_{it}$ is a dummy variable for whether the ferry link $i$ in year $t$ is a CT; $SKM_{it}$ is the number of sailed kilometers per year ferry link $i$ in year $t$, $PCE_{it}$ is the number of Passenger Car Equivalents (PCE[5]) transported, $L_i$ is the length of link $i$, $SW_i$ is a dummy variable indicating if the link traverses sheltered or open waters, $FSIZE_{it}$ is the average age of ferries, $NFERRIES_{it}$ is the number of ferries operated at the link, $AGE_{it}$ is the average age of ferries, $t$ is a time trend, $FCRIS_t$ is a dummy indicating the year the financial crisis started (=1 in 2009) and $e_{it}$ an error term. Both the average ferry size and age were calculated as an unweighted arithmetic average of all ferries serving each link. The terms $PCE_{it}$, $PCE_{it} \times L_i$ are the same as in the specification of Mathisen and Jørgensen (2012). However, they also included the sheltered waters ($SW_i$) in their specification, as higher service requirements apply in open waters that might raise costs. When using the fixed effects estimator, one cannot estimate the effect of variables that remain constant within each panel unit. Consequently, we use the interaction variable between open waters and the number of PCE-transported ($PCE_{it} \times SW_i$) to correct for possibly higher sailing costs in open waters. Using the interaction, the variable is not constant within each panel unit, which enables us to align more closely with the framework of Mathisen and Jørgensen (2012). Mathisen and Jørgensen (2012) were able to estimate a parameter using only the indicator variable of sheltered waters ($SW_i$) as they used the Random Effects estimator which enables identification of variables that are constant within each panel unit. Their model had fewer variables than ours, and data from 1995 to 2000 and 2003–2005 which might explain why a different estimator was deemed appropriate.

Moreover, we have also added several other variables based on the availability in our dataset and their possible relevance for costs, for which we now provide a short rationale. The number of sailed kilometers is important to include as not just the length, but also the frequency of the link influence costs. A number of variables describing the state of the ferries are added. The average size of ferries in PCE as larger ferries may be more expensive to operate, the average age of ferries as older ferries may exhibit a different cost levels than newer ones, and the number of ferries serving the link as more ferries may exploit possible economics of scale in maintenance facilities etc. The dummy indicating the start of the financial crisis is included to control for any possible changes in demand or costs due to the overall shock to the economy as a result of the crisis. We are the first to include this factor into a model of ferry costs, as previous studies used data prior to the financial crisis. A time trend is included to capture any trend in costs, such as increasing real prices or wages.

Our parameter of interest is $\omega$, which may be interpreted as the average reduction in costs when tendering a link in levels (1A) or percent (1B). Even though our model is a simple linear regression, it contains more variables having a possible impact on costs than earlier studies.

While we use Equation (1) to assess the impact of CT on operational costs, it has two shortcomings that can be resolved given adequate data, which, in the present study, were not available. The first is that it does not account for the "bad control problem" as has been addressed in the literature of modern econometrics; see, for instance, Angrist and Pischke (2008). The "bad control problem" occurs if an explanatory variable is correlated with both of the other explanatory variables, such as CT, and with the independent variable, which in our case is the operational costs. If such variables are not controlled for, the parameter estimates will be statistically biased. A potential "bad control problem" situation in our case study is that, whereas we are interested in the "pure" effect of a change from cost norms to CT on operational costs, implementing CT can, in some instances, also entail a change in the service levels to be provided, e.g., increased frequency. Thus, if the "bad control problem" was not considered, the results would be a mix of two different effects as follows: (i) efficiency could be increased, leading to a lower cost level (pure effect); and (ii) if the change to CT also entailed a change in the service levels, the effect of implementing the CTs will, in addition to (i), be transmitted to changes in, e.g., the number of sailed kilometers and/or frequency of ferries (the additional effect). The sum of (i) and (ii) would be the "total" effect. Thus, one could argue that the "total" effect of a CT is lower than the "pure" efficiency effect since operational variables are altered during the tendering process, which, if it entails increased frequency, will ceteris paribus increase costs and, hence, decrease the efficiency of CT. With appropriate data, it is possible to

[3] Currently (2018), 100 NOK is equal to 11.87 US dollars and 10.27 euros.

[4] We would like to thank an anonymous reviewer for pointing out this possibility to us.

[5] PCE is a normalized measure of a vehicle's size. One passenger car (less than 6 m) is the same as 1.025 PCE. A heavy truck is equivalent to 10.682 PCE, according to Jørgensen and Solvoll (2018).

disentangle these two effects using SEM analysis, as addressed in, e.g., Bardal and Mathisen (2015) in a transport-related application. However, in our case, the data on the extent to which moving from cost norms to CT also entailed changes in service levels were not readily available. Available data on the earlier implementations of CT, however, revealed that CT did actually entail changes in service levels., see e.g., Hervik and Sunde (2001). Thus, we apply Equation (1) but warn that the results must be interpreted with some care; they account only for the "pure" effect of tendering and does not include effects of changes in the service levels.

The second issue worth noting is that we do not employ nonlinearity in the cost function. The linear model will, however, provide a simple interpretation of the coefficients and a parsimonious alternative to more complex models. However, we have tested a non-linear model yielding consistent results, but could not be proven preferable to the linear one, which is documented in appendix 1. The linear model should provide a sufficient approximation since our variable of interest is the effect of CT, which is a categorical variable. Possible non-linearity is more likely to be associated with the number of kilometers sailed and the number of vehicles transported due to economics of scale. Thus, we proceed assuming linearity in the cost function.

There are two approaches to the estimation of Equation (1): the fixed-effects (FE) and the random-effects (RE) regression models. The RE regression model assumes that the individual ferry link-specific characteristics are correlated with the explanatory variables, whereas the FE regression models assumes they are not. If the former is appropriate, then the feasible generalized least squares (FGLS) estimation is the suitable approach, and if the latter is appropriate, then the least squares dummy variable (LSDV) estimation is the appropriate approach; see, for instance, Baltagi (2005). To decide between the FE and RE, researchers have relied on Hausman's test; see, for instance, Hausman (1978) .[6] Hausman's test controls for the violation of the random effects modeling assumption that there is a correlation between the independent variables and the unit effects. If this correlation does not exist, for instance, in relation to Equation (1), the estimates of our variable of interest ($\omega$) in the fixed effects model ( $\widehat{\omega}_{FE}$) should be similar to the estimates of $\theta$ in the random effects model ($\widehat{\omega}_{RE}$). The test statistic $H$ for the Hausman test is a measure of the difference between the two estimates:

$$H = (\widehat{\omega}_{RE} - \widehat{\omega}_{FE})^{'} [Var(\widehat{\omega}_{FE}) - Var(\widehat{\omega}_{RE})]^{-1} (\widehat{\omega}_{RE} - \widehat{\omega}_{FE}) \tag{2}$$

$H$ above has a Chi-square distribution with degrees of freedom equal to the number of regressors in the model. A significant $p$ value is taken to be evidence that the two models are significantly different; hence, the RE model is rejected in favor of the FE model. Although we use this test in this paper, it should be mentioned that this test has several weaknesses; see, for instance, Clark and Linzer (2015) for further explanations these weaknesses. Further, to correct for dependency between observations at each link over time, we use so-called panel-robust standard errors (see, e.g., Baltagi (2005)). Such standard errors correct for observations at the same link perhaps being correlated over time. We assess both variants of model 1 using the Hausman-test. That is, both the log-transformed equation and the one in levels.

### 4.2. The effect of CT on market concentration

The next issue that we examine in this paper is whether the implementation of CT has affected market concentration in the Norwegian ferry market. Market concentration is a function of the number of ferry companies and their respective shares of total production in the ferry market. A measure of market concentration is useful because it may reflect the degree of competition in the market. If CT leads to market concentration, it means that the degree of competition may have been reduced and vice versa. In the former case, the results of the implementation of CT will be the reverse of the intention with CT; the aim of CT is to increase competition, but it leads to less competition through market concentration through the formation of duopolistic and/or oligopolistic markets.

Next, to measure the degree of market of concentration across time, including the periods before and after the implementation of CT, we used the most commonly accepted measure of market concentration, the Herfindahl-Hirschman Index (HHI); see, for instance, Rhoades (1993). The HHI is the sum of squared market shares for all companies, calculated as follows:

$$HHI_t = \sum_{i=1}^{m} S_{it}^2 \tag{3}$$

where $S_{it}^2$ is the market share of the $i$th firm in year $t$ and where $m$ is the number of firms. In our case, we defined the market share as the percentage of links managed by a given firm or company. An interpretation of HHI is that the lower that the index is, the closer that the market is to being in perfect competition, and the higher that the index is, the closer the market is to be a monopolistic/oligopolistic market.

We have calculated the HHI-index in two different ways. In the first instance (HHI by each company), we do not adjust for ownership groups but simply consider the development in individual companies' market shares according to equation (3). In the second instance (HHI-index by ownership group), we adjusted for ownership groupings. Ownership grouping in this case imply that companies are grouped together under their parent company. For instance, companies Fjord1 MRF and Fjord1 Fylkesbaatane were subsidiaries under a parent company named Fjord1. Consequently, as they were owned by the same parent company, it can be argued that they should be viewed as one single company. Therefore, when calculating the HHI-index by ownership group the market share of the parent company is calculated by combining the shares of the subsidiaries. In cases where separate companies were not owned by the same parent company but were later merged into one large as in the case of Fosen Trafikklag and Namsos Trafikklag which were merged into FosenNamsos, they are considered as separate prior to the merger and the same company after the merger. The same consideration was made with regards situation where smaller companies were bought by larger companies as in the case where Torghatten bought Hurtigrutens ferry operations.

We however, encountered a potential problem when calculating the HHI-indices as described above. For some links that were initially served by separate companies, our data set contained only the name of the parent company which they were merged into at a later point in time. For instance, we know that some links were operated by *either* HSD[7] *or* Stavangerske prior to their merger into Norled in 2007, but we do not know exactly which company of the two operated the link. Therefore, in such a case we assumed that the links were operated by the same (Norled in this example) both before and after 2007. This represents a potential weakness with our analyses because the market concentration index will be inflated. However, as we are primarily interested in the change in the HHI-index, it actually gives a more conservative estimate on that change. To see this, consider a situation where two links were previously operated by two different companies and later merged into one company, but we regard them as merged before and after the tendering. In such a case, the *change* in market concentration will not be counted for. Consequently, our estimates on the overall change in HHI-index present a lower bound. As we are interested in the impact of tendering on market concentration as measured by the HHI-index, our conclusions will most likely remain valid but on the conservative side, i.e., if changes in market concentration (HHI-indices) are found. Lastly, there was no

[6] The Hausman-test is used for both model 1A (cost in levels) and 1B (cost in natural logarithms).

[7] HSD was named Tide Sjø between 2006 and 2007.

record of any break-up of companies in contrast to merging in our data set.

## 5. Data

The Norwegian Public Roads Administration (NPRA), which is charged with overseeing the construction and management of public roads within the trunk system, including the ferry links that serve the road system, provided the data for the present analyses. These data are available from the NPRA upon request at www.vegvesen.no. The relevant data for the present analyses includes operational data per ferry link for each individual year from before and after CT was implemented. The period under study is 2003–2010. One might thus ask why data that are even more recent were not used. The answer is that, after 2010, the operational data were no longer reported to the NPRA. A request for more recent data from the companies that operate links proved useless; companies regard their data as confidential given that CT has been implemented. However, we believe that the available data for the period of 2003–2010 are sufficiently large and cover a sufficiently long period to infer the impact of CT on the operational costs of ferry links. Before 2003 only a small percentage i.e., only 6 ferry links had been subjected to trial tendering. It was as from 2003 that tendering in the ferry sector was officially inaugurated.

The data set that we use to assess the impact of tendering and market concentration derives from one database but divides into two parts as follows:

- **Main dataset**: This is the main dataset discussed in the preceding paragraph and contains information on operational cost by ferry links, tendering status and identities of companies that operate individual ferry link at every point in time. It is used to assess the overall impact of tendering.
- **Dataset for market concentration:** This is an extract of the main dataset based on which companies operated which ferry links before and after tendering was inaugurated and; who merged with after competitive was inaugurated. It is used to derive the HHI- indices.

Our data comprise 53 ferry links out of 130, representing approximately 41% of all of the trunk road ferry links that existed in Norway in 2009 (the second last year in our sample). The links were selected based on (i) data availability concerning our variables of interest and (ii) whether the links were in operation for the whole time period that we study; as some were replaced by fixed links (roads or tunnels) in that period.

However, it is important to underline two points: Firstly, as some links are significantly larger than others, it is important to not only look at the number of crossing but also the total amount of PCE (demand) that is covered by our sample. Measured as the number of PCE transported, our dataset covers 69% percent of PCEs transported in the Norwegian trunk road system as of 2010. Earlier studies as in e.g., Secondly, Mathisen and Jørgensen (2012) and Jørgensen and Mathisen (2010) had a lower number of links in their studies at 40 and 51 respectively[8] are hence, lower percentage PCE than ours.

Given that we have a time series of 8 years, the total number of observations (8 × 53) is 424, which is statistically sufficient to derive robust conclusions; depending on the number of variables included in the equations. It should be noted however, that only five of the links we consider in this study included links that served three ports and only two links served more four or more ports. Consequently, our results are mainly valid for ferry links that call at only two ports, which is typical for Norway.

[8] Jørgensen & Mathisen (2010) had a different number of links for each year, as links were closed during their period of study. 51 is the number we have interpreted as giving them a balanced panel.

Summary statistics of the variables used in this paper are shown in Table 1. The operational costs and the variables that explain it vary a great deal as can be observed in the min, max and the standard deviation. The mean value for the dummy for CT is low at 0.13 while the max is at 1 (as it is a dummy variable). It is evident that the dataset is comprised of both small and large links. For instance, the largest link transported 2.3 million PCE's in 2010 while the smallest one transported only 32,000 PCE. A closer examination of data set showed that two links were significantly larger than the others and hence, could be regarded as outliers. These links are Halhjem-Sandvikvåg and Mortavika-Arsvågen with respectively 2.5 and 4 standard deviations above the mean. Therefore, as robust check, we run our analysis both with and without these links, to ascertain whether the effect can be attributed only to the largest crossings. These robustness results are addressed and presented in appendix 2.

## 6. Estimation results

Recall that the objective of this paper is to infer two different issues regarding the workings of CT as follows: (i) CT's impact on the operational costs of ferry links and (ii) CT's impact on the market concentration of ferry link services. Below, we present the results according to these three objectives.

### 6.1. *CT's impact on operational costs of car ferry services*

To infer the impact of CT on operational costs, we estimated two alternative models where, one assumed cost in levels (model 1A) and the other assumed cost in logarithms (model 1B). The results of both the models are presented in Table 2. Further robustness tests of these models are addressed and reported in the appendix 1 and 2.

We first performed an initial statistical test for both model 1A and model 1B to find whether it is the fixed-effects (FE) or the random -effects (RE) regression model that applies. We used the Hausman test, as explained in the methodology section. With 5 degrees of freedom in both cases, the H-statistic for the Hausman test were 28.69 and 76.87 for models 1A and 1B respectively. The concurrent p-values were both at 0.0001 respectively; meaning that for both models, RE formulation was rejected in favor of the FE formulation. We henceforth proceeded with the FE formulation as the most appropriate formulation for both model 1A and 1B.

Consider next, the explanatory power of both the models as measured by the adjusted-$R^2$ shown in the lowermost part of the table. Model 1A shows a respectable adjusted-$R^2$ at 0.84 which implies that it explains 84% of the variations observed in operational costs. Model 1B on the other hand has an adjusted-$R^2$ of 39%, which is significantly lower than for model 1A. Both models exhibit relatively high F-statistics indicating that both models are valid, but model 1A outperforms model 1B in terms of explanatory power.

The estimated effect of CT on operational costs as measured by the coefficient of the dummy variable $D_{it}$, is an average reduction of 2.25 million NOK/year per link for model 1A. It is important to emphasize that these effects are in levels and hence, are an average over a wide range of total costs, from 3 to 139 million NOK. Therefore, it might be interesting to assess the effect in percentage terms which, is possible by inspecting the coefficient of $D_{it}$ in model 1B. The results show that implementing CT reduced operational costs on average by about 8% (−0.0807 *100) which, is a sizable impact. A next relevant question to ask is how these different values derived by the two models compare. The mean operational cost level in the sample is 27.7 million NOK. Using a back-of-the-envelope calculation one can compare the estimates by cost in levels (model 1A) versus logarithmic transformation (model 1B). Model 1A gave 2.25 million NOK/year per link. It implies that it gives 2.25/27.7 = 8.1% in cost reduction. This is exactly the same result as in model 1B, implying that the two methods give similar results as far at the impact on cost reduction is concerned.

**Table 1**
Summary statistics.

| Variable | Symbol | Unit | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|---|---|
| Operational costs | $y_{it}$ | NOK/year | 424 | 2,77,00,000 | 2,14,00,000 | 34,08,202 | 13,90,00,000 |
| Tender (0 = non CT; 1 = CT) | $D_{it}$ | Is/is not | 424 | 0.13 | 0.34 | 0.00 | 1.00 |
| Number of sailed kilometers | $SKM_{it}$ | Kilometers/year | 424 | 1,01,156 | 81,621 | 12,369 | 5,59,838 |
| Number of passenger vehicles transported (in PCE) | $PCE_{it}$ | PCE/year | 424 | 375898 | 387217 | 23115 | 2343754 |
| Average ferry capacity | $FSIZE_{it}$ | PCE capacity/ferry | 424 | 68 | 37 | 9 | 224 |
| Time trend | t | Year | 424 | 5 | 2 | 1 | 8 |
| Financial Crisis (0 = otherwise, 1 = 2009) | $FCRIS_t$ | Is/is not | 424 | 0.13 | 0.33 | 0.00 | 1.00 |
| Number of Ferries serving link | $NFERRIES_{it}$ | Ferries/link | 424 | 1.37 | 0.62 | 1.00 | 4.00 |
| Type of waters traversed (0 = open sea, 1 = sheltered waters) | $SW_i$ | Is/is not | 424 | 0.09 | 0.29 | 0.00 | 1.00 |
| Average age of ferries at links | $AGE_{it}$ | Years | 424 | 29 | 12 | 5 | 51 |
| Length of link | $L_i$ | Kilometers | 424 | 12 | 14 | 2 | 60 |
| Average number of links operated by each company | | Links | 96 | 6.61 | N/A | 1 | 20 |

**Table 2**
Estimation results.

| Variable | Symbol | Model 1A | Model 1B |
|---|---|---|---|
| | | Cost in levels $y_{it}$ | Cost in logarithm ln ($y_{it}$) |
| CT (1 if CT; 0 otherwise) | $D_{it}$ | −2253858.1* (-2.26) | −0.0807* (-2.18) |
| Sailed km | $SKM_{it}$ | 53.47* 2.05 | 0.00000157* 2.11 |
| No. of vehicles transported (PCE) | $PCE_{it}$ | 18.86** 2.73 | −0.000000161 (-0.89) |
| Average ferry size (PCE-capacity) | $FSIZE_{it}$ | −8714.3 (-0.24) | −0.000822 (-1.07) |
| PCE-kilometers transported | $PCE_{it} \times L_i$ | 0.211 0.55 | 0.0000000173 1.36 |
| Trend (time) | $t$ | 1203873.2*** 7.11 | 0.0575*** 10.13 |
| Financial crisis (1 if 2009; 0 otherwise) | $FCRIS_t$ | −853944.4 (-1.50) | −0.0335* (-2.04) |
| No. of ferries serving link | $NFERRIES_{it}$ | 5971821.6* 2.15 | 0.055 1.3 |
| Interaction between open waters and PCE | $PCE_{it} \times SW_i$ | 51.59** 3.4 | 0.000000298 0.88 |
| Average age of ferries serving link | $AGE_{it}$ | 39382.7 0.64 | 0.000282 0.13 |
| Constant | | −1435006.7 (-0.18) | 16.47*** 120.61 |
| *N* | | 424 | 424 |
| (Overall) adj. $R^2$ | | 0.8446 | 0.3963 |
| F | | 26.33 | 30.34 |

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$.
*z*-values below estimates.

Since both these results are statistically significant at the 5-percent level, they suggest that we have obtained a reasonable estimate of the effect of CT on operational costs. Thus, a clear conclusion is that the implementation of CT led to a significant reduction in operational costs in the Norwegian ferry sector, for those links that were subjected to it. See appendix 1 and 2 for specification and robustness checks of our finding.

Next, the impact of all the other variables included in the models on operational costs must be addressed. The significant variables with regards to model 1A are as follows: (i) operational costs increase with increase in number of sailed kilometers ($SKM_{it}$), (ii) operational costs increase with number of vehicles transported ($PCE_{it}$), (iii) Operational cost increases over time as measured by the time trend ($t$), (iv) operational costs increases by the number of ferries serving the link ($NFERRIES_{it}$) and, (v) transporting large number of vehicles in open waters ($PCE_{it}\ x\ SW_i$) incur higher operational costs as compared transporting vehicles in sheltered waters. All these results are plausible because they all are all cost drivers. All the other variables not mentioned above were found to be insignificant in explaining the variation in observed operational costs.

### 6.2. CT's impact on the market concentration of ferry link services

The previous sub-section established that the implementation of CT reduces operational costs. Another equally related issue to consider is the CT's effect on market concentration. If CT leads to market concentration, it could in the long run lead to monopoly/monopsony. All other things equal, this would be the opposite of the intentions with CT; to enhance competition. We proceed to measure the degree to which CT has led to market concentration according to the two HHI-indices developed in section 4.2. The dataset used to calculate the HHI-indices is as described in section 5 and is comprised of the 53 links. Fig. 1 plots the two HHI -indices (HHI by each company and HHI by ownership group) against the relative proportion of tendered links in the sample; this proportion is the most relevant to consider when inferring the impact of CT on market concentration. It is important to underline that the calculation of the two HHI-indices are based on the links that are also used to estimate the effect of CT on operational cost. This ensures consistency between the CT's impact on operational cost and CT's impact on market concentration. Thus, it should be noted that our results are most valid for the sample of links used in this study. However, these results can to a certain extent be generalized to yield in Norway since the data set comprised about 41% of ferry links in the trunk road system and, 69% of total PCEs transported in the trunk road system.

Consider first the curves for both the HHI - indices. The two curves are strikingly alike in shape and the only clear differences between them is their percentages which, is explained by that one considers individual companies and the other groups of companies. Naturally, the HHI index that accounts for groups (HHI by ownership group) will have a higher percentage index.

From Fig. 1, it is evident that the curves for the HHI's which, are very similar in shape, are relatively stable until 2006, picks up until 2008 and then stabilizes. This indicates that the market concentration increased from 2006 and stabilized in 2008. Next, consider the curve for the proportion of (CT) tendered links. The proportion of tendered links increased from 2004, increased steeply in 2006 and increased even more steeply in 2008, 2009 and 2010. The explanations of these trends are logical and are related to the developments in the Norwegian ferry link procurement system, as explained in section 2. The year when CT became the norm for procuring ferry link transportation services was 2003 hence, the proportion of CT links increased, while operators were


**Fig. 1.** HHI indices versus relative proportions of tendered links.

still determining how to strengthen their market positions. Then came 2006/7 when a larger proportion of links as compared to 2004, were set out on tenders. At this time, some operators had managed to merge or buy out smaller operators to strengthen their market positions; market shares as measured by HHI-indices increased beginning in this year. When an even larger proportion of ferry links, approximately 50%, were set out on CT in 2008–2009, mergers and buyouts to strengthen market positions reached their maximum. The HHI- indices stabilized from then on, as is evident in the figure.

Note, however, that market concentration and tendering in our data set are not correlated perfectly, especially in the years after 2004; see Fig. 1. A possible explanation is that a correlation between tendering and market concentration is dependent on the links that are subjected to tendering. If the market is concentrated from prior to CT reform, an increase in tendering might not necessarily lead to a market concentration if large operators are more likely to win tenders at links that they already operate. One might therefore expect only a mild increase in the market concentration when the smaller operators lose their tenders at the hands of the larger operators. A closer observation of our data set revealed that, before the larger tendering process commenced in 2006, the three largest operators controlled 83% and 70% of the market as measured by the number of links operated; when considering and not considering ownership groupings, respectively. This finding means that a high concentration was already present, notably from 2004. Accounting for the correlations between the three largest operators and the HHI-indices, we found a correlation coefficient of 0.89 and 0.84, when considering ownership groupings and individual companies, respectively. This suggests the possibility that the larger operators are the main drivers of the changes in the market concentration index.

## 7. Concluding remarks

We have used an econometric framework to address the impact of CT on operational costs and market concentration using data from the Norwegian ferry sector. The motivation was that previous studies in the literature have been inconclusive on whether CT truly reduces operational costs of transportation services in both the short term and the long term. In addressing the matter, we have heavily relied on econometrics, i.e., the applications of statistics on economic data, to provide empirical evidence.

The results revealed that CT significantly lowered operational costs. Specifically, and according to the preferred model, implementation of CT reduced the operational costs by 2.25 million NOK per year or 8%, on average. However, we find evidence that CT may also have led to increased market concentration. This latter result is not promising for the working of CT in the long run because market concentration can lead to a monopolistic or duopolistic situation, which is much less efficient than CT. For example, if there exist benefits on the regional level associated with shared maintenance facilities, regional monopolies may arise. Vigren (2020) found the distance from the operators to the contract area to reduce the likelihood of placing a bid in the bus sector. Even though differences between the ferry and bus sector exist, the same mechanism may play a role in the former sector as well. With the understanding that monopoly/duopoly is not as cost efficient for public authorities as CT is, the long-term effect could be an increase in operational costs. However, some moderating comments are as follows:

1. In the current situation, almost all links are operated by Norwegian firms (Oslo Economics, 2019). However, if profits rise due to monopolistic competition, international firms might be attracted to enter the market. This could increase competition.
2. There is also the threat of monopolistic competition (see. e.g. Baumol, 1977, p. 394) in which local monopolies could arise, having a hold on a specific geographical area.

3. If the number of bidders is reduced over time, it could increase the fierceness of competition, which in turn would reduce the tendering bids, maintaining lower costs for the government.[9] According to Oslo Economics (2019), the number of bidders was reduced between 2004 and 2015, from 3,1 to 1,9 per contract.
4. There is also the possibility of a winner's curse, in which companies either miscalculates or bids strategically to enter the market, below the long-term economically viable price level. Consequently, in the long run, costs may increase as firms seek to avoid the winner's curse or scale down their (possible) strategic bidding practices.
5. If the market is contestable, in the sense of Baumol (1986), it is the cost distribution of potential entrants that determines the fierceness of competition; if highly competitive firms threaten to enter the market, the incumbent firms may increase their efforts to stay competitive. Consequently, a highly concentrated market may remain efficient, if there exists a credible threat of entry by more efficient firm[10]. However, a contestable market requires costless entry and exit for each firm (in the extreme). In the tendering process, there are costs associated with entering the market connected to submitting a bid and provide facilities to maintain vessels. Consequently, it remains uncertain to what extent the ferry market is contestable.

We have also discussed the so-called 'bad control' problem when estimating the effects of reforms on outcomes. The effect of CTs might also be transmitted through changes in service levels (i.e., our explanatory variables). However, in this study, we did not estimate such an effect, but it might be an interesting opportunity for further research.

Our results must be compared to earlier findings in the transportation literature. The first finding is that CT works to reduce operational costs, at least in the short term, as has been argued in theory by economists. Wallis, Bray, & Webster (2010) warned, however, that one size does not fit all; in some circumstances, CT might be the best option, whereas in other situations, negotiations with the incumbent might be the best option to reduce costs. Underlying this conclusion is that real competition with many bidders might occur in some situations but not in others. In the latter case, negotiation is likely the best option. A similar issue was addressed by Hensher and Wallis (2005), who warned that, under CT, there is some danger of excessive consolidation of the supplier market among a few large operators, with risks of excessive market power and possible collusion. They suggested that this danger could be minimized by imposing market share or equivalent limits on any one operator in an area; this suggestion has not been followed in the Norwegian ferry sector. However, in the early trial period with CT in the Norwegian ferry sector, Bråthen et al. (2004, p. 408) warned that one should critically assess the market imperfections and that some markets could lead to collusive behaviors, e.g., leading to monopolies. They noted that limited CT in combination with negotiations would be a worthwhile endeavor; note the similarity of their conclusions to those of Hensher and Wallis (2005) published a year later. Our results seem to concur with these previous authors' suggestions; we also conclude that one size does not fit all and that, before implementing CT, one should evaluate whether there will be continued competition in all circumstances; if not, continued negotiation is likely the best option. However, we must caution that our results covers a part of the Norwegian ferry market, and one must be careful not to generalize our results too far without further investigation.

Finally, our assessment concludes that CT might not lead to free competition in the long run as intended but rather to monopolistic/duopolistic tendencies, which are not the goals of implementing CT. Our policy recommendation is therefore clear: reconsider CT because its results could be counter to intentions in the long term.

## Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## CRediT authorship contribution statement

**James Odeck:** Conceptualization, Data curation, Supervision, Writing - review & editing. **Harald Høyem:** Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing.

## Acknowledgments

We would like to thank Joachim Rønnevik, Edvard Sandvik Thonstad and Thor Harald Eliassen at the NPRA for clarifications on the different contract systems and for excellent provision of data. Furthermore, we thank Professor Svein Bråthen at Molde University College for insightful comments on the operation of Norwegian ferry links. We would also like to thank to anonymous reviewers for useful comments. We would like to thank the Norwegian Public Roads Administration (grant 19/145723-2) and Urbanet analyse (Norway) for funding this research.

## Appendix 1. Specification test of non-linearity

To assess whether the specification used is appropriate for studying the cost structure of ferry operation, we the F-test (Kennedy, 2008, p. 52) where two models are compared against one another.

First define the term $M$ which contains a selection of the variables in equation (1) as:

$$M_{it} = \omega D_{it} + \beta_1 SKM_{it} + \beta_2 PCE_{it} + \beta_3 PCE_{it} \times L_i + \beta_4 PCE_{it} \times SW_i \\ \beta_5 FSIZE_{it} + \beta_6 NFERRIES_{it} + \beta_7 AGE_{it} + \beta_8 FCRIS_{it} \tag{A1}$$

Our preferred specification may now be written as:

$$y_{it} = \alpha + \varphi_i + M + \lambda t + e_{it}, \tag{A2}$$

Equation (A2) is the same as equation (1) in the text, corresponding to a linear model of costs. We wish to investigate whether this is an appropriate approximation to the data. We formulate a non-linear specification we may use to test against equation (A2) as follows:

$$y_{it} = \alpha + \varphi_i + \omega D_{it} + \gamma_1 SKM_{it}^2 + \gamma_2 PCE_{it}^2 + \gamma_3 (PCE_{it} \times L_i)^2 + M_{it} + \lambda t + e_{it}, \tag{A3}$$

In equation (A3), we have added squared terms for the number of sailed kilometers ($SKM_{it}$), the number of PCE transported ($PCE_{it}$) and the number of PCE kilometers ($PCE_{it} \times L_i$). These variables are chosen as they are the ones who are directly linked with the production of services at each link: how

[9] We would like to thank an anonymous reviewer for pointing out these three examples.
[10] We would like to thank an anonymous reviewer for pointing out this example to us.

many PCE that are transported, how far they are carried and how many kilometers the ferry travels.

Equation (A2) is the linear model, while equation (A3) is the non-linear one. Other specifications or variable transformations could have been chosen, such as log-transforming the variables. However, we wish to assess if equation (A2) differs significantly from (A3), using a specification in which one model is nested within the other, enables us to use an F-test to achieve this.

Using the F-test we first formulate a null hypothesis that $\gamma_1 = \gamma_2 = \gamma_3 = 0$, that is, all the squared terms are not statistically significant, which renders model (A2) equal to (A3). The F-test derives the likelihood that at least one parameter is different from zero, using the F-statistic, defined as (Kennedy, 2008):

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \tag{A4}$$

Where $J$ (=3) is the number of constraints, $N$ (= 424) is the number of observations and $K$ (= 14) is the number of parameters to estimate. We use the computer package STATA and the command "test" to perform the F-test using model (A2) and (A3). The F-statistic equal to 0.75, which yields a p-value of 0.52. Consequently, we cannot reject the null hypothesis that all $\gamma$ are different from zero.

The F-test does not refute the validity of a non-linear model, but indicates that using our data, we cannot say that a non-linear model is preferable to our linear one.

Moreover, the estimate of the effect of tendering estimating equation (A3) comes out at −2.2 million NOK, as an average per crossing with a p-value of 2%, as shown in table A1. Consequently, we arrive at a similar estimate using the non-linear model, as in the linear one.

## Appendix 2. Robustness checks

To assess the robustness of the results, we provide some additional regressions. The following estimations are made:

- **Model (1A-1B-1AR-1BR) -** *Running the model when the two largest links are omitted.* The purpose is to assess if removing the two largest links alter the results in any way. If larger links dominate the effect of tendering, the estimated effect should diminish when removing them from the sample. One reason why larger links could influence the effect substantially is that any percentage change in savings will constitute a larger effect on costs in levels, as compared to smaller links. However, using the logarithmic version, such an effect is expected to have a lesser influence. We run this test using both the levels and logarithmic version of the model.
  - **Model 1A**: Fixed effects estimation using cost in levels including the two largest crossings. The same as model (1A) in Table 2.
  - **Model 1B**: Fixed effects estimation using cost in logarithms including the two largest crossings. The same as model (1B) in Table 2.
  - **Model 1AR**: Fixed effects estimation using cost in levels excluding the two largest crossings.
  - **Model 1BR**: Fixed effects estimation using cost in logarithms excluding the two largest crossings.
- **Model (2)** - *Running a pooled version of the model, using OLS.* The purpose is to assess the difference between the estimate when not correcting for possible fixed cost effects at link level. All the observations are included in the analysis.

The results are shown in table A2. Starting with the OLS results (model 2), we see that the effect is almost twice as large as the one estimated by Fixed Effects. Consequently, this highlights the practical importance of using the Fixed Effects estimator to attain an unbiased and consistent estimate of the result.

Looking at model 1A-1AR and 1B-1BR we note that all estimates of effects are reasonably similar. The models where costs are given in levels (1A & 1AR), have very similar coefficients of 2.2 and 2.6 million NOK with and without the two largest links, respectively. The effects seem relatively precisely estimated in both cases with low p-values. Moreover, the models where costs are given in logarithms (1B & 1BR), have similar estimates of the effect of tendering on operational costs. However, dropping the two largest crossings from the sample, reduces the precision on the estimate, with a p-value slightly exceeding the 5-percent level. However, as all other models have yielded a consistent result, and the exceedance is rather small, we view our robustness test as confirmatory of our results in the main body of the manuscript.

**Table A1**
Regression table of the linear and non-linear model * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

| Variable | Symbol | Model (A2) | Model (A3) |
|---|---|---|---|
| | | Cost in levels $y_{it}$ | Cost in levels $y_{it}$ |
| CT (1 if CT; 0 otherwise) | $D_{it}$ | −2253858.1*<br>(-2.26) | −2155636.3*<br>(-2.23) |
| Sailed km | $SKM_{it}$ | 53.47*<br>2.05 | 22.66<br>0.63 |
| No. of vehicles transported (PCE) | $PCE_{it}$ | 18.86**<br>2.73 | 16.52<br>1.28 |
| Average ferry size (PCE-capacity) | $FSIZE_{it}$ | −8714.3<br>−0.24 | 4045.6<br>0.13 |
| PCE-kilometers transported | $PCE_{it} \times L_i$ | 0.211<br>0.55 | 1.29<br>1.46 |
| Trend (time) | $t$ | 1203873.2***<br>7.11 | 1170113.2***<br>6.71 |

(continued on next page)

**Table A1** (*continued*)

| Variable | Symbol | Model (A2) | Model (A3) |
|---|---|---|---|
| | | Cost in levels $y_{it}$ | Cost in levels $y_{it}$ |
| Financial crisis (1 if 2009; 0 otherwise) | $FCRIS_t$ | −853944.4<br>−1.5 | −865120.8<br>−1.58 |
| No. of ferries serving link | $NFERRIES_{it}$ | 5971821.6*<br>2.15 | 5159571.9<br>1.95 |
| Interaction between open waters and PCE | $PCE_{it} \times SW_i$ | 51.59**<br>3.4 | 74.36*<br>2.15 |
| Average age of ferries serving link | $AGE_{it}$ | 39382.7<br>0.64 | 47391.9<br>0.77 |
| Number of sailed km. Squared | $SKM_{it}^2$ | N/A<br>N/A | 0.0000803<br>1.09 |
| Number of PCE km. squared | $(PCE_{it} \times L_i)2$ | N/A<br>N/A | −4.57E-08<br>−1.06 |
| Number of PCE transported squared | $PCE_{it}^2$ | N/A<br>N/A | 0.00000134<br>0.36 |
| Constant | | −1435006.7<br>−0.18 | −2102540.4<br>−0.27 |
| $N$ | | 424 | 424 |
| (Overall) adj. $R^2$ | | 0.8446 | 0.8085 |
| F | | 26.33 | 130.21 |

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$.
$z$-values below estimates.

**Table A2**
Regression with robustness checks on the two largest links excluded/included and OLS estimation * $p < 0.05$, $**p < 0.01$, $***p < 0.001$

| Variable | Two largest links included? | Model (1A) | Model (1B) | Model (1AR) | Model (1BR) | Model (2) |
|---|---|---|---|---|---|---|
| | | YES | YES | NO | NO | YES |
| | Estimator | FE | FE | FE | FE | OLS |
| | Symbol | Cost in levels $y_{it}$ | Cost in logarithm ln ($y_{it}$) | Cost in levels $y_{it}$ | Cost in logarithm ln ($y_{it}$) | Cost in levels $y_{it}$ |
| CT (1 if CT; 0 otherwise) | $D_{it}$ | −2253858.1*<br>(0.028) | −0.0807*<br>(0.034) | −2645263.2*<br>(0.02) | −0.0814<br>(0.055) | −4298460.1**<br>(0.002) |
| Sailed km | $SKM_{it}$ | 53.47*<br>0.045 | 0.00000157*<br>0.04 | 54.83*<br>0.049 | 0.00000161*<br>0.039 | 117.5***<br>0 |
| No. of vehicles transported (PCE) | $PCE_{it}$ | 18.86**<br>0.009 | −0.000000161<br>0.379 | 24.35*<br>0.013 | −0.000000172<br>0.497 | 8.854**<br>0.002 |
| Average ferry size (PCE-capacity) | $FSIZE_{it}$ | −8714.3<br>(0.809) | −0.000822<br>(0.288) | −3274.4<br>(0.942) | −0.00114<br>(0.227) | 102211.8***<br>(0) |
| PCE-kilometers transported | $PCE_{it} \times L_i$ | 0.211<br>(0.584) | 1.73E-08<br>(0.181) | 0.142<br>(0.735) | 1.68E-08<br>(0.176) | 0.125<br>(0.648) |
| Trend (time) | $t$ | 1203873.2***<br>(0) | 0.0575***<br>(0) | 1152510.5***<br>(0) | 0.0572***<br>(0) | 1487970.8***<br>(0) |
| Financial crisis (1 if 2009; 0 otherwise) | $FCRIS_t$ | −853944.4<br>(0.14) | −0.0335*<br>(0.046) | −803811.4<br>(0.154) | −0.0360*<br>(0.036) | −1409330.3*<br>(0.011) |
| No. of ferries serving link | $NFERRIES_{it}$ | 5971821.6*<br>(0.036) | 0.055<br>(0.2) | 5756254.9<br>(0.088) | 0.0429<br>(0.385) | 8046642.0***<br>(0) |
| Interaction between open waters and PCE | $PCE_{it} \times SW_i$ | 51.59**<br>(0.001) | 0.000000298<br>(0.383) | 56.39<br>(0.329) | 0.000000955<br>(0.401) | 4.033<br>(0.55) |
| Average age of ferries serving link | $AGE_{it}$ | 39382.7<br>(0.528) | 0.000282<br>(0.898) | 17416.2<br>(0.795) | −0.000134<br>(0.956) | −28263.3<br>(0.555) |

(*continued on next page*)

**Table A2** (*continued*)

| Variable | Two largest links included? | Model (1A) | Model (1B) | Model (1AR) | Model (1BR) | Model (2) |
|---|---|---|---|---|---|---|
| | | YES | YES | NO | NO | YES |
| | Estimator | FE | FE | FE | FE | OLS |
| | Symbol | Cost in levels $y_{it}$ | Cost in logarithm ln ($y_{it}$) | Cost in levels $y_{it}$ | Cost in logarithm ln ($y_{it}$) | Cost in levels $y_{it}$ |
| Constant | | −1435006.7<br>(0.856) | 16.47***<br>(0) | −1872087.6<br>(0.837) | 16.49***<br>(0) | −11249524.9***<br>(0) |
| *N* | | 424 | 424 | 408 | 408 | 424 |
| (Overall) adj. $R^2$ | | 0.8446 | 0.3963 | 0.8057 | 0.2531 | 0.9372 |
| F | | 26.33 | 30.34 | 18.73 | 31.27 | 790.4 |

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
*p*-values are in parenthesis.

## References


Aarhaug, J., & Fearnley, N. (2016). Deregulation of the Norwegian long distance express coach market. *Transport Policy, 46*, 1–6.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, New Jersey: Princeton university press.

Baird, A. J., & Wilmsmeier, G. (2011). Public tendering of ferry services in Europe. *European Transport/Trasporti Europei*, (49), 90–111.

Baltagi, B. (2005). *Econometric analysis of panel data* (3rd ed.). Chichester, England: John Wiley & Sons Ltd.

Bardal, K. G., & Mathisen, T. A. (2015). Winter problems on mountain passes–Implications for cost-benefit analysis. *Transportation Research Part A: Policy and Practice, 74*, 59–72.

Baumol, W. J. (1977). *Economic theory and operations analysis* (4th ed.). New Jersey: Prentice-Hall.

Baumol, W. J. (1986). Contestable markets: An uprising in the theory of industry structure. *Microtheory: Applications and Origins*, 40–54.

Bråthen, S., Fuglseth, A., Hervik, A., Larsen, O. I., Lyche, L., Odeck, J., & Sandvik, E. T. (2004). *Anbud i ferjesektoren-erfaringer og utsikter. (Tendering in the ferry industry–Experiences and expectations). Molde.* Norway: Møre Research.

Bråthen, S., Hervik, A., Odeck, J., & Sunde, O. (2004). Competitive tendering versus subsidy schemes in public transport services: Comparing efficiency in the provision of ferry services. In *Proceedings of the European transport conference 2004*.

Bray, D., & Wallis, I. (2008). Adelaide bus service reform: impacts, achievements and lessons. *Research in Transportation Economics, 22*(1), 126–136.

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods, 3*, 399–408, 02.

Economic and Policy Services Pty Ltd, & Ian Wallis Associates Ltd. (2007). *Review of the performance of bus reform in Adelaide*. Adelaide: Department for Transport, Energy and Infrastructure, Public Transport Division (September).

Førsund, F. R. (1992). A comparison of parametric and non-parametric efficiency measures: The case of Norwegian ferries. In *International applications of productivity and efficiency analysis* (pp. 21–39). Springer Netherlands.

Hart, O, Schleifer, A, & Vishny, RW (1997). The proper scope of government: Theory and an application to prisons. *The Quarterly Journal of Economics, 112*(4), 1127–1162.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251–1271.

Hensher, D. A., & Stanley, J. (2008). Transacting under a performance-based contract: The role of negotiation and competitive tendering. *Transportation Research Part A: Policy and Practice, 42*(9), 1143e1151.

Hensher, D. A., & Stanley, J. (2010). Contracting regimes for bus services: What have we learnt after 20 years? *Research in Transportation Economics, 29*(1), 140–144.

Hensher, D. A., & Wallis, I. P. (2005). Competitive tendering as a contracting mechanism for subsidising transport: The bus experience. *Journal of Transport Economics and Policy, 39*(3), 295–322.

Hervik, A., & Sunde, Ø. (2001). An evaluation of competitive tendering in the ferry sector in Norway. In *International conference series on competition and ownership in land passenger transport – 2001 – Molde, Norway– Thredbo 7*.

Houghton, E., & Hensher, D. A. (2005). Negotiated and competitively tendered performance based contracts. In K. J. Button, & D. A. Hensher (Eds.), *Transport strategy, policy and institutions, handbooks in transport* (Vol. 6, p. 527e546). Oxford: Elsevier Science.

Jørgensen, F., & Mathisen, T. A. (2010). Using standardized revenue and cost norm analyses to reveal subsidy fraud in contracted public transport services. *Transport Reviews, 30*(3), 299–313.

Jørgensen, F., Mathisen, T. A., & Larsen, B. (2011). Evaluating transport user benefits and social surplus in a transport market – the case of the Norwegian ferries. *Transport Policy, 18*(1), 76–84.

Jørgensen, F., & Solvoll, G. (2018). Determining optimal frequency at ferry crossings. *Transport Policy, 63*, 200–208.

Kain, P. (2006). The pitfalls in competitive tendering: Addressing the risks revealed by experience in Australia and Britain. In *Papers of the competitive tendering of rail passenger services: Experience to date EMCT workshop. Paris: ECMT. 12 January*.

Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Malden MA: Blackwell Publishing.

Krugman, P. (2003). *Thanks for the M.R.E.s. New York times, 12 August, section A*.

Mathisen, T. A. (2016). Competitive tendering and cross-shareholding in public passenger transport. *Transport Policy, 48*, 45–48.

Mathisen, T. A., & Jørgensen, F. (2012). Panel data analysis of operating costs in the Norwegian car ferry industry. *Maritime Economics & Logistics, 14*(2), 249–263.

Myers, J., & Ashmore, D. (2007). When to tender, when to negotiate? Why are we ignoring the elephant in the room?. In *Papers of the 10th International Conference on Competition and Ownership in Land Passenger Transport. Hamilton Island 12e17 August*.

Nash, C., & Wolański, M. (2010). Workshop report–Benchmarking the outcome of competitive tendering. *Research in Transportation Economics, 29*(1), 6–10.

Norheim, B. (1999). Competitive pressure as an alternative to competitive tendering? The development of a performance contract in Oslo. In *Papers of the 6th international conference on competition and ownership in land passenger transport. Cape town, South Africa. September*.

Odeck, J. (2008). Efficiency measurement of ferries serving road networks in Norway: A stochastic frontier approach. *Maritime Economics & Logistics, 10*(4), 409–428.

Odeck, J., & Bråthen, S. (2009). The efficiency of Norwegian ferries in providing public transport services. *International Journal of Transport Economics/Rivista Internazionale di Economia dei Trasporti*, 121–139.

Oslo Economics. (2016). *Premissanalyser – tiltaksanalyse på for utvikling av fergemarkedet på lang sikt (English: Analyses of means to develop the ferry market on a long term basis)*. Oslo Economics-report 2016-23.

Oslo Economics. (2019). *Analyse av markedet for innenriks sjøfart (English: Analysis of the market for domestic ferry services)*. Oslo Economics-memo, unserialized. 22.03.2019.

Pollitt, C., & Bouckaert, G. (2004). *Public management reform: A comparative analysis.* Oxford university press.

Rhoades, S. A. (1993). Herfindahl-hirschman index. *Federal Reserve Bulletin, 79*, 188.

Savas, E. S. (2000). *Privatization and public-private partnerships*. New Jersey: Chatham House Press.

Stanley, J., & van de Velde, D. (2008). Risk and reward in public transport contracting. *Research in Transportation Economics, 22*(1), 20–25.

Statistics Norway. (2020). https://www.ssb.no/en/sok?hovedemner=National+account+and+business+cycles&sok=Adjusting+for+inflation.

Vigren, A. (2020). The Distance Factor in Swedish Bus Contracts How far are operators willing to go? *Transportation Research Part A: Policy and Practice, 132*, 188–204.

Wallis, I., Bray, D., & Webster, H. (2010). To competitively tender or to negotiate–Weighing up the choices in a mature market. *Research in Transportation Economics, 29*(1), 89–98.

Walters, J., & Cloete, D. (2008). The South African experience with negotiated versus competitively tendered bus contracts. *Transportation Research Part A: Policy and Practice, 42*(9).

# Paper II

Research in Transportation Economics 82 (2020) 100878

Contents lists available at ScienceDirect

Research in Transportation Economics

journal homepage: http://www.elsevier.com/locate/retrec

# Optimal public transit frequency under stochastic demand and fixed vehicle size: Application in the Norwegian car ferry sector

Harald Høyem [a,b,*], James Odeck [a,c]

[a] Department of Civil and Environmental Engineering, NTNU - Norwegian University of Science and Technology, 7491, Trondheim, Norway
[b] Urbanet Analyse, Po Box 337, 0101, Oslo, Norway
[c] Department of Economics and Logistics, Molde University College, P.O. Box 6405, Molde, Norway

ARTICLE INFO

*JEL classification:*
R40
R41
R42
R48

*Keywords:*
Transport sector
Car ferry services
Optimization
Optimal frequency

ABSTRACT

When capacity constraints are present in public transportation services, some travelers may not be able to board the first vessel they desire. Consequently, a type of waiting time cost is introduced, as passengers must wait for (at least) another full headway before the next service arrives. This waiting time has, to a lesser extent, been accounted for in the literature focusing on the optimal frequency of departures in transport. Not including this waiting time cost may lead to an underestimation of the optimal departure frequency. In this paper, we develop a novel optimization framework to incorporate the effect of this waiting time when there is a limited vessel capacity. Two different formulations are applied, which yield consistent results. Our policy recommendation is clear: Both the theoretical and empirical findings suggests that a "third" waiting time component of not being able to board the first vessel is highly relevant for decision-makers when vessel size constraints exists. Not including the extra waiting time costs stemming from capacity constraints leads to starkly different optimal service levels, which clearly indicates the relevance of our findings for practical policy making.

## 1. Introduction

Decision-makers that control, provide and subsidize the level of services offered by public transport agencies require tools for estimating the optimal level of services to be offered by the transport agents before selecting the agency that offers the best solution. One particular tool that is needed by decision-makers is one that is used to determine the best/optimal frequency. Such a tool must be able to account for all the factors that impact the costs and benefits for operators and users such that the option with the greatest positive difference between net benefits and net costs is chosen.

Waiting time is an important cost element for transit users that must be addressed adequately by optimizing the departure frequencies in any given public transport circumstance. Time is readily evaluated in monetary terms, and when public transport system users are not able to depart at their preferred time due to an inappropriate departure frequency and/or capacity limitations, they incur losses that must be accounted for.

There is, however, an additional type of waiting time that has not been appropriately considered in the literature when designing the optimal departure frequency for public transport, at least in car ferry transport. When there are capacity constraints on a public transport link, some travelers may not be able to board the first vessel they desire. Consequently, those travelers who are not able to board must wait until the next vessel arrives, which is a cost that must be accounted for when designing the optimal frequency. Next, even if this waiting time cost was accounted for, the frequency would need to be increased to reduce it and increasing the frequency may also increase the capacity. Thus, accounting for this type of waiting time requires a framework that can help decision-makers to properly account for the waiting time cost and capacity increases, such that the optimal frequency provision can be reached. In the case of Norwegian ferries, at approximately 58% of the links,[1] more than 1% of the vehicles waiting to board the ferries are not able to board the first arriving vessel, as a yearly average. Of these links,

* Corresponding author. Department of Civil and Environmental Engineering, NTNU - Norwegian University of Science and Technology, 7491, Trondheim, Norway.

*E-mail address:* Hho@urbanet.no (H. Høyem).

[1] Weighted by the number of vehicles transported at each link.

https://doi.org/10.1016/j.retrec.2020.100878
Received 31 October 2019; Received in revised form 21 January 2020; Accepted 22 January 2020
0739-8859/

the proportion of vehicles that have to wait until the next vessel arrives is 3%, as a yearly average,[2] indicating that some users indeed experience extraordinary waiting times.

The purpose of this paper is to develop a framework to determine a socially optimal frequency at a ferry crossing when we include the users' time costs when they are unable to board their desired departure. Consequently, we develop a frequency optimization procedure that incorporates the abovementioned waiting time while accounting for the increased sailing cost. To the best of our knowledge, the method that we present is new to the literature, at least in the car ferry sector, although it builds on the framework of Mohring (1972) and Jaria-Diaz & Gschwender (2003). As will be shown later on, Mohring (1972) is a special case in our framework, in which infinite capacity is assumed.

Furthermore, we develop and compare two different optimization problems that depend on the assumption underlying the arrival time, i.e., (i) users arrive randomly and (ii) users plan their arrivals at the quay. Development of these problems is also a contribution to the literature.

The rest of the paper is organized as follows: Section 2 provides a literature review. Section 3 briefly outlines Mohring (1972)'s classic model. Section 4 presents our model formulations. Section 5 uses numerical examples to demonstrate the usefulness of the models developed, and some concluding remarks are provided in Section 6.

## 2. Literature review

The question of how to optimize different policy decision variables, such as frequency, in public transport has been investigated in the literature of transport economics for decades. Mohring (1972) provided the first formulation and study of optimal frequency in the context of a fixed-service public transit line. His model was later refined by several authors, e.g., Jansson (1980), Furth (1981), Jaria-Diaz & Gschwender (2003; 2009; 2017) and Jørgensen and Solvoll (2018), to incorporate additional dimensions, such as crowding, financial constraints, multiple periods, multiple lines and the application to car ferries.

Researchers in the context of car ferries have assessed the importance of waiting time costs when designing the optimal frequency of ferry services. In their research, they divided waiting time costs into two different categories, i.e., the so-called hidden waiting time costs that arise when users are not able to depart at their preferred time but must instead follow a schedule and open waiting time, which is related to the minutes spent at a specific terminal before the ferry arrives; see, e.g., Sandberg-Hanssen, Jørgensen, and Larsen (2019) or Andersen and Tørset (2018).

A common implicit assumption in the optimal frequency literature is that there are no extra costs associated with users not being able to board the first vessel that arrives at the quay.[3] As an example, consider that some public transport users will plan their travels and arrive on time at the quay and, hence, will board on time, incurring zero (open) waiting time. However, some users will arrive randomly, hoping to board the first departure, but will fail to do so due to the prevailing capacity constraint, as the users arriving before them have already boarded the vessel, which is at its full capacity. The last arriving users will, thus, incur a waiting cost that must be accounted for when designing the optimal frequency for the transport link concerned. By setting the capacity to be only sufficient to cover some level (e.g., mean value), extra waiting time costs will arise if the realized demand exceeds the capacity. Some users will not be able to board the vessel and will need to wait for an extra cycle before they are able to board, which implies that an additional, nonzero waiting time cost is present when less-than-sufficient capacity is involved,[4] i.e., a "third" waiting time component. Thus, the third waiting time component can be defined as the *open waiting time exceeding the headway when capacity restrictions apply*.[5]

The above problem was noted by Oldfield & Bly (1987), who argued that bus sizes should be increased to limit the extra cost incurred by users when the capacity proved to be less than sufficient. However, the frequency was determined by the bus size and capacity constraints in their model, providing only an explicit expression for bus size. From the perspective of a planner who wishes to provide sufficient capacity, one may increase the sizes of the vessels, the number of departures, or their combination. In particular, there are cases in which changing the vessel size is difficult, e.g., when equipment has been purchased that is meant to last a long time, and the demand increases more than expected. In such cases, adjusting the frequency might be the only option available, which adds to the importance of investigating the frequency as an endogenous optimization variable.

Further, Oldfield & Bly (1987) relied on parametric forms when estimating the probability that demand exceeds capacity. Thus, one must estimate such parameters during each application of the model; moreover, it is unclear how this estimation relates to the other formulations the literature (i.e., the tradition begun by Mohring (1972)). It would then be preferable, from our perspective, to have a unified framework, where all costs could be addressed in a consistent framework, where one model could be derived from another, with the appropriate assumptions being made.

Jaria-Diaz & Gschwender (2003) formulated a model containing the third-waiting time component in the objective function. However, they did not provide any analytical or numerical solutions for *how* the optimal frequency, including the additional costs, may be found. Their objective function was intended to point in the *direction* in which a general microeconomic model of the optimal transit frequency should be developed, while our paper attempts to determine how the optimal frequency can be found.

A limitation of their model was the absence of an *explicit* expression for differential weighting of the waiting time for users who (a) are able to board immediately (cost of half the headway) and (b) must wait for the next service (cost of half the headway plus a full headway), which is amended in our paper. As the cost to users who are left behind is significantly higher than that to those who are not, it is vital to treat this cost properly and completely, which is a central part of our contribution to the literature.

Pedersen (2003) also studied the case with capacity constraints in conjunction with an analysis of optimal price setting. The analysis assumed demand to be deterministic, and the different pricing strategies were examined under different assumptions, depending on whether the constraints were binding or the capital was utilized efficiently.

Herbon and Hadas (2015) and Hadas and Shnaiderman (2012) both developed frameworks where optimization was applied to determine both the bus size and frequency while treating demand as stochastic. Both studies used an objective function that seeks to minimize the under- and overprovision of capacity. From the perspective of a transit operator, providing too much capacity is a waste of resources, while providing too little increases the user cost, which is important to the transit authority. However, none of the models have approached the problem from an economic perspective, that is, minimizing the sum of the user cost (as defined by half the headway) and operator cost simultaneously, focusing on capacity surplus and shortage rather than on waiting times for user costs. Further, Herbon & Yadas rely on a

[2] Based on 2018 data gathered from The Norwegian Public Roads Administration's database "Ferjedatabanken" on 06.12.2019. It is measured as "vehicles" divided by the "number of cars that have to sit back".

[3] Many studies use bus transportation as an example, where "vehicles" and "stops" are considered instead.

[4] A recent empirical work by Hanssen et al. (2019) suggests that, in the case of Norwegian car ferries, 20% of users' waiting times are explained by not being able to board the first ferry that arrives, demonstrating the practical importance of relaxing the assumptions mentioned above.

[5] We would like to thank an anonymous reviewer for suggesting this definition.

parametric form to calculate the number of users that are unable to board the first arriving bus, such that the parameters have to be estimated using some statistical method.

As we aim to include the waiting time cost of not being able to board the first vessel arriving at a quay, a pertinent issue is the question of how to estimate the users' waiting cost. Fosgerau (2009) survey the literature on this subject and develop a framework whereby users either arrive randomly or actively plan their arrival time at the quay. We use Fosgerau and Karlström (2010)'s approach in conjunction with Fosgerau (2009) as a point of departure to evaluate the waiting time.

## 3. Optimal frequency without the third waiting time component

To provide an outline of the standard method for estimating the optimal frequency in the transport economics literature, we use Mohring's seminal model (Mohring, 1972) as a point of departure. His model of optimal frequency assumes inelastic demand and minimizes the Value of Resources Consumed (*VRC*) given by the following equation:

$$VRC = \omega \frac{Y}{2\lambda} + L \times q_L \times \lambda, \tag{1}$$

where $\omega$ is the value of the waiting time (NOK/hour), $Y$ is the level of demand (demand/hour[6]), $\lambda$ is the frequency (departures/hour), $L$ is the round-trip length (twice the length) and $q_L$ is the operators' cost per kilometer sailed (NOK/km). The first part of the equation is the user cost, with the standard assumption of the average waiting time being half the headway. The second part is the operators' cost. The equation is minimized with respect to $\lambda$ (frequency) such that one obtains some algebraic expression for the optimal $\lambda$.

The optimal frequency is given by the following equation:

$$\lambda = \sqrt{Y \frac{\omega}{2Lq_L}} \tag{2}$$

This is known as the "square root formula." The model implicitly assumes that the total level of demand is fixed at $Y$ and does not include any costs related to the possibility that some users may not be able to board, as the vessel capacity is absent from the mathematical expression.

## 4. Optimal frequency with the third waiting time component included

Next, we present our optimization models that include the third waiting time component. We begin with the formulation in which users arrive randomly at the quay and then proceed with the formulation where users plan their arrival time.

The majority of models, starting with Mohring's (1972) seminal work, make the following two central assumptions, which we will carry forward in our formulation:

- First, demand is exogenous, meaning that altering the service level does not change the level of demand. This is arguably a strong assumption, and one possible extension of our work is to relax it.
- Second, pricing is omitted from the optimization, meaning that the fiscal impacts are measured solely based on the operational costs associated with an increasing frequency, which is in line with previous studies; see, e.g., Jaria-Diaz & Gschwender (2003) and Jørgensen and Solvoll (2018).

The second assumption provides another example of a possible avenue for further investigation, i.e., modifying the model. The consequence of carrying forward these assumptions may be that the estimated service levels differ from the ones resulting from a "complete" model. The direction would most likely depend on a user's sensitivity to fares and service levels. Thus, our model applies first and foremost to the case in which the overall level of demand may be relatively insensitive to service level changes and should be regarded as a first step in formulating a framework that fully incorporates the third waiting time component.

Moreover, it is relevant to discuss the differences in the long- and short-run marginal costs and their importance for the optimal frequency. The short-run marginal costs in the ferry sector can be interpreted as the cost of increasing the frequency by one unit when there are enough vessels available. Long-run marginal costs are present when the capacity is increased by acquiring more and/or larger vessels.

According to standard economic theory, the capacity should be extended if the marginal willingness to pay exceeds the long-term marginal costs. We do not include any capital costs of acquiring new vessels in our models. We effectively assume that more capacity can be provided by increasing the frequency with the given number of vessels available, such that long-term marginal costs are not relevant. Including this factor into the model would be a possible extension of it.

### 4.1. Formulation 1: random arrival

A main feature of this model is its ability to differentiate between the following two groups:

- **Group A:** Users who are able to board the first vessel arriving at the quay.
- **Group B:** Users who are unable to board the first vessel arriving due to capacity limitations.

We will denote these groups as A and B, respectively, throughout the paper. The important assumptions within this framework are as follows:

- Demand is normally distributed.
- The model does not assume any queuing at the quay. Consequently, arriving early (prior to a specific departure) does not mean one has a higher probability of boarding the vessel prior to other users. Usually, car lanes are used at a quay, such that queues are indeed formed. Formulation 2 investigates the effect on the optimal frequency when queuing at the quay is introduced.
- Our model examines the demand and frequency per hour and creates an *average and equal probability* of not being able to board at each specific departure within an hour. However, it might be that only those arriving prior to the first departure will have to wait, while those arriving just after the first departure are able to board directly.[7]

Let $f(x;\mu,\sigma)$ be the probability density function (pdf) of demand (Passenger Car Equivalents (PCE)[8]/hour), with mean $\mu$ and standard deviation $\sigma$. Further, let $k$ denote the capacity of each vessel measured in PCE(which is taken as fixed) and let $\lambda$ denote the departure frequency (departures per hour). The expected cost of group A (the one able to board immediately after service arrival) is equal to the following:

$$C_A(\lambda) = \frac{\omega}{2\lambda} \int_0^{k\lambda} xf(x)dx \tag{3}$$

where $\omega$ denotes the value of time associated with waiting for the service arrival, measured in NOK/hour. The expected number of users in group A is $E(A) = \int_0^{k\lambda} xf(x)dx$, that is, the part of demand less than the

[6] Measured in PCE.

[7] We would like to thank an anonymous reviewer for suggesting this example.

[8] Passenger Car Equivalent is a normalized measure of vehicle size. According to Solvoll & Jørgensen (2018), one passenger car (less than 6 m) is equivalent to 1.025 PCE, while a heavy truck (longer than 19 m) amounts to 10.682 PCE.

capacity $k\lambda$.

The division by 2 reflects the standard assumption of the average waiting time being half the headway; see, e.g., Osuna and Newell (1972). Demand is integrated from 0 to $k\lambda$, which is the current capacity of the service, within one time period.

Hanssen et al. (2019) and Andersen and Tørset (2018) found that car ferry passengers in Norway tend to adjust to the scheduled departure times as the headway increases, suggesting lower waiting times at the terminal than implied by half of the headway. However, the crucial relationship is the difference between the desired start time of a trip and the departure time of the service. Assuming a uniform distribution of the desired departure times, ½ of the headway remains the average cost. Moreover, if users wait a larger proportion of their time at home before departing, their valuation of this time may differ from the valuation of time spent waiting for the service at the terminal. Our analysis thus assumes that users value the time spent at the terminal and the time spent at home equally.

To calculate the costs for group B (the ones that are "left behind" at the first service arrival), we will use a concept from the supply chain management literature called the "loss function" (see, e.g., Nahmias, 2005; Thoneman, 2005). The loss function is defined as follows (Thoneman, 2005):

$$J(x) = \int_{y=x}^{\infty} (y-x)f(y)dy = G\left(\frac{x-\mu}{\sigma}\right)\sigma = G(z)\sigma \tag{4}$$

where G(z) is the so-called standardized loss function (detailed in appendix 1), $\mu$ (PCE/hour) is the expected value and $\sigma$ is the standard error of demand. Using this definition, we now write the costs for group B as the integral of demand over the demand distribution that *exceeds* the capacity (expected number of users larger than $k\lambda$) minus the capacity itself, which equals the expected number of users above the capacity[9]:

$$C_B(\lambda) = \frac{3\omega}{2\lambda}\sigma \int_{k\lambda}^{\infty} (x-k\lambda)f(x)dx \tag{5}$$

This yields the expected number of users that exceed the capacity $k\lambda$, by integrating from $k\lambda$ to $\infty$.

Group B will have costs that are 1 higher than group A. Users that are not able to board will first wait half the headway when arriving at the quay and another full headway (1) for the next service, giving a total waiting time of $1+\frac{1}{2}=\frac{3}{2}$ in terms of the headway. An important assumption here is that the users can board on their "second attempt," on average.

We now have the costs for the two different groups weighted by their relative importance, which is derived from the following two parts:

(1) The share of total demand (found by integration), and
(2) The relative disutility of the waiting time (found by the weights, 1/2 and 3/2, in the expressions).

We are able to write the resulting objective function in a compact manner and define the total cost (user and operating) as the function $\Omega(\lambda)$ (The derivation of the objective function is given in appendix 1)[10]:

$$\min_{\lambda}\Omega(\lambda) = \frac{\omega}{2\lambda}[\mu + 2J(\lambda)] + \lambda q_L L \tag{6}$$

The operating cost is added to the equation, where $q_L$ is the cost per kilometer and $L$ is the round trip length (two times the length in kilometers). An increase in the frequency may require more vessels, thus yielding higher capital costs, which is not a part of this cost formulation.

The function is minimized with respect to the frequency, $\lambda$. The model is a nonlinear convex function of $\lambda$, as the terms $\frac{\omega}{2\lambda}$ and $J(\lambda)$ are both convex functions, and $\lambda q_L L$ is only a linear function. The product of two convex functions is also convex, so $\Omega(\lambda)$ is globally convex. A description of the numerical solution method is given in appendix 3.

The model's formulation now allows us to directly estimate the 'optimal' share of the demand "left behind," which is the expected number of users that must wait an extra turn ($J(\lambda)$) divided by the expected users in total ($\mu$), as follows:

$$\gamma^*(\lambda^*, \mu) = \frac{J(\lambda)}{\mu} \tag{7}$$

This is the optimal share of the demand left behind ($\gamma^*$), which is a function of the optimal frequency ($\lambda^*$) and expected demand ($\mu$).

We will now examine some special features of this model; i.e., how it might be changed to incorporate the delay cost, how Mohring's (1972) model is a special case of it, and the consequences of setting the capacity at the mean demand.

Thus far, the model has only included a uniform weighing of the different time components that users are facing. However, the empirical research suggests that the waiting time and delays are treated differently by travelers, where delays are considered more costly to users compared to waiting time (see, e.g., Samstad et al. (2010)). One can modify the model to incorporate this facet by replacing the cost term $3\omega/2\lambda$ with a weighted term, as follows:

$$C_B(\lambda) = \frac{\omega}{\lambda}\left(v_D + \frac{v_H}{2}\right)\sigma \int_{k\lambda}^{\infty} (x-k\lambda)f(x)dx \tag{8}$$

where, $v_H$ is the weight associated with the waiting time, that is, the number of minutes of in-vessel time each waiting time minute corresponds to. Similarly, $v_D$ is the weight of delays, that is, the number of in-vessel minutes each delayed minute corresponds to. The underlying assumption in this formulation is that waiting for the next vessel after being "left behind" corresponds to a delay that users have not planned for. One could argue that if the service continually has too low a capacity, users might take this into account when deciding upon their departure time. However, we argue that the cost of departing prior to a preferred time might also be interpreted as a cost, along the lines of Fosgerau and Karlström (2010) below. Using the extra delay cost will raise the optimal capacity compared to the level without extra delay costs. We have not used the assumption of delay weighting in our empirical application, as this section is given as an example of the model's flexibility for extensions.

An interesting and novel feature of the model is its limiting behavior when frequency, and hence capacity, approaches infinity, that is $k\to\infty$. We will now demonstrate that the present model includes the "standard" model as a special case in its limit. The result is derived in appendix 4 and reads as follows:

$$\lim_{k\to\infty} C_A + \lim_{k\to\infty} C_B = \frac{\omega}{2\lambda}\mu \tag{9}$$

This result is the same as in the standard model of Mohring (1972). That is, the cost is equal to half of the headway, multiplied by the value of time ($\omega$) and demand ($Y=\mu$). Consequently, our model contains the standard one as a special case, showing that the two models are consistent with each other and yield the same results when the same assumptions apply. Moreover, it also shows how the standard approach may sometimes yield results that differ from our framework, particularly when less than sufficient capacity is provided.

We shall now briefly touch upon the consequences of setting the capacity equal to the expected (mean) demand. Using our above framework but neglecting the cost of users not being able to board, $C_B(\lambda)$, we obtain user costs equal to (the derivation of cost is given in appendix 2) $C = \frac{\omega}{2\lambda}$, which is the "standard" half-headway rule.

[9] As mentioned by one of the reviewers, the integration should be performed to $2k\lambda$ instead of $\infty$. However, when testing the effect of changing this assumption, we found the results to be identical. As it is difficult to change the integration limits, we did not change the expressions to avoid unnecessary complications that do not affect the results.

[10] The term $J(\lambda)$ is defined as.$G(z)\sigma$.

However, when applying our model (that is, including $C_B(\lambda)$), we find that the costs actually accrue to $C = \frac{\omega}{\lambda}$, which is twice the value, and hence, the cost is $\frac{\omega}{2\lambda}$. Thus, setting the capacity at the mean demand and not taking into account the cost of users "left behind" yields only half of the actual cost under our assumptions. Assuming demand to be normally distributed is what drives the result given above. If one assumed the demand to follow a nonsymmetrical distribution, the result would most likely be different.

### 4.2. *Formulation 2: planned arrival*

Our second formulation uses a reliability-oriented framework inspired by Fosgerau and Karlström (2010). In this framework, users are not arriving randomly at the quays but plan ahead, selecting an optimal arrival time, which is a compromise between the disutility of having to abandon an activity earlier than preferred and the disutility of not being able to board the first arrival at the station. Arriving earlier at the quay will reduce the likelihood of not being able to board the first arriving vessel. Underlying this mechanism is the assumption that users are boarding the vessel in a First-In-First-Out (FIFO) queue, thus it is desirable to arrive early. We omit any strategic behavior on the users' part in this formulation, meaning that the users do not take into account that other users might also be thinking strategically, which is a possible extension of the model.

Formulation 1 assumed that the users arrive randomly at the quay. However, some users may plan to select an "optimal" time to arrive at the quay. Formulation 2 is introduced to assess whether changing the assumptions regarding the user's decision-making process influences the optimal frequency level. As some ferry links have rather low frequency, it might be pertinent to include an explicit model in the analysis. With low-frequency services, arriving randomly (without consideration of the timetable/planning) may yield a higher "open" waiting time compared to that of planning ahead. If users weight the open waiting time differently to the hidden one, the total user cost will be affected.

Further, Mohring's framework does not assume any queueing at the quay. However, as most ferry links in Norway employ car lanes at the quays, some queuing does indeed occur. Consequently, users arriving early will have a higher open waiting time but a lower probability of experiencing an excess open waiting time (the "third" waiting time component); the opposite is true for users arriving late.

For example, a user arriving "first"; i.e., a full headway prior to departure, will board the vessel with a waiting time greater than half the headway, but the user is certain to board the vessel. Users arriving just before departure will have an open waiting time less than half the headway but may have a (expected) higher excess open waiting time (the "third" waiting time component), as the vessel might be full already.[11] Formulation 2 attempts to make this distinction explicit. As noted in chapter 4.1, Mohring's approach does not assume any queuing. Thus, in the empirical applications, we can use formulation 2 to assess how big an impact the no queuing assumption in Mohring's model (and formulation 1) exerts on the results by comparing formulation 1 and 2.

In the model, users balance the following two different costs in this formulation:

(1) The cost of having to leave their current activity earlier than preferred (i.e., open waiting time at the quay) and
(2) The probability and cost of not being able to board the first vessel arriving.

The earlier the users leave their current activities and travel to the quay, the lower the probability of their not being able to board becomes, assuming a FIFO queue. However, users view this time as an inconvenience. In the optimum, the costs are balanced against each other to achieve a total cost minimum. As in formulation 1, we assume that the probability of not being able to board is the same for all departures within a time period (i.e., each hour), and normally distributed demand.

Start by defining the expected number of vehicles arriving per hour as $Y$ (PCE/hour). With a frequency of $\lambda$ (departures/hour), the expected number of vehicles arriving in the interval between two consecutive arrivals is equal to $Y/\lambda$. A vehicle which arrives $t$ hours prior to the next departure will, on average, have $\mu(t) = (1/\lambda - t)Y$ vehicles arriving before it. A lower $t$ (closer to the departure time of the vessels) will yield a higher-than-expected number of vehicles that have arrived at the quay, reaching a maximum of $Y/\lambda$ vehicles, when arriving at the quay just before departure. Consequently, by observing the FIFO-principle, the lower the $t$, the more likely that the vessel is at capacity, as more vehicles have arrived. At the same time, a higher $t$ yields a higher cost of early departure from the current activity the users occupying the vehicles are undertaking (open waiting time). In optimum, these costs must balance.

Now, define the probability of not being able to board as a function of the number of hours (in decimal) before vessel departure, $P(t)$, as follows:

$$P(t) = \int_k^\infty f(x; \mu(t))dx \tag{10}$$

where, $\mu(t)$ is defined as in the text above.

Further, define the reliability cost $H(t, \lambda)$ as the sum of the cost of the open waiting time ($t$) and the cost of not being able to board the first vessel that arrives at the quay. The cost is given by the following equation:

$$H(t, \lambda) = t + \frac{1}{\lambda}\int_k^\infty f(x; \mu(t))dx \tag{11}$$

Here, $t$ is the time to the next departure (measured in hours), as defined above. The probability density function, $f(x; \mu(t))$, is defined as in formulation 1. The total user cost is the sum of the time to the next departure ($t$) and the probability of not being able to board the vessel times the headway, $1/\lambda$. We, thus, assume that no user must wait for more than one extra departure.

This is a simplified representation of the user cost, as the time to the next departure $t$ is not dependent on the frequency through the 'normal' waiting time, that is, the costs are only related to not being able to board through frequency. However, we contend that it is interesting to consider the results, as then we can make the users' decisions a part of the model.

We can now use the definition of the probability density function to find an explicit expression for the cost as follows:

$$H(t, \lambda) = \frac{1}{\lambda}\int_k^\infty f(x; \mu(t))dx = \frac{1}{\lambda}(1 - \Phi(z(t, \lambda))) \tag{12}$$

where, $\Phi(.)$ is the standard normal cumulative distribution function. Consequently, the user cost (minutes/PCE) becomes the following:

$$H(t, \lambda) = t + \frac{1}{\lambda}(1 - \Phi(z(t, \lambda))) \tag{13}$$

We can write $z$ as follows:

$$z = \frac{k - \mu(t)}{\sigma} = \frac{k - \left(\frac{1}{\lambda} - t\right)Y}{\sigma} \tag{14}$$

where $\sigma$ is the standard deviation of demand. We are now able to formulate the second optimization problem based on planned arrival as follows

[11] We would like to thank an anonymous reviewer for suggesting this example.

$$\min_{t,\lambda} Y\left(t+\frac{1}{\lambda}(1-\Phi(z(t,\lambda)))\right)+\lambda q_L L \tag{15}$$

s.t.

$$0 < t < \frac{1}{\lambda}$$

As in formulation 1, $q_L$ is the cost per sailed kilometer (NOK/km), with $L$ being the round-trip length (twice the length) measured in kilometers. The one constraint ensures that $t$ is positive (one does not arrive at the quay after departure) and $t < 1/\lambda$, which ensures that one does not arrive at the quay before the preceding departure. A description of the numerical solution method is given in appendix 3.

## 5. Numerical examples

To illustrate our model, we provide some numerical examples based on real-world data. Our application will first demonstrate the consequence of setting the capacity at the expected demand, when the demand, in fact, is random. Second, using different levels of vessel capacity, we show how the optimal frequency changes when the cost of not being able to board is taken into consideration.

### 5.1. Data description

We obtained data on the demand patterns, capacity and length from the Norwegian Public Roads Administrations database (fdb.triona.no) for one month in 2015. Table 1 displays some summary statistics for each crossing, including the length, capacity and peak demand. Table 2 shows the number and sizes of each ferry operated at the crossings. The data are obtained from Fjordfahren (2017). Moss-Horten is the largest crossing in terms of PCEs transported, while Halhjem-Sandvikvåg is the longest. Mortavika-Arsvågen is the shortest one. These crossings are selected because they are the three largest ones currently operating in Norway.

Table 2 shows the capacity in PCE for each ferry. PCE is a composite measure that takes into account the amount of space a vehicle occupies on a ferry in relation to a 'standardized' passenger car (see footnote 7). The average capacity per ferry is 192 PCE. To estimate the average capacity at each crossing, the capacities of all the vessels are combined using a simple arithmetic average at each crossing. Consequently, Mortavika-Arsvågen has the largest capacity at 220 PCE, Halhjem-Sandvikvåg at 212 PCE and Moss-Horten at 167 PCE.

To highlight the difference between our formulation and Mohring's formulation, we adjust the average size of each ferry in some scenarios by a 50% reduction. In all scenarios, capacity, $k$, is set equal to 2 times the size of the vessel, as we calculate the optimal capacity by using the cost of a complete round-trip.

We have set the value of the travel time to 89 NOK/Hour, as measured in 2018 prices (NPRA, 2018), which corresponds to the value of time for leisure-related car trips below 70 km in length with a single individual present in the car. The value of time is normally determined by the average composition of a trip's purpose, number of passengers and trip length, according to the official Norwegian guidelines (NRPA, 2018). For example, a heavy goods vehicle will have a higher time value than a trip to work in a passenger car, according to NPRA (2018).

**Table 1**
Data on ferries operating in the three case studies – speed and capacity at the links (2017).

| Variable | Moss-Horten | Halhjem-Sandvikvåg | Mortavika-Arsvågen |
|---|---|---|---|
| Demand [PCE/day] | 10956 | 5698 | 8707 |
| Round-trip length [km/round trip] | 20 | 44 | 16 |
| Peak demand [PCE/hour] | 874 | 463 | 720 |

**Table 2**
Statistics from www.Fjordfahren.de (accessed July 2017).

| Name | Capacity [PCE] | Link |
|---|---|---|
| Mastafjord | 212 | Mortavika-Arsvågen |
| Stavangerfjord | 212 | Mortavika-Arsvågen |
| Boknafjord | 238 | Mortavika-Arsvågen |
| Bastø I | 200 | Moss-Horten |
| Bastø II | 200 | Moss-Horten |
| Bastø III | 212 | Moss-Horten |
| Bastø VII | 106 | Moss-Horten |
| Bastø VIII | 115 | Moss-Horten |
| Fanafjord | 212 | Halhjem-Sandvikvåg |
| Bergensfjord | 212 | Halhjem-Sandvikvåg |
| Raunefjord | 212 | Halhjem-Sandvikvåg |
| **Average** | **192** | |

The value of time is weighted by a waiting time factor of 1.3, which is the average of the Norwegian Public Roads Administration's guidelines for short and long headways (NPRA, 2018). For waiting times between 0 and 5 min, a factor of 2.3 is used, and for waiting times above 60 min, a factor 0.28 is used.

Other authors (see, e.g., Jørgensen and Solvoll (2018) [12]) have used a higher estimate for the value of time compared to our estimate. However, as we have not included capital costs into the model, a conservative estimate on the benefits seems appropriate. Increasing the frequency only affects the sailing costs in our model. If one needs additional ferries to increase the frequency above a certain point, capital costs will accrue. Although we acknowledge our estimate to be low, we contend it is beneficial, as we do not include all costs. Moreover, the focal point is the comparison of our method against Mohring's. As the same value of time is applied in all the model formulations, we believe the differences between the models to be the most interesting feature.

If these formulas are to be applied in practice, one should perform a sensitivity analysis of the results with respect to costs and values of time. As the value of time is highly sensitive to the specific composition of travelers and their trip purposes, the central input parameters should be viewed with a degree of skepticism and tested. For example, using a value of travel time for longer trips (>70 km) might be relevant in the car ferry sector. The focus in the current paper is on comparing methods, and a sensitivity analysis may be out of its scope, as viewed from the authors' perspective.

At the time of writing this paper, a project to update value of time estimates for Norway was in progress, where among other items, new values for goods transportation were estimated (Halse et al. (2019)). However, the results were not publicly available at the time of writing and is therefore not used.

An important question is whether the waiting time components are valued differently by users. For example, a hidden waiting time may have a lower time value compared to that of an open waiting time. The former can be used for other activities at home, reducing the alternative cost. For example, Jørgensen and Solvoll (2018) used a hidden waiting time of half the value of the open waiting time. Moreover, waiting for the next departure after failing to board the first one might be viewed as a delay from the users' perspective. Consequently, the third waiting time component (excess open waiting time) might be valued higher than both the hidden and "ordinary" open waiting times. In our estimations, we have assumed equal valuations for all components, which is a simplification.

The operational sailing cost per kilometer is set on the basis of a statistical approach. Panel data analysis has proven to be the preferred

[12] Jørgensen and Solvoll (2018) used 353 NOK/hour for the open waiting time and 176 NOK/hour for the hidden waiting time, measured in 2013-NOK; using a 13% price increase (gathered from Statistics Norway 2013–2018), this value amounts to 398 and 198 NOK, respectively. The average of these two values indicates that we have a very conservative value.

econometric method to estimate the marginal cost in the Norwegian ferry sector (see, e.g., Mathisen, 2008; Mathisen & Jørgensen, 2007; Mathisen & Jørgensen, 2012). Our panel model of operational costs is very simple and is only meant to provide an average marginal cost per sailed kilometer. The data in question are based on 53 links over 8 years for a total of 424 observations, from 2002 to 2010, including all the major links in Norway.

The marginal cost is determined by a simple regression of the cost at link $i$ at time $t$ ($K_{it}$), as follows:

$$K_{it} = \alpha + b_i + SKM_{it}\beta + \theta_i + \varepsilon_{it} \tag{16}$$

where $\alpha$ is a common constant at all links and $b_i$ is a fixed effect at link $i$ that represents the constant and unobserved factors that influence the cost. $SKM_{it}$ is the sailed kilometers per year at the link, with $\beta$ as the marginal cost, which is our parameter of interest. The last two terms represent the error structure of our model. In a panel analysis, there is often autocorrelation within each unit. To correct standard errors, researchers often use so-called panel-robust standard errors, where both heteroscedasticity and autocorrelation are controlled at the panel level (Cameron & Trivedi, 2005). The term $\theta_i$ is a correlation of the cost in unobserved factors over time at link $i$, while $\varepsilon_{it}$ is a common error term for all links. A separate model with a constant per mooring was also tried but gave imprecise measurements of the unit costs.

The estimated cost per kilometer was found to be 160 NOK, which has a high degree of significance in the normal statistical sense, measured in 2018-NOK.

## 6. Results

We now give the results, using our framework in contrast to equation (1), starting by assessing the consequences of setting the capacity at the mean demand.

We show the optimal frequency as calculated by our models under different assumptions regarding the total capacity of the vessels. As shown in the methodology section, when sufficient capacity is provided (that is, sufficiently large vessel are being used), Mohring's model and our model should yield similar results; however, when there are bounds on the capacity/vessel size, they should differ.

Two scenarios are tested, i.e., one in which capacity is kept at its current level and one in which it is reduced by 50%, such that there is an insufficient level. Our aim is to show how the model yields different results than the standard Mohring model under assumptions of constraints on capacity. The numerical estimates are not meant to be exact estimations of the optimal frequency levels at these specific crossings, but rather illustrate when and why our model yields different results from the standard one being applied in the literature.

Figs. 1–3 show the mean demand (PCE/hour), one standard deviation and the probability of running full vessels (demand $\geq$ capacity)[13] with the capacity set at the maximum (mean) expected demand (PCE/hour) throughout the day (mean demand in the peak hour). The figures clearly show that setting the capacity equal to the expected number of PCE/hour (the mean), will raise the probability of obtaining a full vessel to approximately 50% at peak demand. For all links, there is a non-negligible likelihood that not all users are able to board on their first attempt during rush hours.

Fig. 4 displays the expected number of PCE in each time period that are not able to board the first vessel arriving at the quay when setting the capacity at the maximum (mean) demand.[14] At peak levels, approximately 55 PCE are not able to board during each hour of operation. Moreover, it should be noted that the variation in demand (standard deviation of PCE/hour) is of such a magnitude that even in periods of lower demand, one expects some instances of the capacity to not be sufficient to cover the demand. This situation highlights the importance of not only considering the mean demand but also taking its variation into account.

Fig. 5 displays the expected share of PCE that are left behind during each day of operation when setting the capacity at the peak mean demand. The main finding is, thus, that setting the capacity at the mean demand in the peak period will yield, on average, 2% of users being left behind at the quay.[15] This percent of users amounts to a maximum of approximately 58 PCEs for Mortavika-Arsvågen, 52 for Moss-Horten and 38 for Halhjem-Sandvikvåg per peak hour. Thus, when using an optimization framework in which the mean demand is being used to optimize the frequency, some costs may not be taken into account, possibly biasing the estimated optimal frequency downwards. This is an important point to consider for policy-makers that employ planning methods to design service levels.

Jørgensen & Solvoll (2007) showed how one could derive the optimal capacity of a ferry for a given level of service quality, measured as the number of users left behind. The current calculations support their view that the service quality should be investigated in an explicit optimization framework. However, we contend that policy-makers should aspire to investigate the optimal service level within a framework that simultaneously determines the frequency and service level on the basis of user and operating costs.

### 6.1. Optimal frequency with current vessel capacity

We now move on to the optimal frequency as estimated by our two different formulations. Figs. 6–8 show, for each link, the optimal number of departures using formulation 1, formulation 2, and the simplified standard model (Mohring, 1972). We did not include extra delay costs or weighing in these estimations. We assume that the capacity of each vessel is its present capacity.

At Mortavika-Arsvågen, formulation 1 and Mohring's model yield approximately the same results, as is expected when there is a sufficient level of capacity provided, which confirms our earlier statements: when using larger ferries, the frequencies as calculated with or without the third waiting time component should converge. The frequency is higher in periods of higher demand. Formulation 2 yields different results, which are discussed further below.

At Halhjem-Sandvikvåg and Moss-Horten, the results are broadly in line with what is observed at Mortavika-Arsvågen, where optimal frequencies as calculated by the Mohring model and our models yield quite similar results.

Moreover, it is worth noting that formulation 2 tends to yield lower frequencies compared to those of the other methods, most likely due to its formulation, which focuses on scheduling costs rather than waiting time directly, as discussed in the methodology section. Consequently, there needs to be a sufficiently large probability of not being able to board for it to yield at least the same level of frequency as the framework of Mohring (1972).

The results presented in this section confirm the statements made in our methodology section; when there is sufficiently high capacity provided (measured as the size of vessels), our formulation 1 and Mohring's will converge to the same values. Thus, when the vessels are 'sufficiently' large, Mohring's formula can be applied to calculate all the

[13] This is calculated as the probability of demand exceeding capacity per hour, given as $1 - \int_0^Q f(x)dx$, where "$x$" is the demand per hour and "$Q$" is the total capacity, i.e., the total number of departures * capacity of each vehicle. Here, "$Q$" is set to be equal to the maximum PCE/hour in the data (i.e., in the peak period).

[14] This is calculated by using equation (7), where maximum denotes the peak period.

[15] In the Norwegian guidelines on service quality, no more than 2% of vehicles should be unable to board the first vessel they desire, as a yearly average.

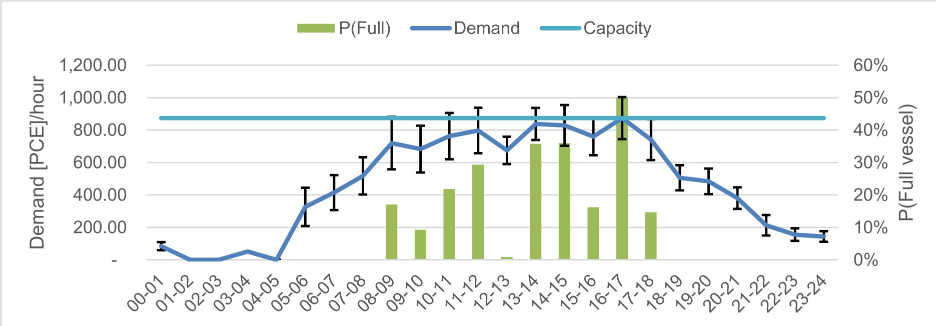



**Fig. 1.** Probability of a full vessel, where the capacity is set equal to the mean demand in the peak period. Moss-Horten ferry link.

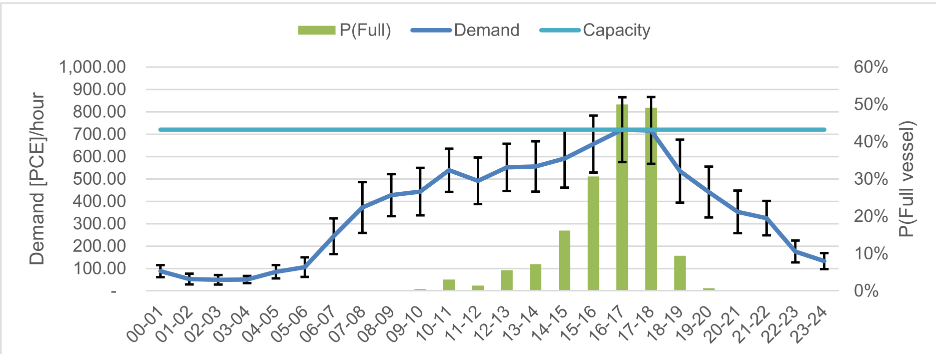



**Fig. 2.** Probability of a full vessel, where the capacity is set equal to the mean demand peak period. Mortavika-Arsvågen ferry link.

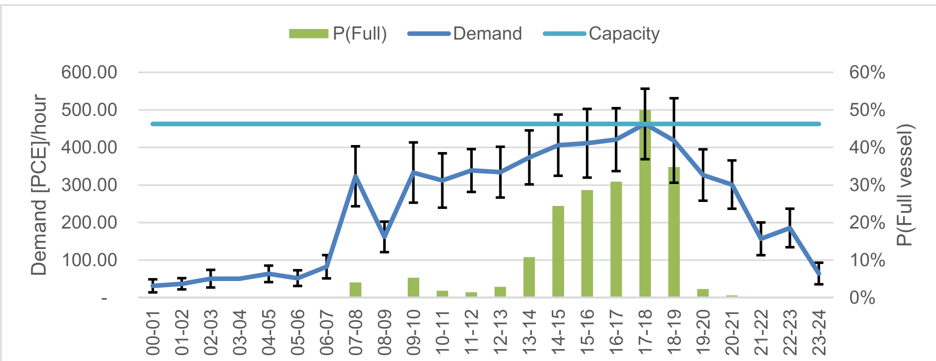



**Fig. 3.** Probability of a full vessel, where the capacity is set equal to the mean demand in the peak period. Halhjem-Sandvikvåg ferry link.

relevant waiting time costs.

As the capacity may have been adjusted to remove most of the passengers left behind at the *specific* crossings we study, it is interesting to investigate the consequence of changing the vessel size to illustrate some *general* points. Our aim is to gauge when and why the model yields different results compared to those of the standard model. Thus, we now move on to assess what happens when there is a lower level of capacity per vessel available.

### 6.2. Optimal frequency with the vessel capacity reduced by 50%

The optimal frequency results, as estimated by our two different formulations, are now presented, with a vessel capacity (the size of each individual ferry, measured in PCE) that is reduced by 50%. Figs. 9–11 show, for each link, the optimal number of departures using formulation 1, formulation 2, and the simplified standard model (Mohring, 1972).

At all links, the frequencies differ. The difference is the smallest at Mortavika-Arsvågen and is somewhat higher at the other two. The former link is shorter, meaning it is less costly to operate at a higher

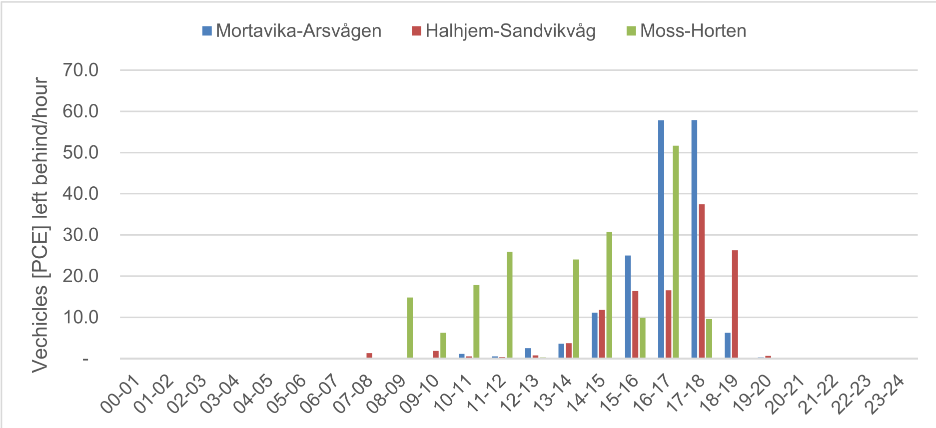



**Fig. 4.** Expected number of vehicles [PCE] left behind/hour when setting the capacity at the maximum of the mean demand throughout the day.

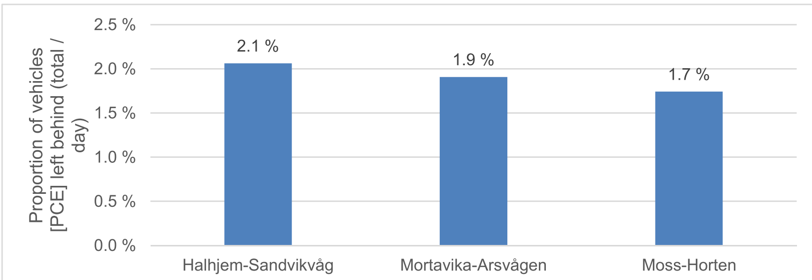



**Fig. 5.** Expected percentage of vehicles [PCE] left behind during one day of operation when setting the capacity at the maximum of the mean demand throughout the day.

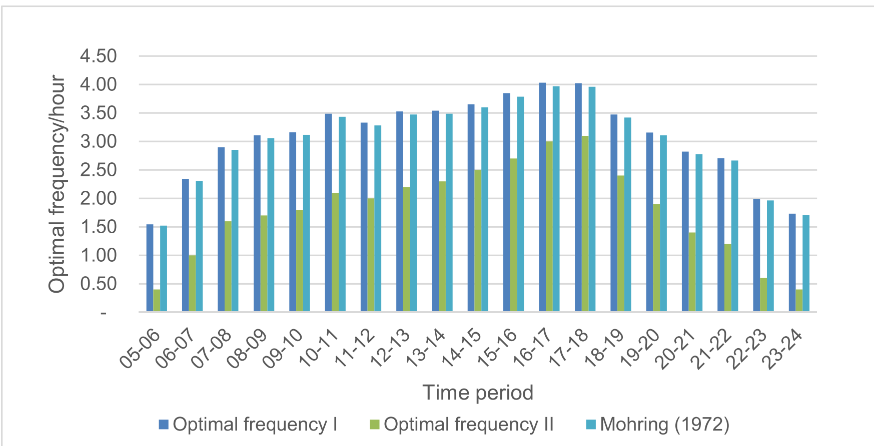



**Fig. 6.** Optimal frequency results given by the different models at Mortavika-Arsvågen with the current vessel capacity.

frequency, which might partly explain the effect. It is then already optimal to have a frequency as high as that derived from Mohring (1972), as the crossing is short.

At Halhjem-Sandvikvåg and Moss-Horten, the differences are more pronounced. Halhjem-Sandvikvåg is the longest crossing, with a round trip length of 44 km. Consequently, it is more costly to increase the frequency; thus, a lower optimal frequency is found when using Mohring (1972)'s framework. However, as the new models are applied, the true

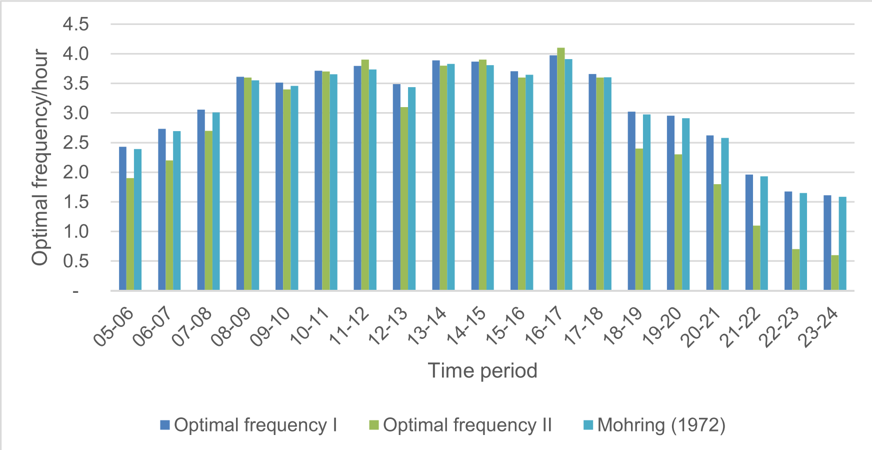



**Fig. 7.** Optimal frequency results given by the different models at Moss-Horten with the current vessel capacity.

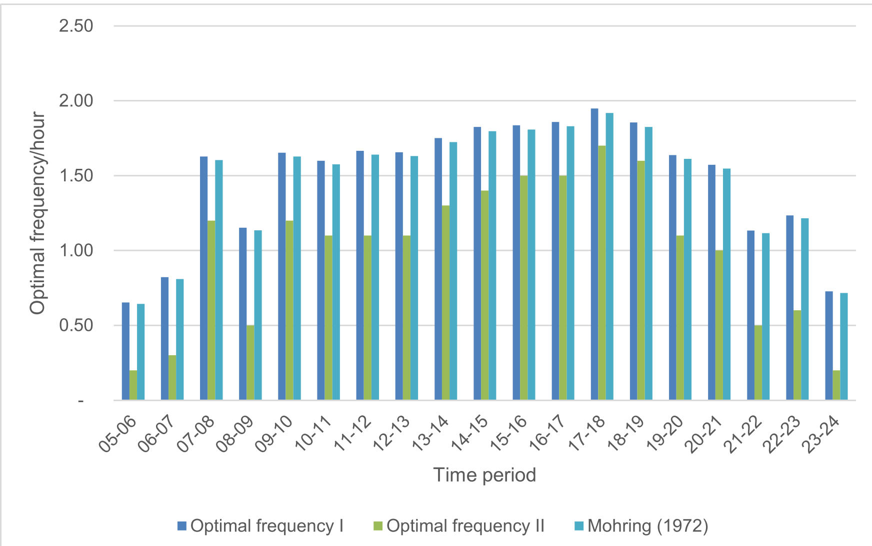



**Fig. 8.** Optimal frequency results given by the different models at Halhjem-Sandvikvåg with the current vessel capacity.

costs of not providing sufficient capacity are captured to a greater extent, which sharply increases the optimal frequency.

Moss-Horten is a relatively short link, with a round trip length of only 20 km, i.e., half that of Halhjem-Sandvikvåg. However, in our dataset, the average ferry operated at Moss-Horten has a lower capacity than the two other links. Consequently, the difference in the optimal frequency result when using the new framework versus Mohring (1972)'s is higher, as more departures are required to transport the same number of vessels when the ferries are smaller (even at 50% of the current capacity).

Finally, we note that both formulation 1 and formulation 2 (users arriving randomly and planning ahead) yield similar optimal frequencies. This result might at first seem somewhat odd, but they are both directly linked to the required capacity to minimize user cost, irrespective of the objective function formulation.

When using formulation 2, our model indicated that it was always optimal to have a frequency (hence, capacity) such that users could arrive at the quay just moments before departure, that is, $t = 1$ min, for all crossings (the minimum value we chose, as users must spend some time buying a ticket for the ferry). When using formulation 1, we saw that it was optimal to provide a capacity level such that virtually all users were able to board on their first try. Thus, both formulations indicate that the user costs, on these specific links, far outweigh the operational costs, irrespective of the objective function, when there are limits on the size of each vessel, as measured in PCE. That is, one should consider providing a capacity such that users do not need to worry about being able to board the first vessel that arrives. As the capacity level required to reach such a level is the same in both formulations, they also yield similar results.

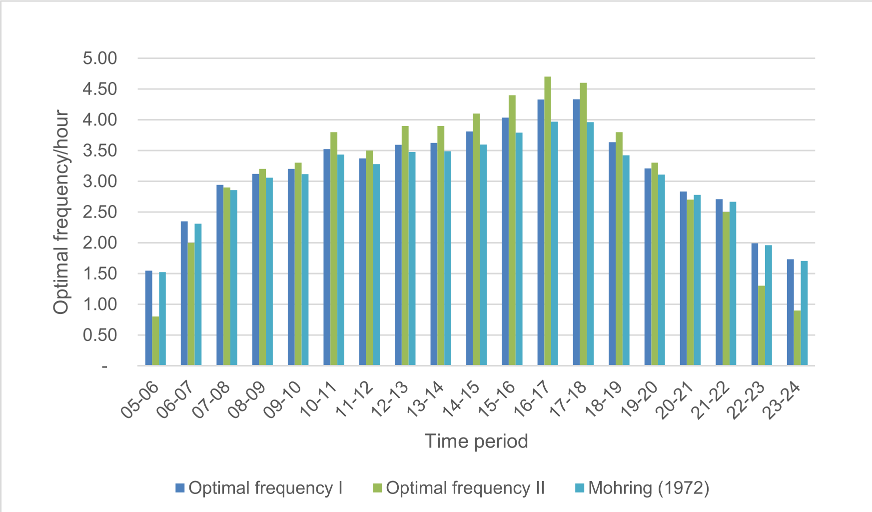



**Fig. 9.** Optimal frequency results given by the different models at Mortavika-Arsvågen.

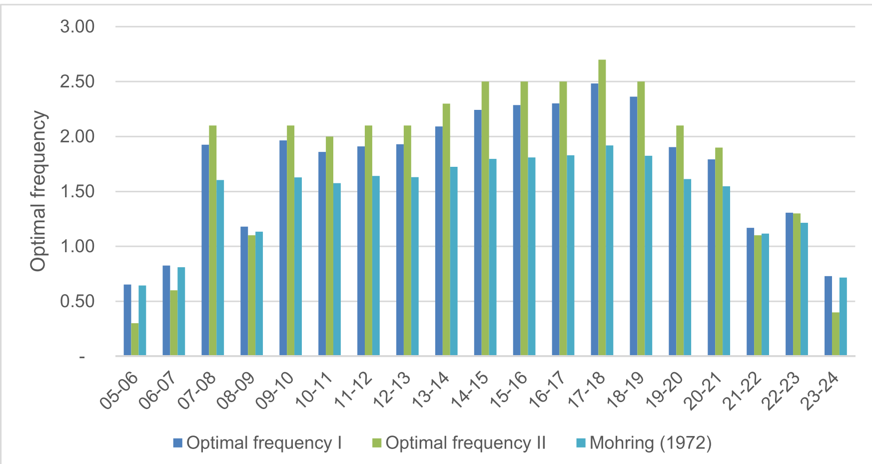



**Fig. 10.** Optimal frequency results given by the different models at Halhjem-Sandvikvåg.

## 7. Conclusion

In this paper, we develop a framework for setting optimal public transit frequencies, with the costs of not being able to board due to capacity restrictions included. To the best of our knowledge, this is the first paper in the literature to treat this problem from social planning and economic perspectives. We developed two different formulations, i.e., one in which users arrive randomly at quays and one in which they select an optimal arrival time.

We show that the standard optimal frequency model effectively underestimates user costs when there are limits on the vessel capacity (the size of each individual ferry). When the extra user costs stemming from not being able to board the first vessel arriving at the quay are taken into account, the optimal frequency is higher.

The model is applied to three different car ferry crossings in Norway to provide numerical illustrations of the methodology. Our empirical results are based on conservative values of time and strongly suggest that the frequency should be increased to promote social efficiency if there are constraints on the vessel size, as compared to Mohring's model. Our example confirms that Mohring's model and our model (formulation 1) give the same results when sufficiently large vessels are used.[16] In general, the two model formulations yield the same results suggesting that formulation 1 (based on Mohring) can be applied even though it does not assume any queuing at the quay (as formulation 2 does).

The policy recommendation is clear: Both the theoretical and empirical findings suggests that the "third" waiting time component of not being able to board the first vessel is highly relevant for decision-

[16] As we only examined one specific example for illustrational purposes, it would be interesting to examine cases in which smaller vessels or a higher demand/capacity ratio is present in future studies.

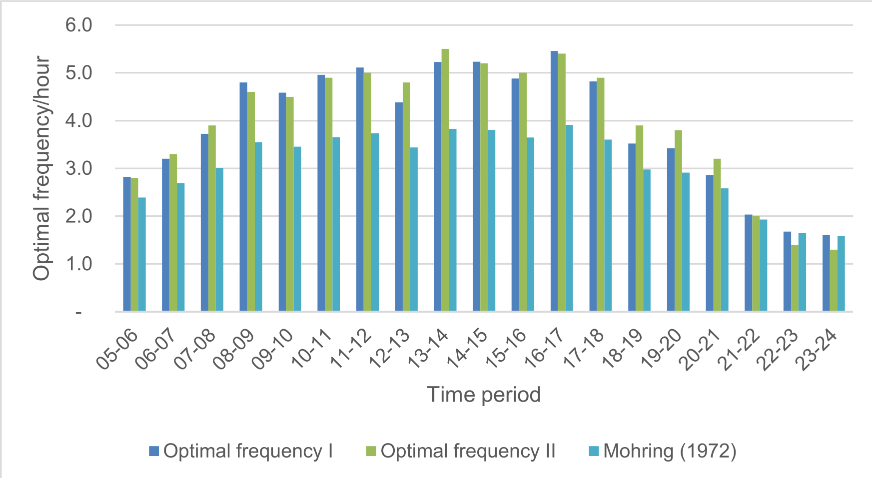



**Fig. 11.** Optimal frequency results given by the different models at Moss-Horten.

makers. Not including the extra waiting time costs stemming from capacity constraints leads to different optimal service levels when the vessel capacity is constrained, which clearly indicates the relevance of our findings for practical policy making.

It is also important to mention some limitations of the framework, which include the following:

- It does not explicitly weight open and hidden waiting times differently;
- It assumes an equal probability of not being able to board for each departure within a specific time period; and
- It assumes no users must wait for more than one departure.

It would be interesting to relax these assumptions in future work.

## Appendices

*Appendix 1. Derivation of the objective function in formulation 1*

To derive numerical expressions for the two cost components in formulation 1, we use standard results from the supply chain literature. The standardized loss function can be written as follows by using a special feature of the normal distribution (see Nahmias, 2005):

$$G(z) = (\varphi(z) - z(1 - \Phi(z))) \tag{A1}$$

where $\varphi(z)$ is the standard normal probability density function (pdf) and $\Phi(z)$ is the corresponding cumulative probability function (cdf). Using this definition, we can now write the number of users in group B as follows[17]:

$$E(B) = \sigma \int_{k\lambda}^{\infty} (x - k\lambda) f(x) dx = \sigma(\varphi(z) - z(1 - \Phi(z))) \tag{A2}$$

Further, we define $z = \frac{k\lambda - \mu}{\sigma}$, with all terms following their prior definitions. Bear in mind that we have separated the total expected demand, $\mu$, into two separate groups, i.e., A and B. We can now use this fact to calculate $E(A)$ by using the following equation:

$$E(A) + E(B) = \mu \tag{A3}$$

Thus, we obtain the number of users that are not "left behind" by using $E(B)$ and its definition, as follows:

$$E(A) = \mu - E(B) = \mu - \sigma(\varphi(z) - z(1 - \Phi(z))) \tag{A4}$$

The total user cost can now be written as follows:

$$C_A(\lambda) + C_B(\lambda) = \frac{\omega}{2\lambda}(\mu - \sigma(\varphi(z) - z(1 - \Phi(z)))) + \frac{3\omega}{2\lambda}\sigma(\varphi(z) - z(1 - \Phi(z))) \tag{A5}$$

and further simplified the equation by using the definition of $E(B)$, resulting in the following:

$$C_A(\lambda) + C_B(\lambda) = \frac{\omega}{2\lambda}[\mu - E(B) + 3E(B)] = \frac{\omega}{2\lambda}[\mu + 2E(B)] \tag{A6}$$

[17] This function is called the loss-function, which is equal to the standardized loss function multiplied by the standard error, $\sigma$. We temporarily use the notation *E(B)* instead of *J(x)* (as in equation (2)) to underline the separation of costs between the two groups (A and B).

We are now able to write the function in a more simplified manner and define the total cost (user and production) as the function $\Omega(\lambda)$ as follows:

$$\min_{\lambda}\Omega(\lambda)=\frac{\omega}{2\lambda}[\mu+2J(\lambda)]+\lambda C_L L \tag{A7}$$

where $J(\lambda)=E(B)$, to emphasize its dependence on frequency. The operating cost is added to the equation, where $C_L$ is the cost per kilometer and $L$ is the length of the route (in one round-trip). An increase in the frequency may require more vessels, thus yielding higher capital costs, which is not a part of this cost formulation.

*Appendix 2. Derivation of the expected cost when setting the vessel capacity equal to the expected demand*

Here, we derive the user costs when setting the capacity equal to the expected (mean) demand, assuming a normally distributed demand. Using our above framework, the user cost ($C_A(\lambda)+C_B(\lambda)$) is defined as follows:

$$\frac{\omega}{2\lambda}\int_0^{k\lambda} xf(x)dx+\frac{3\omega}{2\lambda}\sigma\int_{k\lambda}^{\infty}(x-k\lambda)f(x)dx \tag{A8}$$

Following the assumption of a standard normal distribution, we have a symmetrical distribution, which means that there is an equal probability mass on each "side" of the mean value. Thus, it is equally likely to observe a value that is greater than the mean and one that is lesser than the mean. According to this fact alone, it should be intuitively clear that setting the capacity to the mean demand will result in a capacity that is too low half of the time. Thus, the cost then becomes the following:

$$\frac{\omega}{2\lambda}\int_0^{\mu} xf(x)dx+\frac{3\omega}{2\lambda}\sigma\int_{\mu}^{\infty}(x-k\lambda)f(x)dx=\frac{1}{2}\left(\frac{\omega}{2\lambda}+\frac{3\omega}{2\lambda}\right)=\frac{\omega}{\lambda}>\frac{\omega}{2\lambda} \tag{A9}$$

Using the fact that in a symmetrical distribution $\int_0^{\mu} xf(x)dx$ equals 50% of $\mu$ and $\sigma\int_{\mu}^{\infty}(x-k\lambda)f(x)dx=\mu-\int_0^{\mu}xf(x)dx$, as discussed above, if we assume no random variation in demand and neglect the cost of users being unable to board, $C_B(\lambda)$, we obtain costs equal to the following:

$$C=\frac{\omega}{2\lambda} \tag{A10}$$

which is the "standard" half-headway rule.

*Appendix 3. Solution methods*

The first optimization model is a convex function of $\lambda$; thus, one should be able to use fairly simple numerical methods to find its minimum. As the analytic solution is generally not available, we use numerical optimization techniques in this paper.

A simple and popular method for minimizing a convex function is gradient descent (Nocedal, 2006). This method works by continuously moving the solution in the direction with the greatest descent rate. The optimization variable is updated via the following equation:

$$\lambda_{n+1}=\lambda_n-\alpha\Omega'(\lambda_n) \tag{A11}$$

The variable $\alpha$ is a so-called "step-size" parameter, with $\Omega'(\lambda_n)$ being the gradient of the function to be minimized. The gradient is numerically approximated by the following equation:

$$\Omega'(\lambda_n)\approx\frac{\Omega(\lambda_n+h)-\Omega(\lambda_n)}{\Delta} \tag{A12}$$

where $\Delta$ is a small positive perturbation to the value of $\lambda_n$. The iterative process continues until some convergence criteria are met. We have chosen the convergence criteria of the absolute difference in the objective function value between two successive iterations, $\delta=|\Omega(\lambda_n)-\Omega(\lambda_{n-1})|$. Thus, when $\delta$ falls below a prespecified level, $\varepsilon$, the iterations terminate. We used the following parameters in the optimization: $\Delta=10^{-3}$, $\gamma=10^{-3}$ and $\varepsilon=10^{-5}$.

The objective function of the second optimization model is not generally convex, as we have confirmed by plotting the objective function graphically. As the range of the optimization variables are naturally restricted by the constraint ($t$) and by a reasonable bound on the number of departures ($\lambda$), a simple exhaustive search procedure is applied. For each level of frequency, the arrival times of users from 1 to $1/\lambda$ are tested, and the values yielding the lowest cost are retained.

**Appendix 4. Deriving waiting time cost with infinite capacity**

We first start with the costs of users that are able to board the first vessel arriving at the quay. This limit is simply given by the following

$$\lim_{k\to\infty}C_A=\frac{\omega}{2\lambda}\int_0^{\infty}xf(x)dx=\frac{\omega}{2\lambda}\mu \tag{A13}$$

It can be observed that $\int_{-\infty}^{\infty}xf(x)dx=\mu$ by the very definition of the statistical mean. As one cannot observe a negative demand, one only needs to integrate from 0 to $\infty$ to obtain the mean. The expression $\frac{\omega}{2\lambda}\mu$ is the exact same expression measuring the user cost in the standard model of the optimal frequency. Next, we need to show that the cost of users "left behind" ($C_B$) converges to zero when $k\to\infty$. Note that, in general, the integral from $z$ to $z$ is always zero; $\int_z^z xf(x)dx=0$. When $k\to\infty$, we also have $\lambda k\to\infty$. Consequently, the limit of $C_B$, ($\frac{3\omega}{2\lambda}\sigma\int_k^{\infty}(x-k\lambda)f(x)dx$), now becomes zero.

Thus, the total user cost is now as follows:

$$C_A + C_B = \frac{\omega}{2\lambda}\mu + 0 = \frac{\omega}{2\lambda}\mu \tag{A14}$$

This is the same as in the standard model of Mohring (1972).

## References


Andersen, S. N., & Tørset, T. (2018). In *Waiting time in ferry services: An empirical study from a Norwegian context (No. 18-01401). Transportation research board 97th annual meeting*. Washington DC, United States, 2018-1-7 to 2018-1-11.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge university press.

Fjordfahren. (2017). http://www.fjordfaehren.de/no/fjord_alf.htm.

Fosgerau, M., & Karlström, A. (2010). The value of reliability. *Transportation Research Part B: Methodological, 44*(1), 38–49.

Fosgerau, Mogens (2009). The marginal social cost of headway for a scheduled service. *Transportation Research Part B: Methodological, 43*(8–9), 813–820.

Furth, P. G., & Wilson, N. H. (1981). Setting frequencies on bus routes: Theory and practice. *Transportation Research Record, 818*, 1–7, 1981.

Hadas, Y., & Shnaiderman, M. (2012). Public-transit frequency setting using minimum-cost approach with stochastic demand and travel time. *Transportation Research Part B: Methodological, 46*(8), 1068–1084.

Halse, A., Mjøsund, C., Killi, M., Flügel, S., Jordbakke, G. N., Hovi, I. B., et al. (2019). *Bedrifters verdsetting av raskere og mer pålitelig transport. Den norske verdsettingsstudien for godstransport 2018*. Institute of Transport Economics Technical Report 1680/2019.

Hanssen, T. E. S., Jørgensen, F., & Larsen, B. (2019). Determinants affecting ferry users' waiting time at ferry terminals. *Transportation*, 1–22.

Herbon, A., & Hadas, Y. (2015). Determining optimal frequency and vehicle capacity for public transit routes: A generalized newsvendor model. *Transportation Research Part B: Methodological, 71*, 85–99.

Jansson, J. O. (1980). A simple bus line model for optimisation of service frequency and bus size. *Journal of Transport Economics and Policy*, 53–80.

Jara-Díaz, S., Fielbaum, A., & Gschwender, A. (2017). Optimal fleet size, frequencies and vehicle capacities considering peak and off-peak periods in public transport. *Transportation Research Part A: Policy and Practice, 106*, 65–74.

Jara-Díaz, S., & Gschwender, A. (2003). Towards a general microeconomic model for the operation of public transport. *Transport Reviews, 23*(4), 453–469.

Jara-Díaz, S. R., & Gschwender, A. (2009). The effect of financial constraints on the optimal design of public transport services. *Transportation, 36*(1), 65–75.

Jørgensen, F., Mathisen, T., & Solvoll, G. (2007). *Kapasitet i fergesektoren. Technical report SIB-rapport 1/2007*. Handelshøgskolen i Bodø.

Jørgensen, F., & Solvoll, G. (2018). Determining optimal frequency at ferry crossings. *Transport Policy, 63*, 200–208.

Mathisen, T. A. (2008). Marginal costs and capacity utilization: Calculating short-run, medium-term, and long-run marginal costs in the ferry industry. *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, 373–389.

Mathisen, T., & Jørgensen, F. (2007). Using cost norms analyses to reveal subsidy fraud in out-contracted public transport services. In *Proceedings of the European transport conference 2007* (Leiden, Netherlands).

Mathisen, T. A., & Jørgensen, F. (2012). Panel data analysis of operating costs in the Norwegian car ferry industry. *Maritime Economics & Logistics, 14*(2), 249–263.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *The American Economic Review, 62*(4), 591–604.

Nahmias, S., & Cheng, Y. (2005). *Production and operations analysis* (Vol. 6). New York: McGraw-hill.

National Public Roads Administration. (2018). *Håndbok V712 – konsekvensanalyser*.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Oldfield, R. H., & Bly, P. H. (1988). An analytic investigation of optimal bus size. *Transportation Research Part B: Methodological, 22*(5), 319–337.

Osuna, E. E., & Newell, G. F. (1972). Control strategies for an idealized public transportation system. *Transportation Science, 6*(1), 52–72.

Pedersen, P. A. (2003). On the optimal fare policies in urban transportation. *Transportation Research Part B: Methodological, 37*(5), 423–435.

Samstad, H., Ramjerdi, F., Veisten, K., Navrud, S., Magnussen, K., Flügel, S., & San Martin, O. (2010). Den norske verdsettingsstudien. Sammendragsrapport. *TØI rapport, 1053*, 2010.

Thonemann, U. (2010). *Operations management: Konzepte, methoden und anwendungen*. Pearson Deutschland GmbH.

# Paper III

Case Studies on Transport Policy 10 (2022) 41–56

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

# Assessing the socially optimal capacity at a selection of Norwegian car ferry crossings

Harald Høyem [a,b,*], James Odeck [b,c]

[a] Asplan Viak, 0153 Oslo, Norway
[b] Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway
[c] Norwegian Public Roads Administration, 2605 Lillehammer, Norway

ARTICLE INFO

*Keywords:*
Transport sector
Car ferry services
Optimization
Optimal frequency

ABSTRACT

Car ferry services constitute an important part of the transportation network in several parts of the world, especially in areas with limited alternative modes of transport. A central problem facing decision makers is the socially optimal capacity of ferry services. However, the literature has not examined all the decision variables the are relevant to decision makers in a simultaneous framework, only partially. We add to the literature by treating all the relevant decisions variables in a simultaneous framework, which enables a more complete representation of optimal capacity, than partial frameworks. Our proposed methodology also includes the cost of not being able to board the first arriving departure, which is an essential cost in the case of car ferries with too low capacity. We apply the methodology to a case study of three major ferry crossings in Norway. Results indicate that a too large capacity is provided. Thus, local policy makers should consider revising the current service levels. Other policy makers may enact better decisions based on the findings we provide. Sensitivity test suggests that the method used to estimate the number of users not being able to board due to capacity concerns may be improved. This, however, does not alter our main conclusion.

## 1. Introduction and literature review

Ferry services is an important part of the transportation network in several areas throughout the world carrying both passengers and vehicles. For example, Norway maintains an extensive ferry network to ease accessibility between fjords and Islands. There are over 130 links with an estimated total social surplus of 6.6 billion NOK/year (Jørgensen et al., 2011), amounting to 33 million NOK/service[1] which underlines its importance. Denmark support about 65 ferry services and Greece covers over 360 itineraries (Baird & Wilmsmeier, 2011). Turkey, and in particular Istanbul, has a large operator of ferries for both passenger and car traffic. Ferries of various sizes are also used in Sweden and Scotland, to connect islands to the mainland. In the Americas, Washington (Washington State Department of Transportation, 2020), North Carolina (North Carolina State Department of Transportation, 2020), Alaska state (Department of Transportation and Public Affairs, 2020), and the province of British Columbia (BCFerries, 2020) are some of the governmental entities operating such services, some of whom are serving larger communities not accessible by roads[2]. In the case of Seattle, Zhang et. al. (2017) found ferries to be important in reducing CO2 emissions by limiting the trip distance for cars, enabling a shortcut over water. Further, many waterfront cities use ferry services for increased connectiveness without needing to invest in costly infrastructure including New York (NYC Ferry, 2020), London (Transport for London, 2020) and Toronto (City of Toronto, 2020). The importance of such services on uncongested rivers in major cities in underlined by Bignon & Pojani (2018).

Policy makers are often charged with planning the appropriate level of service and subsidization of transport. To enable policy makers to provide good recommendations about the appropriate quality of ferry services, all relevant costs and benefits to society must be accounted for. Such tools are instrumental in designing good services that meet the criteria set by decision makers, while maintaining efficient use of scarce resources.

A central question when designing ferry services is the level of capacity and quality that should be implemented. Increased quality,

* Corresponding author at: Asplan Viak, 0153 Oslo, Norway.
*E-mail address:* harald.hoyem@asplanviak.no (H. Høyem).

[1] 2019-NOK, adjusted for inflation by 30 % from 2007. 1 EUR ≈ 11 NOK. They included 97 services in their study.

[2] An example is the Alaskan state capital Juneau with over 30 000 inhabitants.

https://doi.org/10.1016/j.cstp.2021.10.008
Received 24 May 2021; Received in revised form 14 October 2021; Accepted 20 October 2021
Available online 16 November 2021
2213-624X/

interpreted as the number of departures per hour, reduces user waiting time costs, but increases production costs. Capacity is defined as the total number of passengers/vehicles transportable per hour as a composite of frequency and vessels size. When too low capacity is offered, waiting times increase as users may not be able to board the first arriving ferry (Findley et al., 2018). A central question for policy makers is thus how to optimize service levels in accordance with economic principles, such that one achieves the largest net difference between benefits and costs.

The transportation literature is scant with respect to optimization of service levels at ferry crossings. Further, there is a number of shortcomings that needs to be addressed before policy makers may be able to estimate optimal capacity at ferry crossings. The literature on ferry services has thus far investigated the question of optimal service levels by treating frequency in isolation (Jørgensen & Solvoll, 2018; Høyem & Odeck, 2020). When capacity calculations are included, only frequency is a variable (Høyem & Odeck, 2020), but not vessel size and number of vessels. Moreover, these studies assume a fixed level of demand and prices. Studies in the bus sector have included capacity costs and fleet size (Jara-Díaz and Gschwender, 2003; Jara-Díaz and Gschwender, 2009, Jara-Díaz et al., 2017; Jara-Díaz et al., 2020, Tirachini et al., 2014; Börjesson et al., 2017; Börjesson et al., 2019). Last, Škurić et al. (2020) investigated optimal capacity and number of ferries, but from a profit-maximizing perspective, not incorporating user costs directly.

The transport economics literature has investigated capacity in two ways (i) by assuming a fixed level of capacity for buses[3] as optimal and (see e.g., Jara-Díaz et al., 2017) (ii) by including on-board crowding costs on buses with respect to demand level and size of each vehicle (see, e.g., Börjesson et al., 2017; Tirachini et al., 2014). However, these approaches highlight two main shortcomings of the literature that remain unresolved when it comes to ferry services.

- **(i) Assuming an exogenous level of total capacity**: Firstly, assuming a fixed level of optimal capacity circumvents any trade-offs between higher production and/or capital costs versus lower user costs by increasing capacity of each vehicle. Consequently, capacity should arguably be an *endogenous* variable in the optimization by fully describing user costs associated with different levels of it. Jørgensen & Solvoll (2018), implicitly makes such an assumption in the case of car ferries.
- **(ii) Using on-board crowding as a metric of user costs related to capacity**: Studies that use on-board crowding as the user cost associated with a certain level of capacity, maintains lesser relevance in the ferry sector for the following reasons.
  - o *Firstly*, fewer alternative modes of transportation normally exist at ferry crossing or their cost is substantially higher (because of connections to islands or between either side of fjords). This increases the relative importance of *being able to board* a certain vessel, compared to the on-board crowding cost.
  - o *Secondly*, many ferries carry cars in which congestion is naturally limited by the physical stature of vehicles, and congestion between passengers is smaller as they may wait in cars or enjoy any extra amenities (such as cafeteria or walking about). This makes on-board crowding, which has been studied in the literature in the case of buses and trains, *less relevant for ferry operations*. Consequently, optimal capacity is more dependent on the probability of being able to board the vessel than on-board crowding. This reduces the transferability of results from land-based public transport to the ferry sector.

Consequently, a research gap exists, which if solved, could enable policy makers to make better recommendations to decision makers fostering more efficient use of scarce resources.

The contribution of the paper to the scientific literature is as follows: We develop an optimization framework to assess the optimal capacity for ferry services in which frequency, vessel size and number is determined simultaneously. To the best of our knowledge, our paper is the first to apply such a complete model adapted to ferry services. Further, we model the optimal capacity using costs relevant to ferry services, as the current state of literature does not describe such cost in an appropriate way. It is also worth noting that several previous studies (Jara-Díaz and Gschwender, 2003; Jara-Díaz and Gschwender, 2009; Jara-Díaz et al., 2017; Jørgensen & Solvoll, 2018; Høyem & Odeck, 2020) have assumed that demand and prices are kept fixed. In this paper, we relax both of these assumptions in the context of car ferries. Last, the model is applied to three case studies to estimate the optimal composition of vessel capacity, number of vessels and frequency and prices. The case study approach will investigate if current service levels are optimal in three of the major crossings in Norway, providing an illustration of practical application of the model and its implications for policy.

Our results contribute by contrasting the optimal with current service levels, which is especially useful for local policy makers, but is also valuable for other policy makers elsewhere seeking to optimize service levels of their local ferry services. Consequently, our study first and foremost contributes as a case study of optimal capacity, which could be of interest to policy makers and the industry, using a relatively standard microeconomic optimization model.

The rest of the paper is organized as follows. Section 2 describes the optimization model used. Section 3 describes data used in some empirical example. Section 4 describes the scenarios that are investigated and the result of using the model on some empirical examples. Section 5 provides as discussion of results while section 6 concludes.

## 2. Methodology

We now present our formulation of the optimal capacity problem, using the model of Jansson et al. (2015) as our point of departure. It is modified into containing both the number of vessels and their size, while incorporating aspects relevant for ferry transportation.

The model of Jansson et al. (2015) uses a welfare economic framework. From an economic perspective, maximization of total welfare, consisting of producer and user surplus, is the goal when optimizing service levels at a ferry crossing. A private operator would focus solely on the costs and ticket income, whereas an economic planner would take into account the user's non-monetary costs (e.g., waiting time) as well (Jara-Díaz and Gschwender, 2009). As pointed out by Holmgren (2018), this is highly relevant for public transit services, as they often require a subsidy from the public sector. Consequently, we assume that the planner seeks to maximize total (economic) welfare. Our objective is to maximize the total welfare of operating a link with a given service and price level. Total welfare consists of two main components (i) consumer's surplus (*CS*) minus (ii) the net subsidy (equal to negative of the producer's surplus) required to provide a given service level (*NS*). We now define each component before presenting the complete problem. Together, these terms constitute the social surplus of operating a ferry service.

### 2.1. The consumer's surplus

We first define the user costs associated with running a ferry service, and then define the consumer's surplus on the basis of this.

Let $x$ be demand measured as PCE[4]/hour. The model assumes that a

[3] Defined as a composite of seats and standing area, dependent on which costs that are included.

[4] Capacity is measured in Passenger Car Equivalents (PCE), a multiproduct measure of production, used to normalize all vehicle sizes in the number of passenger cars. See Jørgensen & Solvoll (2018).

given number of vessels is operated in shifts of a specific length. Let $H$ be the set of all shifts, and let any shift be defined as a tuple of consecutive time periods, $t$, such that shift $i$ is $H_i = \{t_n, \cdots, t_m\}$. We assume there are two shifts; peak and off-peak. Further, define $V_i$ as the number of ferries operated in shift $i$ and $k$ the average capacity of each vessels measured in PCE. User costs, less the price, is defined as as the sum of three components: Open waiting time costs (*OW*), hidden waiting time costs (*HW*) (Andersen & Tørset, 2019) and excess waiting time costs (*EW*) (Høyem & Odeck, 2020):

$$\theta(x_t, V_i; k) = OW(V_i) + HW(V_i) + EW(x_t, V_i; k) \tag{1}$$

User costs is assumed to be a function of the number of vessels ($V$), average capacity of the vessels (PCE/vessel) and demand ($x$ PCE/hour). The components are discussed in more detail in the following sections[5].

The consumer's surplus is defined as the net difference in aggregate willingness to pay for the service, minus the generalized cost of using the service and ticket fare (Jansson et al., 2015). The aggregate willingness to pay is defined as the integral from 0 to $x$ of the marginal willingness to pay, $\bar{h}(x)_t$ , for using the service in each time period $t$ , summed over all periods $t \in [1, T]$ measured in hours. The consumer's surplus for a single day of operation is now defined as (see, e.g. Holmgren (2018) for a more detailed explanation):

$$CS = \sum_{t=1}^{T} \left( \int_0^x \bar{h}(x)_t dx - (\theta(x_t, V_i, k) + p_t)^* x_t \right) \tag{2}$$

The leftmost part inside the parentheses measures the aggregate willingness to pay. The rightmost part measures the total user costs, by multiplying costs per user ($\theta + p_t$) by the total number of users in each period ($x_t$)[6].

We now discuss the user costs components in greater detail.

#### 2.1.1. Open and hidden waiting time costs

Waiting time costs arise in transportation with scheduled services as the difference between the desired departure time of a user and the possible departures times of the scheduled service. In our analysis, we distinguish between so-called "open" and "hidden" waiting time (Høyem & Odeck, 2020; Andersen & Tørset, 2019; Hanssen et al., 2019; Jørgensen & Solvoll, 2018). *Open waiting time* is time spent waiting at the quay, while *hidden waiting time* is time spent elsewhere and represents and inconvenience in which the travelers are not being able to depart at their preferred time.

To attain a realistic estimate of the benefits of increased frequency, the cost of hidden ($HW(V)$) and open waiting time ($OW(V)$) is separated and valued independently by using the following equation:

$$OW(V) + HW(V) = \frac{\omega_H}{2f(V)} + (\omega_O - \omega_H) W_O(V) \tag{3}$$

All costs time components are scaled by their corresponding value of time, measured in NOK/hour (hidden: $\omega_H$ , open: $\omega_O$). The open waiting time is given as a function of number of vessels through frequency $f(V)$ in which it is assumed that $W_O(V)/\partial V < 0$ , that is, more departures (i.e., more vessels being used) decreases the open waiting time. Intuitively, increasing frequency should lower the safety margin one needs in arriving prior to a departure, which has also been observed empirically by Andersen & Tørset (2019) and Hanssen et al. (2019). The derivation of equation (3) is relegated to appendix A2, along with the form of $W_O(V)$ . Further, the equation of frequency given the number of vessels, $(V)$ , is defined in Section 2.2.4 below.

[5] If the model is applied to the bus sector, crowding costs could likely be substituted in for excess waiting time cost.

[6] For a discussion on the assumptions used when estimating a consumer's surplus see, e.g., Nyborg (2014), Varian (1992) or Slesnick (1998).

#### 2.1.2. Excess waiting time costs

Excess waiting time costs arise in the ferry sector when demand is greater than the capacity of a crossing (Høyem & Odeck, 2020). The cost of not being able to board the first, desired departure is equal to the duration of a complete headway. Let $\omega_E$ be the value of time for excess waiting time measured in NOK/hour. Costs associated with excess waiting time for direction $j$ is defined as:

$$EW(V; k) = \frac{\omega_E}{f(V)} \times \eta(x, V, k), \quad j \in IT \tag{4}$$

where $\eta(V, k)$ is the number of users left behind on average in each time period, which is a function of $V, k$ and demand; . The fractional part of the equation gives the monetary valuation of not boarding the first arriving vessel, assuming users must wait one full additional headway $(1/f(V))$. The derivation of $\eta(x, V, k)$ is relegated to appendix A3.

### 2.2. Net subsidy

We now discuss the net subsidy defined as the difference between operating costs and profit for each day of operation. We start by discussing costs and then moving on to the subsidy itself. The model assumes there are two time shifts: *peak* and *off peak*, with number of vessels operated given as $V_{OP}$ (off-peak) and $V_P$ (peak). Total costs ($C$) is comprised of two cost components: *Capital investment costs* ($\phi_I$) and *operating costs* ($\phi_O$) (see, e.g. Jara-Díaz et al. (2017)) for a similar cost composition):

$$C(V_{OP}, V_P, k) = \phi_I(V_P, k) + \phi_O^{OP}(V_{OP}, k) + \phi_O^P(V_P, k) \tag{5}$$

Capital investment cost is given as a function of the number of vessels employed in the peak period ($V_P$) and capacity of each vessel ($k$). As we assume a different number of vessels may be run during peak and off-peak, total operating cost is separated into peak ($\phi_O^P$) and off-peak ($\phi_O^{OP}$), being dependent on vessel number (through frequency) and capacity (as a larger vessels have higher running costs).

Define $p_t$ as the price per PCE in time period. The net subsidy is given by the total costs, less the income generated through all the time periods (, i.e., hours of the day):

$$NS = C(V_{OP}, V_P, k) - \sum_{t=1}^{T} p_t x_t \tag{6}$$

We now explain the different costs components in more detail.

#### 2.2.1. Operating costs

Operating cost for shift $i$ is defined as the sum of administrative and repair costs ($C^{AdminRepair}$), crew costs ($C^{Crew}$) and sailing costs ($C^{Sailing}$) all measured in NOK/day.

$$\phi_O^i(V_i, k) = C_i^{Crew} + C_i^{Sailing} + C_i^{AdminRepair} \tag{7}$$

Crew costs are defined as follows: Denote the number of sailors as $A$ [7]and the hourly wage per sailor as . All sailors in a specific shift $i$ work in a shift of length $s_i = |H_i|$ hours, where "||" denotes the number of time periods in shift $i$ [8]. Crew cost for shift $i$ then becomes:

$$C_i^{Crew} = V_i \times A \times w \times s_i \tag{8}$$

Sailing costs are defined as follows: Let $c_l(k)$ be the cost per sailed kilometers as a positive function of vessel size $k$ (larger vessels are more costly to run) and let $l$ be the length of the crossings (one round trip)

[7] In our analyses, we assume a fixed crew size for all vessels, which is motivated in relatively low variance in crew size for the vessels we consider

[8] More formally, this is known as the cardinality of the set $i$. We assume that the off-peak varies from 21:00–05:59, and peak from 06:00–20:59. Within each "shift", there may be changes to personnel, but the overall number of sailors stays constant.

measured in kilometers. The total sailing cost in shift $i$ is given by the cost of one round trip ($l\times c_l(k)$) multiplied by the total number of departures in shift :

$$C_i^{Sailing} = l\times c_l(k)\sum_{\forall t\in i} f_t(V_i) \tag{9}$$

Thus, the sailing cost is directly proportional to the frequency ($f$) and is also dependent on the size of each vessel, . Thus, running larger vessels and/or more often yields higher cost. To find the total cost for each day, the cost for all time periods are summed. The function $c_l(k)$ is described in appendix A4.

#### 2.2.2. *Administrative and maintenance costs*

Besides the direct costs related to operating the ferry services, there often exists both administrative costs (i.e. payroll, hiring, certifications, e.g.) and maintenance costs. We use a very simple methodology to estimate these costs, by scaling the cost per sailed kilometer by a factor $\alpha_{Sailing}$ , which is the share of operating costs (less crew costs) determined by sailing cost alone:

$$C_t^{AdimRepair} = C_i^{Sailing}\frac{\alpha_{Sailing}}{1-\alpha_{Sailing}} \tag{10}$$

We follow Svendsen et al. (2017) in defining the share $\alpha_{Sailing}$ as dependent upon the number of ferries at the crossing. Based on their methodology, we use the following equation to estimate $\alpha_{Sailing}$ , whose details of implementation are relegated to the appendix A4:

$$\alpha_{Sailing} = \min(a_1 + a_2{}^*V_{\text{Peak}}, \alpha_{Max}) \tag{11}$$

Here, $a_1$ is a constant, $a_2$ a slope coefficient interpreted as the change in the share of sailing costs when an extra ferry is added to the crossing, $V_{\text{Peak}}$ the total number of ferries at the crossing (equal to the peak) and $\alpha_{Max}$ the maximum value of $\alpha_{Sailing}$ . As the number of ferries increase, the administrative and maintenance cost's share decreases.

#### 2.2.3. *Capital investment costs*

Capital investment costs are calculated based on estimated investment cost which is given as a function of vessel size.

Let $I(k)$ be the total investment cost measured in NOK/vessel for a given vessel of size $k$ PCE. Yearly capital costs are found by adjusting for amortization by the interest rate ($r$) and vessel life-time ($n$) through the factor $a(r,n) = r/(1-e^{-r^*n})$, and multiplied by the number of vessels operated in the peak period ($V_{\text{Peak}}$). As all other components of the model, we adjust by dividing yearly cost by the number of operating days per year ($OD$).

$$\phi_I(V_2,k) = \frac{V_{Peak}\times I(k)}{OD}a(r,n) \tag{12}$$

Equation (12) now estimates the capital costs associated with operating $V_{Peak}$ ferries at an average capacity of . It does not have a time subscript, as the cost is measured per day and not per hour. Details on the investment cost function, $(k)$ , is given in appendix A5.

#### 2.2.4. *Frequency*

Frequency is determined indirectly from the number of vessels employed in each period. Observe that the number of vessels required to operate a frequency of $f$ given a total round-trip time of $t$ equals:

$$V = t^*f \tag{13}$$

Total time per round-trip ($t$) is given by the length of the link (one roundtrip) in kilometers ($l$), speed of the vessels in kilometers per hour ($s$), time spent in mooring at the two quays ($m$) and time used to load and unload a PCE ($q$) measured in hours, and demand and frequency per hour ($x_t$ and $f$).

$$t = \frac{l}{s} + m + q\frac{x_t}{f} \tag{14}$$

For each time period, frequency is defined as the maximum permissible frequency given the number of vessels employed is the found by inserting (14) into (13) and solving for :

$$f_t(V_i) = \frac{(V_i-(q^*x_t))}{(l/s+m)} \tag{15}$$

Assuming that (14) holds in all shifts, we use (15) to estimate the frequency employed in each time period $t$ given by the number of vessels in the shift which time period $t$ belongs to; $V_i$ .

### 2.3. *The optimal capacity problem*

We now move on to defining the objective functions and the conditions for the welfare optimum. Whether or not to include a capacity requirement is of special interest when formulating the optimization problem. We thus start by briefly discussing capacity requirements in optimization before moving onto the definitions itself.

The literature on optimization of public transport services have used different approaches to capacity requirements. Jansson (1980), Jara-Díaz and Gschwender (2003), Jara-Díaz and Gschwender (2009), Jara-Díaz et al. (2017), Jara-Díaz et al. (2020), Pedersen (2003), Tirachini et al. (2014), Tirachini and Antoniou (2020), have used some form of capacity constraint in which demand should be covered up to an exogenously given factor. However, in the case of bus services Börjesson et al. (2017), Asplund and Pyddoke (2020) did not assume any constraint, but rather included crowding costs.

Consequently, the literature has presented different principles with regard to including capacity constraints as a part of the optimization problem. We will demonstrate that the case without a capacity constraint is a special case of the one where a constraint is enforced. More specifically, we will show that the unconstrained optimum is the same as the constrained optimum, where the socially optimal capacity is set into the constraint. Another relevant aspect is the long- versus short run marginal cost. In this framework, the long-run marginal cost comes into play through the investment cost and operational cost functions. However, in the first-best case (without a capacity requirement) the long-run marginal cost does not enter the price equation directly. This is commented upon in Section 2.3.3 below.

We start by deriving optimality conditions for the first case, then the second. Last, we show how the two formulations are related to one another and under what conditions they are equal. The solution to the equation system using several variables is the method proposed in this study.

#### 2.3.1. *Without a capacity requirement*

We now present the complete optimization problem and derive conditions for the welfare optimum without a capacity requirement. Social surplus ($SS_1$) equals the difference between the consumer's surplus ($CS$) and net subsidy ($NS$)[9]:

$$SS_1 = CS - NS \tag{16}$$

Inserting for $CS$ and $NS$ into (16) and rearranging terms, one obtains the following complete expression for social surplus:

$$SS_1 = \sum_{t=1}^{T}\left(\int_0^x \bar{h}(x)_t dx - \theta(x_t, V_i, k)^* x_t\right) - C(V_{OP}, V_P, k) \tag{17}$$

The planner maximizes this problem with respect to prices $= \{p_1, \cdots, p_T\}$ , capacity, $k$ , and vessels used in peak and off-peak $V_{OP}\&V_P$ . To find

[9] Observe that defining social surplus as CS – NS is the same as defining it as CS + PS where PS is the producer's surplus as NS = - PS. The expression could also have included the marginal cost of public funds, which measures the total welfare loss per NOK spent over public budgets. However, for the sake of brevity, we follow Börjesson et. al. (2017) and do not include it.

the conditions for the optimum, we take the derivative of $V_i$ for all shifts , $k$ and $x_t$ for all time periods $t$ . We then obtain a system of equations characterising the optimum. The derivative of social surplus with respect to the number of vessels employed equals[10]:

$$\frac{\partial SS_1}{\partial V_i} = \frac{\partial \phi_O^i}{\partial V_i} + \sum_{t \in i} \frac{\partial \theta_t}{\partial V_i} * x_t = 0, \forall i \quad (18)$$

Equation (18) states that the marginal change in operating costs from employing more vessels in shift $i$ ($\frac{\partial \phi_O^i}{\partial V_i}$), should equal the marginal change to total user costs for all time periods contained within shift $i$ (sum of $\frac{\partial \theta_t}{\partial V_i} * x_t$). The derivative of social surplus with respect to the capacity employed equals:

$$\frac{\partial SS_1}{\partial k} = \frac{\partial \phi_I}{\partial k} + \frac{\partial \phi_O^P}{\partial k} + \frac{\partial \phi_O^{OP}}{\partial k} + \sum_t \frac{\partial \theta_t}{\partial k} * x_t = 0 \quad (19)$$

Equation (19) states that the marginal change in capital ($\frac{\partial \phi_I}{\partial k}$) and operating costs ($\frac{\partial \phi_O^P}{\partial k} + \frac{\partial \phi_O^{OP}}{\partial k}$) from employing slightly larger vessels, should equal the marginal change to total user costs for all time periods ($\sum_t \frac{\partial \theta_t}{\partial k} * x_t$) when employing slightly larger vessels. The derivative of social surplus with respect to demand in each period equals:

$$\frac{\partial SS_1}{\partial x} = p_t^* - x_t \frac{\partial \theta}{\partial x_t} = 0, \forall t \quad (20)$$

Equation (20) states that optimal price per hour is equal to the marginal change to total user costs for the given time period ($\frac{\partial \theta}{\partial x_t} x_t$). The marginal cost rise as another user enters the system, which will lower the overall probability of being able to board with fixed capacity[11]. Our price expression is very similar to the one derived in Börjesson et al. (2017) for bus operations. Jørgensen et al. (2004) obtained a different expression by applying the principle of Ramsey prices (Ramsey, 1927) for ferry services. However, such prices are relevant in second-best situations only, where a certain amount of revenue must be generated (i.e. to limit subsidization needs). However, we concern ourselves with the first-best case, leaving the second-best with a financial constraints as a possible extension of the model given local preferences for subsidization levels. In any case, we contend the first-best serves as a useful benchmark.

All equations used are linear, expect for $\theta(x_t, V_i; k)$ which is convex, meaning the second-order conditions for a minimum are satisfied. The functions $\phi_O^P$ and $\phi_O^{OP}$ are also checked numerically to certify that they are indeed linear.

### 2.3.2. With a capacity requirement

The case with a capacity requirement is found by adding a constraint to the optimization problem. Let $LB_t(x, V, k)$ be the percentage of users not being able to board on their first attempt ("left behind") in period $t$ , and $LB_{Max}$ be the maximum percentage that is allowed to be left behind (during a whole day). Equation (21) display the objective function to be maximized in which a constraint with Lagrange multiplier $\rho$ is subtracted from the social surplus without any capacity requirement, $S_1$ :

$$SS_2 = SS_1 - \rho \left( \sum_{t=1}^{T} LB(x, V, k) - LB_{Max} \right) \quad (21)$$

[10] If $= Peak$ , the term $\frac{\partial \phi_I}{\partial V_i}$ is added to equation (18), which is the marginal effect on investment cost form adding another ferry.

[11] Some caveats are important to mention, as the externality likely depends on the queue type, i.e. a First-In-First-Out), or an "unordered" queue. Høyem & Odeck (2020) assessed the relevance of such considerations at the aggregate level used in this study, and found that queue type did not significantly affect the results.

We now obtain a new set of equations characterising the optimum[12]:

$$\frac{\partial SS_2}{\partial V_i} = \frac{\partial \phi_O^i}{\partial V_i} + \sum_{t \in i} \left( \frac{\partial \theta_t}{\partial V_i} * x_t - \rho \frac{\partial LB}{\partial V_i} \right) = 0, \forall i \quad (22)$$

In the vessel number equation (equation (22)) the term $\rho \frac{\partial LB}{\partial V_i}$ is now added. $\frac{\partial LB}{\partial V_i}$ is the change in the number of users left behind from increasing the number of vessels used. A larger number of vessels will decrease the number of users left behind ($\frac{\partial LB}{\partial V_i} < 0$). Thus, if $\rho > 0$ , it indicates that a smaller number of ferries should be used as $\theta_t$ is a convex function of $V_t$, as compared to the case without a capacity requirement – all else equal.

$$\frac{\partial SS_2}{\partial k} = \frac{\partial \phi_I}{\partial k} + \frac{\partial \phi_O^P}{\partial k} + \frac{\partial \phi_O^{OP}}{\partial k} + \sum_t \left( \frac{\partial \theta_t}{\partial k} * x_t - \rho \frac{\partial LB}{\partial k} \right) = 0 \quad (23)$$

In the vessel size equation (equation (23)), the term $\rho \frac{\partial LB}{\partial k}$ is now added. $\frac{\partial LB}{\partial k}$ is the change in the number of users left behind from increasing the capacity of the vessels used. Larger vessels will decrease the number of users left behind ($\frac{\partial LB}{\partial k} < 0$). If $\rho > 0$ , it indicates that smaller vessels should be used as $\theta_t$ is a convex function of $k$, as compared to the case without a capacity requirement – all else equal.

$$\frac{\partial SS_2}{\partial x_t} = p_t^* - x_t \frac{\partial \theta}{\partial x_t} - \rho \frac{\partial LB}{\partial x_t} = 0, \forall t \quad (24)$$

In the price equation (equation (24)) the term $\rho \frac{\partial LB}{\partial x_t}$ is now added. $\frac{\partial LB}{\partial x_t}$ is the change in the number of users left behind from increasing the demand. A higher demand will increase the number of users left behind ($\frac{\partial LB}{\partial x} > 0$). If $\rho > 0$ , it indicates that a higher price should be charged , as compared to the case without a capacity requirement – all else equal. Thus, prices with a capacity constraint are higher as compared to the first-best case if $\rho > 0$ (too much capacity is supplied, which is discussed below).

Last, the derivative of the Lagrange multiplier implies that:

$$\frac{\partial SS_2}{\partial \rho} = \sum_{t=1}^{T} LB(x, V, k)_t - LB_{Max} = 0 \quad (25)$$

### 2.3.3. Comparing the two formulations

The essential component when comparing the two different formulations is the Lagrange multiplier.

Thus, we start by briefly discussing the Lagrange multiplier associated with the capacity requirement.

The multiplier has a direct interpretation: As $\frac{\partial SS}{\partial LB_{Max}} = \rho$ , we know that an increase in the maximum number of users allowed to be left behind, will increase total welfare by $\rho$. In other words, if $\rho > 0$ then total welfare may be increased by allowing more users to be left behind. If $\rho < 0$ , then total welfare may be by increased by allowing fewer users to be left behind. I.e., by estimating the sign of $\rho$, we may directly infer whether or not too much or too little capacity is provided.

Consequently, the formulation without a capacity constraint, is simply a special case of the formulation with a capacity constraint, in which the optimal capacity requirement (operationalized by the share of users left behind) is enforced. That is, solving the problem without a capacity constraint yields an optimal share of users left behind $B^*$ . If $LB_{Max} = LB^*$ when solving the problem with a capacity constraint, then $\rho = 0$ at the optimum, and the two problems yield the same solution. However, if $B_{Max} \neq LB^*$ , the planner enforces a level of capacity that is not optimal.

The reader should now be able to recognize that the formulation with a capacity constraint is indeed the second-best formulation concerning

[12] See footnote in equation (18) for details regarding the derivative in the peak period with respect to investment cost.


**Fig. 1.** Map of the different crossings in our analysis.

the problem of deriving optimal capacity at a ferry crossing, in which the formulation without a constraint is the first best. In the second-best formulation, constraints are introduced such that the pareto-conditions cannot be attained (Lipsey & Lancaster, 1956). The constraint concerning the number of users left behind is the constraint in question in our particular case. Further, a larger, the higher the level of "excess" capacity is being provided.

Moreover, in the first-best case, the only pricing relevant cost is the marginal, external cost conferred upon other users, plus the marginal cost. When a capacity constraint is included, one effectively assumes the planner "guarantees" a certain quality of the service. For example, that users will have a journey that guarantees a probability of being left behind no larger than a given value. Naturally, the user needs to pay for this guarantee somehow, which is reflected in model where a capacity constraint is included. Thus, $\rho \frac{\partial LB}{\partial x_t}$ gives the price of this guarantee, which will reflect the long-run cost of increasing capacity. However, when no quality of service is guaranteed, the user is not obliged to pay for a certain level of quality.

The user instead now "pays" through two different channels:

- **Directly**: The price levied that reflects the external costs they impose on the operator and other users $p^*$ .
- **Indirectly**: Crowding disutility included in the generalized cost will increase, when demand increases. With a service quality guarantee (capacity constraint), one pays directly for this guarantee as increased demand yields a higher capacity directly. Without a service quality guarantee (capacity constraint), one pays indirectly through increased generalized costs.

It is important to keep this distinction in mind when interpreting the results of the first-best model solution.

### *2.3.4. Estimating the welfare optimum*

Eqs. ((18)–(20)) together define a system of non-linear equations which enables solving for the optimal prices, vessels and capacity for a given level of demand. The model is solved numerically by iteratively estimating the optimal value of decision variables and adjusting demand until convergence to the equilibrium. See appendix A6 for details.

## 3. Case studies and data

We now present the case studies that the model is applied to along with the data gathered and parameter values chosen.

### *3.1. Crossings*

The case studies are based on different car ferry crossings in Norway. Out of approximately 130 active crossings, we chose 3 of the largest ones in terms of PCE transported who are situated along the E39 highway. E39 is the main trunk road and transport corridor in the western part of Norway. The ferry crossings constitute an integrated part of E39, underlining their importance. Fig. 1 shows the geographical location of the crossings in our study. The black line corresponds to E39, while the blue ones are other major trunk roads in Norway. The crossing's positions are marked by a red circle, and major cities marked by a triangle.

Table 1 includes key information on relevant variables for the crossings in our case study. The average length of each crossing is 14.6 km per roundtrip, with a speed of 28.3 km, 56.3 departures per day, 4.3 ferries 178.1 PCE and 1.1 berths per quay. There were no crossings serving more than two quays ("triangle") in our data. All crossings are situated in the south western part of Norway, where the major part of ferry related activities take place. Consequently, our results will mainly

**Table 1**
Information on crossings.

| Crossing | Length [km/round trip] | Speed [km/h] | Departures/day | Ferries | Average ferry size (PCE) | Number of berths |
|---|---|---|---|---|---|---|
| Halhjem-Sandvikvåg | 43.4 | 33 | 54 | 5 | 180 | 1 |
| Stavanger – Mortavika | 18.5 | 25 | 70 | 4 | 226 | 2 |
| Molde – Vestnes | 23.0 | 20 | 45 | 4 | 128 | 1 |
| Average | 28.3 | 25.8 | 56.3 | 4.3 | 178.1 | 1.3 |

**Table 2**
Value of time, persons per vehicle and share of each trip purpose. 2019-NOK

| Trip purpose | Share (%) | Persons per vehicle | VOT/hour (2019-NOK) |
|---|---|---|---|
| Business | 4 % | 1.15 | 480 |
| Work | 21 % | 1.1 | 107 |
| Leisure | 75 % | 1.9 | 91 |

remain valid for this part of the country's ferry sector. Further, the selected crossings are all among the larger ones in Norway, with respect to demand and capacity. Consequently, it is especially interesting to investigate if their capacity level is an optimal one.

Information on the length of each crossing was gathered from the NPRA's[13] databank on ferries ("ferjedatabanken"), as was the average speed of the vessels. The number of departures per day was gathered from the companies' posted timetables, as was in general the number of ferries and their capacities[14]. The capacity of two ferries were gathered by contacting the companies directly. The number of berths was found by inspecting satellite images of the quays. Data on ticket prices were gathered from the companies' websites, in which we used the price for 1 PCE as a representative price.

Data on demand was supplied by the NPRA for 2019. We use average demand per weekday, excluding weekends (when traffic and service levels is lower). The dataset covers number of vehicles in different ticketing categories for each hour within a whole year. All vehicle counts were transformed into PCE units using a conversion table between ticketing classes and PCE supplied by the NPRA's ferry division. There are two different systems for data collection. In the "Autopass" system, cars are registered electronically by a chip mounted in the front mirror. In the "Riksregulativ" system, motorists pay to a ticketer, either on shore or at sea. If demand is high and the crossing is short, the ticketer might not be able to register all motorists. However, we suspect this is a minor issue in our data. Current prices were set equal to the price of the smallest vehicle class (equal to 1.025 PCE), as an approximation to the average price per PCE. Separate prices exists for electric and fossil fueled vehicles, where an assumption of 9.2 % electric cars is used (Statistics Norway, 2021). Due to lack of data, other discounts, such as a loyalty program were not included in the price calculation[15].

### 3.2. Parameter values and value of time

The parameter values chosen for our application of the model is given in table A1. Our model contains many parameters, and the complete set is displayed in appendix 1 together with symbols and in which equation they appear first.

We explain how the value of time is estimated in this section, as it is perhaps the most important parameter. For all other parameters, please see table A1 in appendix 1 for details. Table 2 shows the value of in vehicle time for car users (2019-NOK), share of trips and persons per vehicle gathered from NPRA (2018)[16]. To obtain an average value per vehicle, a weighted average of trip purpose share, persons per vehicle is taken of VOT for each trip purpose. The average value of time for a passenger car (assumed to be 1 PCE) is 176 2019-NOK/Hour[17]. Based on NPRA (2018), the value of time for a heavy truck is 723 2019-NOK[18]. According to Jørgensen & Solvoll (2018), heavy trucks constitute around 15 % of total traffic on average. To convert the value of time per vehicle to value of time per PCE, the value of time for heavy trucks is divided by 6.17, which is the average PCE per large vehicle (longer than 5.6 m based on our dataset containing demand data). The average value of (in vehicle) time, is then 167 2019-NOK.

The value of open waiting time (at the quay) is estimated by multiplying the value of in vehicle time by 1.3, which is based on data from NPRA (2018) and estimated by (Høyem & Odeck, 2020) as an average over different headways. The value of hidden waiting time (not at the quay) is defined as 50 % of open waiting time, in line with Jørgensen & Solvoll (2018). Last, the value of excess waiting time is assumed to be 1.7, based on Flügel et al. (2018). The value of open waiting time is then 218 NOK/hour, hidden waiting time; 109 NOK/hour and excess waiting time; 285 NOK/hour. It is important to underline that these are average values, that may differ between crossings due to share of heavy vehicles, purposes and local factors such as income, etc. Consequently, our calculations will highlight tendencies, rather than exact, local levels.

The model was coded in the Python programming language, where the equation system is solved using the "fsolve" function available in Python Scipy (Virtanen et al., 2020).

## 4. Results

### 4.1. Investigated scenarios

The model is applied to the case study crossings with the parameter values and assumptions as detailed in the preceding sections. First, a base scenario is run. Then, a number of scenarios where different assumptions being altered are run. To assess the scientific validity, the model's results should be subjected to sensitivity tests, to gauge its robustness to alternative assumptions. The number of users being left behind is of major importance when estimating optimal capacity, to which we devote special attention. Thus, we also test the sensitivity of our result, if the number of users left behind is being underestimated in the base scenario.

In total, we run four different scenarios:

- **Base scenario:** The base scenario is simply running the model with the parameter values as declared in the text. Consequently, this scenario reflects the model's result using our best estimates of current costs and benefits relevant to optimizing service levels at a ferry crossing.
- **Electric propellant:** There are three primary propellants used in ferries: Diesel, Electricity and natural gas (LNG). The NPRA uses the same cost per kilometre for Diesel and LNG ferries in cost benefit analyses (NPRA, 2015). As many Norwegian ferry services have converted to electricity as the main propellant, it is interesting to consider how this influences the optimal service levels. DNV GL (2015) suggests electric ferries have higher capital requirements, and around 50 % lower propellant cost per kilometre. We use a 50 % lower cost per sailed kilometer in this scenario.
- **Increased value of excess waiting time (+100%):** We increase the value of excess waiting time by 100 % to assess how sensitive the model's results is to this assumption. Currently, the value of excess waiting time is relatively low according to the official Norwegian estimates, only 30 % larger than open waiting time. If users have a larger willingness to pay for not being left behind, this will affect our results.
- **Increased number of users that are left behind and increased value of excess waiting time:** We use a simplified methodology to estimate the number of users left behind. It is based on the highly

[13] Norwegian Public Roads Administration.

[14] Demand data was available for 2019 while timetable data for 2020. However, this small discrepancy is not likely to affect the results in a significant way, given the time difference is only one year.

[15] Under this program, one pays an amount up front and receives a discount up to 50 %. However, we do not have data on the extent to which this program is used.

[16] The values are an overall average over all trip lengths as given in NPRA (2018) and adjusted from 2016 NOK by the Norwegian Consumer Price Index to 2019-NOK, and a weighted average of each trip purpose and persons per vehicle given in table 2.

[17] Adjusted by the Norwegian Consumer Price Index from 2016 to 2019 NOK.

[18] Adjusted by the Norwegian Consumer Price Index from 2016 to 2019 NOK.

**Table 3**
Results for the different scenarios by each crossing examined.

| **HALHJEM-SANDVIKVÅG** | | | | | |
|---|---|---|---|---|---|
| Variable | **Current** | **Base solution** | **Electric propellant** | **Value of (excess) time+ 100 %** | **Users left behind & Value of time+ 100 %** |
| Price S1 | 299 | 33 | 16 | 33 | 62 |
| Price S2 | 299 | 43 | 29 | 46 | 77 |
| Frequency S1 | 1.0 | 1.3 | 1.5 | 1.2 | 1.3 |
| Frequency S2 | 2.9 | 3.3 | 3.7 | 3.3 | 3.5 |
| Capacity | 180 | 71 | 75 | 85 | 121 |
| Demand S1 | 625 | 1,069 | 1,189 | 1,067 | 983 |
| Demand S2 | 5,087 | 7,085 | 7,630 | 7,100 | 6,479 |
| Total capacity | 19,080 | 8,733 | 10,386 | 10,188 | 15,741 |
| Total demand | 5,713 | 8,154 | 8,819 | 8,167 | 7,462 |
| ΔUser benefit (NOK/day) | | 17,79,339 | 20,37,233 | 17,63,917 | 14,69,766 |
| ΔOperative cost (NOK/day) | | -2,57,234 | -43,944 | -2,22,800 | -32,679 |
| ΔInvestment cost (NOK/day) | | -97,230 | -82,278 | -85,384 | -41,607 |
| ΔPublic revenue (NOK/day) | | -13,70,305 | -14,66,781 | -13,44,138 | -11,45,104 |
| Capacity utilization (%) | 30% | 93% | 85% | 80% | 47% |
| Left behind (%) | 0% | 9% | 5% | 3% | 10% |
| **MORTAVIKA-ARSVÅGEN** | | | | | |
| Variable | **Current** | **Base solution** | **Electric propellant** | **Value of (excess) time+ 100 %** | **Users left behind & Value of time+ 100 %** |
| Price S1 | 245 | 22 | 11 | 22 | 41 |
| Price S2 | 245 | 28 | 19 | 30 | 49 |
| Frequency S1 | 1.7 | 2.3 | 2.7 | 2.2 | 2.3 |
| Frequency S2 | 3.8 | 5.3 | 5.8 | 5.2 | 5.6 |
| Capacity | 226 | 67 | 72 | 80 | 115 |
| Demand S1 | 918 | 1,698 | 1,863 | 1,702 | 1,575 |
| Demand S2 | 7,178 | 10,711 | 11,515 | 10,792 | 9,783 |
| Total capacity | 31,640 | 13,551 | 16,157 | 15,763 | 24,413 |
| Total demand | 8,096 | 12,409 | 13,378 | 12,495 | 11,358 |
| ΔUser benefit (NOK/day) | | 23,29,041 | 26,05,119 | 23,31,561 | 20,47,771 |
| ΔOperative cost (NOK/day) | | -1,43,082 | 26,369 | -1,10,739 | 38,531 |
| ΔInvestment cost (NOK/day) | | -96,789 | -83,264 | -85,442 | -44,703 |
| ΔPublic revenue (NOK/day) | | -16,48,316 | -17,39,542 | -16,22,144 | -14,40,031 |
| Capacity utilization (%) | 26% | 92% | 83% | 79% | 47% |
| Left behind (%) | 0% | 9% | 5% | 3% | 9% |
| **MOLDE-VESTNES** | | | | | |
| Variable | **Current** | **Base solution** | **Electric propellant** | **Value of (excess) time+ 100 %** | **Users left behind & Value of time+ 100 %** |
| Price S1 | 150 | 34 | 21 | 39 | 59 |
| Price S2 | 150 | 38 | 26 | 41 | 69 |
| Frequency S1 | 1.4 | 1.3 | 1.5 | 1.3 | 1.3 |
| Frequency S2 | 2.8 | 3.2 | 3.6 | 3.1 | 3.5 |
| Capacity | 128 | 64 | 69 | 78 | 110 |
| Demand S1 | 529 | 860 | 940 | 858 | 817 |
| Demand S2 | 4,588 | 5,850 | 6,337 | 5,862 | 5,618 |
| Total capacity | 13,056 | 7,684 | 9,322 | 9,157 | 14,313 |
| Total demand | 5,117 | 6,709 | 7,278 | 6,719 | 6,434 |
| ΔUser benefit (NOK/day) | | 6,80,752 | 8,65,133 | 7,01,459 | 7,48,091 |
| ΔOperative cost (NOK/day) | | -1,61,458 | -17,072 | -1,29,997 | 37,207 |
| ΔInvestment cost (NOK/day) | | -39,893 | -27,303 | -30,126 | 7,667 |
| ΔPublic revenue (NOK/day) | | -5,15,185 | -5,85,066 | -4,94,553 | -3,32,915 |
| Capacity utilization (%) | 39% | 87% | 78% | 73% | 45% |
| Left behind (%) | 0% | 9% | 5% | 3% | 9% |

popular «newsvendor» model from the supply chain literature and adapted to public transport by Høyem & Odeck (2020). Consequently, it is important to assess what happens if the number of users left behind is underestimated with respect to service levels, in particular the ferry size. We then assume that real capacity is 50 % of nominal when estimating the number of users being left behind. That is, 1 000 PCE per hour is regarded as only 500 PCE under this scenario. In addition, we increase the value of excess waiting time, such that users have a 100 % higher willingness to pay to avoid being left behind. The purpose is to tests the sensitivity if both the probability and consequence of being left behind is increased markedly.

We compare the different service level variables against each other, along with the ones currently in operation.

### 4.2. *Scenario results*

We now present the results of our model runs. Table 3 shows the estimated optimal levels for the three crossings. The current values of the decision variables are also displayed at the leftmost part of the table. Last, we examine how the optimal level of capacity compares to current ones. Changes in user benefits and costs are compared to the current levels.

#### 4.2.1. *Base scenario*

The base scenario as compared to the current one, generally shows the same pattern across all our crossings. Our model suggests there is too much capacity offered at all three. Currently, there is an estimated capacity utilization of 26 – 39%. Capacity is the composite of frequency and vessels size. The estimated optimal frequency (as derived by

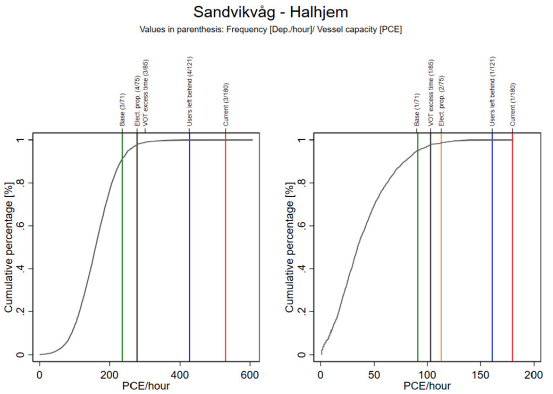



**Fig. 2.** The cumulative distribution function of demand [PCE/hour] and total capacity (Vessel capacity * frequency) in peak (left) and off-peak (right) for the different scenarios. Values in parenthesis indicate frequency and vessel size, respectively, rounded to the nearest integer. Halhjem-Sandvikvåg.

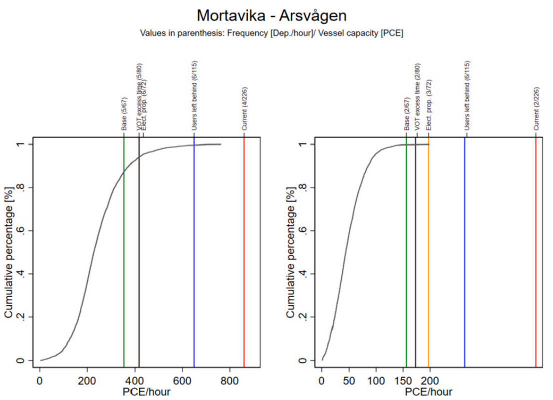



**Fig. 3.** The cumulative distribution function of demand [PCE/hour] and total capacity (Vessel capacity * frequency) in peak (left) and off-peak (right) for the different scenarios. Values in parenthesis indicate frequency and vessel size, respectively, rounded to the nearest integer. Mortavika-Arsvågen.

equation (15) through the number of vessels), is somewhat higher in the base scenario, suggesting more weight should be put on reducing the users open and hidden waiting time costs, as these costs accrue to *all* users, not just the ones being left behind. Consequently, to lower total capacity, smaller ferries should be used, as gauged from an economic perspective, where the user's costs are compared to the social costs of providing a given service level. The reduced vessel size yields a higher number of users being left behind at 9 %. However, the reduction in capacity is quite large. If our model underestimates the true number of users left behind, service levels will also be biased. Optimal capacity utilization is close to 100 %, which is likely too high. It is therefore important to assess how sensitive this result is with respect to the number of users being left behind, which is done below (Section 4.2.4).

Our model also suggests that the prices currently levied are too high. The calculations are solely based on the externality that new users confer upon existing ones when entering the queue, and no financial constraint is included (i.e., the first best). That is, more users will tend to increase the probability that some users are being left behind. As demonstrated by the lower optimal capacity level, current valuations of the user's willingness to pay is not large enough to sustain current

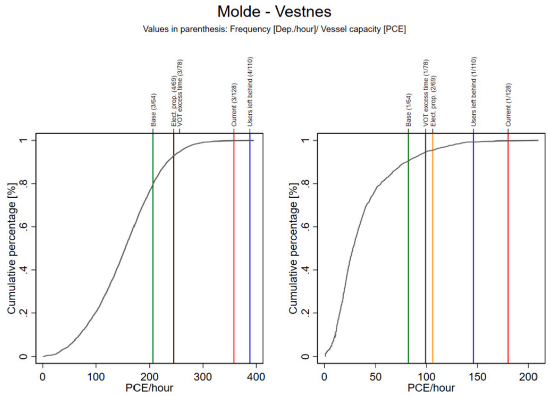



**Fig. 4.** The cumulative distribution function of demand [PCE/hour] and total capacity (Vessel capacity * frequency) in peak (left) and off-peak (right) for the different scenarios. Values in parenthesis indicate frequency and vessel size, respectively, rounded to the nearest integer. Molde-Vestnes.

capacity levels. Consequently, the externality conferred upon other users from a new one entering the system, is also expected to be relatively low, which explains why optimal prices are lower (see equation (20)). Moreover, there is no requirement on the level of subsidisation in our model (i.e. how much of the cost that should be covered by public expenditure). Including such a constraint would most likely increase prices markedly[19].

Employing the optimal service levels would lead to an increase in demand, mainly due to lowered prices. Peak period demand would increase by ca. 30–50 %, while off-peak demand would increase by ca. 60–85 %. Consequently, capacity utilization should optimally be increased by allowing a substantially higher demand in the off-peak period in which there is currently excess capacity. As stated above, this relies on number of users left behind, which needs to be tested further.

User benefits[20] are estimated to increase by 700 000 – 2 300 000 million NOK/weekday, whereas operational and capital costs are projected to fall. The reduction in prices explains the increase in user benefits, even though service levels are lower. Thus, the findings indicate that too much capacity and too high prices are levied, in comparison to the users' costs associated with having to sit back.

### *4.2.2. Electric propellant (lower operating cost)*

Under the electric propellant scenario, the operating cost related to fuel is halved, impacting the direct cost of maintaining a given frequency. Both frequency and vessels size increases as it is now cheaper to maintain a higher service level. However, a 50 % reduction in fuel costs does not correspond to a 100 % increase in frequency, but rather 10 % (compared to the base scenario). Labour constitute the major part of operating costs. Moreover, larger vessels are being used, which will partly offset the effect of lower fuel costs.

### *4.2.3. Value of excess waiting time*

Increasing the value of excess waiting time by 100 % has a relatively minor effect on the optimal frequency. Prices are slightly higher, as the externality by one more user entering the system is higher as excess waiting time makes the cost of being left behind greater. A larger effect is

[19] Under the current system, prices are set to meet a certain budgetary criterion, according to Jørgensen et. al. (2004), whereas ours are not, partly explaining the difference between current and optimal ones.

[20] User benefits are estimated using the "rule-of-half",

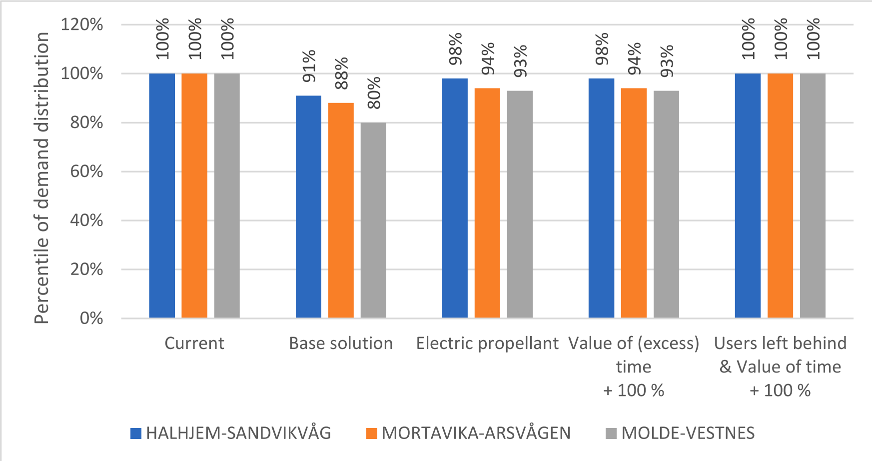



**Fig. 5.** Percentiles of the demand distribution at different capacity levels.

seen on the vessel size, which increases by around 20 % compared to the base scenario, lowering capacity utilization to about 80 %. This in turn, leads to a lower number of users being left behind.

### *4.2.4. Increased number of users left behind and increased value of excess waiting time*

As stated above, the number of users being left behind is a crucial component of the model. Using the base parameters, we observe very high capacity utilization levels. Our model operates on three assumptions that may underestimate the number of users being left behind.

First, we use yearly averages of demand, whereas demand typically fluctuates, being substantially higher during the summer for some crossings in Norway. Thus, by using a yearly average, the demand peaks are "smoothed", which may lead to a lower number of users being left behind. Secondly, demand is also averaged across each hour. However, if peak demand is very limited to a short time period within each hour, our method will underestimate the true number of users being left behind as queuing is "smoothed" down as well. Third, our model implicitly assumes all users are able to board a vessel within one hour (there is no excess demand carried forward to the next time period). In reality, queues may form during peak periods, such that this assumption is violated.

Consequently, it seems prudent to perform a sensitivity analysis in light of the marked difference between optimal and current capacity. We run a scenario effectively assuming that the true capacity is halved when estimating the number of users being left behind. Thus, a capacity of 1 000 PCE/hour, will now be regarded as 500 PCE/hour when using equation (A7) to estimate the number of users being left behind. Moreover, we carry forward the assumption that excess waiting time cost is doubled. Thus, this scenario will test the sensitivity of our results if both the probability and severity of being left behind is doubled, to see if the results from the base scenario changes.

Frequency and vessels size are both increased as compared to the base case. Firstly, a higher frequency will increase capacity such that fewer users are being left behind (i.e., affect the probability). Secondly, a higher frequency will reduce the excess waiting time for those users who must sit back. Vessel size will only affect user cost through the number of users left behind (i.e. the "capacity" channel) but has no direct effect on waiting time for those who are left behind. Consequently, frequency may be increased less percentage-wise to attain the same effect on excess waiting time as vessel size. However, increasing frequency is more costly, as labour costs are increased on a one-to-one basis. Increasing the size of ferries to raise capacity is cheaper. The net effect is that vessels size is increased by a larger proportion as compared to frequency, relative to the base solution.

Observe that compared to the scenario with only the value of excess time increased, capacity utilization goes down, but the number of users left behind goes up. This may seem contradictory, but it is not. Increasing the number of users left behind by reducing nominal capacity by half, means that for the same level of capacity, the number of users left behind will increase. At the same time, it also becomes more costly on the margin to reduce the number of users left behind. As a result, it is possible that the optimal number of users left behind is higher, even though the capacity utilization is lower. As each unit of capacity utilization corresponds to a higher number of users left behind, when capacity is 50 % of nominal.

Further, prices increase by a factor of 80 – 70 %, as the externality from more users entering the system increases when the nominal capacity is diminished. At the same time, capacity utilization drops significantly, to about 50 %. The total capacity is increased by around 80 %, compared to the base scenario. However, even with a high cost and number of users being left behind, the current service levels generally are still estimated to be too high as compared to the optimum, except for Molde-Vestnes in the peak period. A part of the explanation is that although total capacity is lower, optimal frequency is 20–50 % higher during the peak period. Thus, the actual waiting time costs are reduced, which is a gain for *all users*, not only those who are left behind. Even though there is a larger number of users having to "sit back", their excess waiting time when left behind is also smaller as compared to the base scenario.

User benefits fall as compared to the base case for Mortavika-Arsvågen and Halhjem-Sandvikvåg, except for Molde-Vestnes. Operational costs have now increased (as compared to the base scenario).

The sensitivity test illustrates that even if we assume a substantially higher cost associated with capacity shortages, it is still optimal to run smaller vessels, but run them more often. Doing so will lower the user cost of all users and lessen the waiting time experienced when having to sit back a departure.

### *4.2.5. Comparing capacity with the current demand distributions*

Our findings can be regarded as relatively surprising, as there is a sizable drop in the optimal capacity. It warrants a closer inspection to explain why they are achieved. Consequently, we have examined the demand distribution at all crossings compared to the capacity levels per hour under the different scenarios.

Figs. 2–4 shows a cumulative percentage plot over demand for each hour of operation in 2019[21]. The scenarios are depicted as follows:

- Current capacity is the red line
- The base scenario is the green line
- The scenario with electric propellant is orange
- The scenario with an increase value of excess waiting time is **black**
- The scenario with a higher number of users left behind and excess waiting time valuation is the blue line (the scenario discussed in Section 4.2.4).

It is evident that demand levels are below current capacity for a majority of time the ferries are in operation. Demand exceeds capacity in only a small fraction of the hours for which we have data. This can be seen by the few observations that lie to the right of the red line in Figs. 2-4. Consequently, for most departures, smaller ferries could be used, and it is only at the extremes of the demand distribution where larger ferries, as operated, is necessary. The pattern is observed at all crossings we examine.

The base scenario generally leads to a capacity that will increase the number of users being left behind. When looking at the figures, it is important to keep in mind that even though capacity exceeds demand, not all demand will have to sit back. That is, for those departures in which demand is larger than capacity, only a proportion of that demand will be left behind. Consequently, one cannot directly compare the proportions in the figure to the estimated number of users being left behind[22].

However, one could argue that the base scenario underestimates the number of users being left behind to a certain extent. Looking at the case in which real capacity is set to 50 % of nominal (more users being left behind) and the value of excess waiting time is increased by 100 %, the model still indicates that current capacity is too high, except for Molde-Vestnes in the peak period. However, it is closer to the current capacity levels. Thus, even if our model overestimates "real capacity" and the value of excess waiting time by 100 %, it generally still indicates too much capacity is currently provided. As a result, smaller ferries should be used to limit excess capacity. At the same time, those ferries should be run slightly more often, to reduce the cost of being left behind.

When considering the scenario of electric propellant and increased value of excess time, they indicate a slightly higher capacity, as compared to the base scenario. Changing the value of time (user benefit) or cost per kilometer sailed (operating cost), does not significantly alter the conclusions. Even though the total capacity is similar, the composition is somewhat altered. When the cost per sailed kilometer is lowered (electric propellant scenario), the frequency is higher, but the capacity of each vessel is smaller, as compared to increased valuation of excess waiting time (user benefit). Consequently, although the total capacity is not very different, the optimal mix of frequency and capacity per vessel is sensitive to the assumptions that are made (see table 3).

Fig. 5 displays the percentiles of the demand distributions [PCE/hour] corresponding to the capacity at each scenario for the different crossings, in the peak period. The current capacity exceeds the top percentile of all crossings, which indicate that a high coverage of demand is offered. However, note that these are theoretical calculations based on the data. As mentioned in Section 3.1, data quality may be lower at the tails of the distribution. Consequently, one should put more weight on the levels and changes of these calculations, rather than the exact numerical values.

The base scenario suggests that approximately 1/5 & 1/10 h of operation would have a total demand exceeding capacity. Changing the user benefits of increased capacity (value of excess time + 100 %) or reduced operating costs (electric propellant) suggests that 1/20 h of operation would have a total demand that exceeds capacity. Consequently, although the total capacity is not too different between the scenarios, the proportion of departures that run at maximum capacity is somewhat sensitive. Last, applying the assumptions that both value of excess time and the number of users being left behind is severely underestimated by our model, the coverage approaches a level similar to the current situation. However, as noted in Figs. 2-4, this still corresponds to a lower capacity overall. Consequently, it seems that a too high capacity is offered, even under the most restrictive set of assumptions that we apply. Last, please remember, as noted above, that this figure does not give the number of users left behind directly (see the 4th paragraph and footnote in this section for details).

## 5. Discussion and limitations

As with any study, ours suffers from weaknesses that are crucial to highlight in a scientific context. Consequently, we discuss some aspects that were not included in our analyses.

We use average values of time, taken from the Norwegian Public Roads Administration's handbook on project evaluation (NPRA, 2018). If the preferences of local travelers at a crossing differs from the national average, our model will yield biased results. Factors like composition of trip purposes and the share of heavy vehicles may differ between the crossings. However, as they are all situated on the E39 trunk road, we expect there is a reasonably similar composition of traffic between the crossings. Our estimations should nevertheless be regarded as examples and approximations based on uncertain assumptions about time values and operating costs, rather than exact measurements. The general direction and magnitude of changes is more interesting than the actual numerical results.

The methodology to estimate the number of users left behind is quite simple. For example, it employs *hours* as the unit and not departures (se Section 4.2.4 for additional details). Using yearly averages also may create challenges when there is substantial seasonal variation. Our model indicated a sizeable gap between current and optimal vessel size. However, such large differences should be viewed with skepticism, and it may be that our model underestimates the number of users being left behind. Separating periods with very high demand levels (such as the summer period) in the optimization may yield a better representation of capacity and the number of users left behind. Jara-Diaz et. al. (2017) investigated optimal fleet size, frequencies and vehicle capacities in peak and off-peak periods. When vehicle capacity was constant in both peak and off-peak (as is the case with ferries), peak frequencies would increase and vehicles would be smaller, as compared to a single-period case. Consequently, our results would most likely be extended further in the direction observed, if we used an additional separation between periods of high and low demand. However, the study did not consider user costs of capacity, which may affect the conclusions to a certain degree.

Our method assumes normally distributed demand and does not explicitly take queuing behavior at the ferry quays into account. Sensitivity tests do indicate that results are sensitive to the number of users being left behind. However, the main finding that too high capacity is being enforced, is not. It remains valid even though we assume that real capacity is 50 % of nominal and the value of excess waiting time is doubled as compared to the base case – which may possibly be viewed as quite strong assumptions.

[21] For simplicity, the current demand levels are used, and not the ones resulting from the optimal capacity change. However, the figures will indicate the optimal capacity given current demand levels. We only show plots for a single direction, as the pattern is the same on both directions for all crossings.

[22] Another way to frame this problem is by observing that the probability of demand exceeding capacity is not the same as the share of users being left behind. That is, *P(Demand > Capacity) = (Share of users left behind)* is only valid if capacity is close to zero, as both then approaches 1. Otherwise *P(Demand > Capacity) > (Share of users left behind)* since only a fraction of total demand will not be able to board, i.e. (*Demand – Capacity)/ Demand.*

Using a case study approach has the advantage of yielding detailed knowledge of specific cases. However, generalizability of the results to other cases and countries is a possible drawback. We have investigated a total of three crossings, which is only a fraction of approximately 130 active crossings in Norway. Consequently, applying our model to ferry operations in other countries and regions is a possible avenue for further research. However, we have demonstrated that the methodology have important policy implications. Some generalization could perhaps be possible, if one observes crossings having similar service levels, demand, length etc. Using such indicators could be beneficial for selecting additional case studies, that could shed light on the appropriate service level of specific crossings, increasing social efficiency.

Our problem formulation did not include any financial constraints. As shown by Jara-Díaz and Gschwender (2009), a financial constraint will generally lead the government operator to put less weight on the user's cost and implicitly act in a manner similar to a private operator, the tighter the constraint is. Moreover, such a constraint would lead to using too large ferries and too low frequency, which is generally what we observe (at least in the case of vessel size). Thus, the contrast to current service levels may be explained by the government planners having other constraints than the one employed here.

Our observation that smaller vessels should be run more often when optimizing social surplus, mirrors findings made by Jansson (1980) and Jara-Díaz and Gschwender, 2009) for the bus sector. However, Börjesson et al. (2017) found that larger buses should be used when taking on-board crowding into consideration. We obtain the opposite result when capacity is taken into account. However, on-board crowding is influenced by frequency only through the "second channel", that is capacity (see chapter 4.2.4). As frequency influences both the probability of not being able to board and the user cost, when having to actually sit back, is has two positive effects in the ferry sector, as compared to the one in bus related applications. This is a likely explanation for the difference between the findings of Börjesson et al. (2017) and ours.

There are no environmental costs in the model. Ferries may use diesel engines, which emits different pollutants, such as greenhouse gases. Consequently, using larger ferries or running them more often may influence emission levels. For example, if the effect of raising frequency is larger than vessel size on emissions, frequency should be smaller than what is found in our estimations.

We have not included any cost of upgrading the physical infrastructure at the quays. If more ferries are to be used, a need for additional berths may arise. Also, using smaller ferries may entail rebuilding the infrastructure. Such costs were not included and may affect the results.

Last, we use a static framework, where demand and investment costs for a single year is used. If there is traffic growth, and one has to commit to a certain vessel size for many years, one should also investigate whether demand growth increases user costs and the need for a higher level of capacity over time. Such a concern is a possible avenue for further research, which could make it optimal to employ larger vessels sizes than our model suggests.

## 6. Conclusion

Ferries constitute an important part of the transport network in many parts of the world. To enable policy makers to provide good recommendations about the appropriate quality of ferry services, costs and benefits to society must be accounted for. A central question when designing ferry service levels is the optimal capacity and quality that should be implemented. The literature on optimization public transportation services, does not adequately address the question of optimal service levels and capacity at ferry crossings. Consequently, obtaining a better understanding of how an appropriate level is found, is of value to policy makers.

In this paper, we have developed a model to optimize capacity at ferry crossings. We have added to the literature in the following ways. We have developed a model that simultaneously determines the optimal capacity and number of vessels, in contrast to other models in the literature that only treats a single decision variable (see, e.g. Jørgensen and Solvoll, 2018; Høyem & Odeck, 2020). Moreover, we have estimated optimal capacity using costs relevant to ferry services, such as excess waiting time cost, as the current state of literature does not describe such cost in an appropriate way.

Further, our model was applied to a case study using three major car ferry links in Norway. Our results suggest that current capacity levels are too high as compared to the socially optimal ones. We note that our methodology may underestimate the cost of supplying too low capacity. However, when subjected to relatively strict sensitivity tests, our result remains valid. Consequently, our results indicate that policy makers need to rethink how service levels for ferries are designed and possibly revise them. A case study approach restricts the generalizability of our results. Nonetheless, the results are probably of interest to a wider audience of policy makers, as ferries are used in large parts of the world. Thus, our study enables planners to make better informed decisions regarding the design of service levels at ferry crossings. Developing a method that is better suited than ours to handle the stochastic nature of demand, in order to estimate the number of users left behind, is a possible avenue for further research. For example, queuing theory may be a fruitful approach. Moreover, environmental concerns may also be relevant to include in future studies[23], as done by Zhang et. al. (2017).

*Acknowledgments*

We would like to thank the Norwegian Public Roads Administration (grant 19/145723-2), Oslo Norway and Urbanet analyse, Oslo, Norway for funding this study.

*Conflict of interest statement*

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## Appendix A

*Appendix 1: Parameter table*

Table A1 gives an overview of the different parameters used in the models. Each row in the table contains the parameter name, as introduced in the text, its symbol, value, unit, source and in equation in which it is first defined/used. All monetary parameters were adjusted by the Norwegian consumer price index (CPI) to 2019-NOK. Each parameter is defined, and its value motivated in greater depth in the equation where it is defined. Consequently, we refer the reader to the text where each equation is situated for additional details.

[23] We would like to thank an anonymous reviewer for suggesting the queuing approach and drawing our attention towards possible environmental issues, which may be interesting in further research.

*A2 Separating open and hidden waiting time (OW & HW)*

Total waiting time (hidden and open) is given as half the headway:

$$W_T = 1/2f(V) \tag{A1}$$

This equals the valued sum of open and hidden time

$$W_T = W_H + W_O \tag{A2}$$

Using the two equations weighted by the appropriate value of time ($\omega_H$ , $\omega_O$), we find that total waiting time costs become:

**Table A1**
Parameters used in the model.

| Parameter | Symbol | Value | Unit | Source | Defined in eq. # |
|---|---|---|---|---|---|
| phi1 (Headway model) | $\phi_1$ | 0.618 | N/A | Own model based on Jørgensen & Solvoll (2018) | A5 |
| phi2 (Headway model) | $\phi_2$ | −0.60 | N/A | Own model based on Jørgensen & Solvoll (2018) | A5 |
| Hours per mooring | $m$ | 0.0667 | Hours | Jørgensen et al. (2007) | 14 |
| Hours per PCE on–off | $q$ | 0.0014 | Hours | Jørgensen et al. (2007) / Own side calculations | 14 |
| Value of time (Open) | $\omega_O$ | 218 | Kr/hour | NPRA (2016b) - Adjusted by CPI + 7 % | 3 |
| Value of time (Hidden) | $\omega_H$ | 109 | Kr/hour | NPRA (2016b) - Adjusted by CPI + 7 % | 3 |
| Value of time (Excess) | $\omega_E$ | 285 | Kr/hour | NPRA (2016b) - Adjusted by CPI + 7 % | 4 |
| Wage rate | $w$ | 413 | NOK/ HOUR | NPRA (2015) | 8 |
| Crew/ferry | $A$ | 6 | Crew/ferry | Own assumption based on NPRA (2015) | 8 |
| Days of operation/year | $OD$ | 365 | Days | Own assumption | 12 |
| Life-time of vessels | $n$ | 30 | Years | Svendsen et al. (2017) | 12 |
| Interest rate (yearly) | $r$ | 4% | | Svendsen et al. (2017) | 12 |
| Specific consumption | $F$ | 0.208 | kg/Kwh | NPRA (2015) | A9 |
| Specific weight marine diesel | $\tau$ | 0.84 | kg/liter | NPRA (2015) | A9 |
| Fuel price | $p_{Diesel}$ | 6.47 | kr/liter | Svendsen et al. (2017) Adjusted by ferry cost index (fuel) + 20.3% | A9 |
| Parameter cost model (constant) | $c_1$ | 33.1 | mill. NOK | Own model based onNPRA (2015) - Adjusted by CPI + 15.5 % | A12 |
| Parameter cost model (slope) | $c_2$ | 1.2 | mill. NOK | Own model based onNPRA (2015) - Adjusted by CPI + 15.5 % | A12 |
| Maximum fuel share of cost (Excluding crew and capital cost) | $\alpha_{max}$ | 70% | | Svendsen et al. (2017) | 11 |
| Parameter fuel cost share (constant) | $a_1$ | 0.4 | | Svendsen et al. (2017) Adjusted by ferry cost index (fuel) + 20.3 % | 11 |
| Parameter fuel cost share (slope) | $a_2$ | 0.1 | | Svendsen et al. (2017) Adjusted by ferry cost index (fuel) + 20.3 % | 11 |
| Price elasticity of demand | $\varepsilon_p$ | −0.2 | | Own assumption (half of average in Balcombe et al. (2004)) | A14 |
| Parameter energy model (constant) | $b_1$ | 872 | kwh | NPRA (2015) | A10 |
| Parameter energy model (slope) | $b_2$ | 20.3 | kwh | NPRA (2015) | A10 |

$$OW(V)+HW(V) = \omega_H W_H + \omega_O W_O = \omega_H(W_T - W_O) + \omega_O W_O = W_O(\omega_O - \omega_H) + \omega_H W_T \tag{A3}$$

Inserting $W_T$ , we get:

$$OW(V)+HW(V) = \frac{\omega_H}{2f(V)} + (\omega_O - \omega_H)W_O(V) \tag{A4}$$

Further, $W_O(V)$ is given by

$$W_O(f(V)) = \phi_1 \exp(\phi_2 {}^* f(V)) \tag{A5}$$

The parameters of $W_O(V)$ is based upon the functional form used by Jørgensen & Solvoll (2018), which was estimated based on a travel survey of ferry users. However, their chosen functional form, has undesirable mathematical properties, which renders the numerical optimization procedure unstable. Therefore, we chose the alternative form, which, given the appropriate coefficient values enables computation of the relationship between open waiting time and frequency as in Jørgensen & Solvoll (2018).

*A3 Estimation of excess waiting time cost*

We use the basic framework of Høyem and Odeck (2020) as a point of departure to estimate users' costs when demand surpasses capacity. The model makes a number of assumptions highlighted in Høyem and Odeck (2020):

- Demand follows a normal distribution.
- There is no queuing at the quay which is not the case in most car ferry operations. However, Høyem & Odeck (2020) found this to be an innocuous assumption that may be maintained in spite of some theoretical counterarguments.

- The time period of analysis is per hour. As such, the model effectively assumes there exists and average and equal probability of being left behind each departure within the given time period (that is, each hour).

We assume there are only two quays, A and B. Define the average demand on a crossing at time $t$ ($\mu_t$) as the mean demand from A to B ($\mu_t^{A\to B}$) and B to A ($\mu_t^{B\to A}$):

$$\mu_t = \frac{1}{2}\left(\mu_t^{A\to B} + \mu_t^{B\to A}\right) \tag{A6}$$

Let the set $IT = \{A\to B, B\to A\}$ be the itinerary for one round trip. Høyem & Odeck (2020) used the concept of a loss function to estimate the share of users being left behind per time period ($\eta_t$) in direction $j$ by the following equation:

$$\eta_t = \sigma_t(\phi(z_t) - z_t(1-\Phi(z_t))) \tag{A7}$$

Here, $\sigma_t$, is the standard deviation of demand for period $t$, $\Phi()$ is the standard normal cumulative distribution function (cdf) and $\phi()$ is the standard normal density function (pdf).

We now define $z_t$ as follows:

$$z_t = \frac{f(V)^*k - \mu_t}{\sigma_t}, j \in IT \tag{A8}$$

where $f(V)$ and $k$ is defined as previously. The standard deviation is for simplicity defined as $\sigma_t = \frac{1}{2}\left(\sigma_t^{A\to B} + \sigma_t^{B\to A}\right)$ .

*A4 Ferry running costs*

Ferry running costs are estimated on the basis of the methodology presented in the official guidelines for CBA in Norway (NPRA, 2015). The cost per sailed kilometer is estimated as:

$$c_l(k) = p_{Diesel}{}^*\frac{F}{\tau}{}^*E(k)\frac{1}{v_{km}} \tag{A9}$$

Here, $p_{Diesel}$ is the price of diesel in NOK/liter, $F$ is the number of kg diesel / kilowat hour (kwh), $\tau$ is the number of kg/liter diesel, $E(k)$ is the number of kwh/hour of engine operation for each ferry and $v_{km}$ is the speed of the vessels in km/hour.

The relationship between energy consumption for each ferry and capacity is found by the following equation:

$$E(k) = b_1 + b_2{}^*k = 20.3^*K + 872 \tag{A10}$$

This equation is based upon an average kwh and capacity for 10 classes of ferries, given in NPRA (2015) - table 54. As such, it may be regarded as an approximation to the average relationship between ferry size and energy usage, as viewed from the official CBA guidelines in Norway.

Last, the share of sailing cost comprised out of total operating costs (less crew costs) is given by the following equation:

$$\alpha_{Sailing} = \min(a_1 + a_2{}^*V_\Sigma, \alpha_{Max}) \tag{A11}$$

Svendsen et al. (2017) use the value of $\alpha_{Sailing} = 0.5$ for $V_\Sigma = 1$ , $\alpha_{Sailing} = 0.6$ for $V_\Sigma = 2$ and $\alpha_{Sailing} = 0.7$ for $V_\Sigma = 3$ . Consequently, the best estimate we may attain, is found by setting $a_1 = 0.4$ , $b = 0.1$ og $\alpha_{Max} = 0.7$ , following the schedule of Svendsen et. al. up to $V_\Sigma = 3$ , and assuming $\alpha_{Sailing} = 0.7$ for $V_\Sigma \geq 3$ . This is arguably a simplification, but we contend it is as appropriate approximation, sufficient for the current purpose of the paper.

*A5 Ferry capital costs*

On the basis of norm-based cost data obtained from NPRA (2015), we have estimated total investment cost as a function of capacity using the following equation:

$$I(k) = c_1 + c_2{}^*k = 33.14 + 1.19^*k \tag{A12}$$

The equation is based on a least-squares approximation to the data in NPRA (2015).

*A6 solution method*

Eqs. ((1)–(3)) together define a system of non-linear equations which enables solving for the optimal prices, vessels and capacity for a given level of demand. The model is solved numerically by iteratively estimating the optimal value of decision variables and adjusting demand until convergence to the equilibrium. We first describe how the model is solved for each level of demand, then the iteration process.

**Solving the equation system**

The equation system is solved as follows: There are two shifts and 24 h each day. In addition to the derivative of the objective function with respect to capacity, this means there are $2 + 24 + 1 = 27$ equations to be solved. Further, there are 27 unknowns ($V_1$ , $V_2$ , $k$ and $p_t, t \in [1, 24]$), such that the system has a solution. The model is solved numerically. First, all equations including derivatives are estimated using the finite-difference approach (Nocedal & Wright, 2006). Secondly, the equation system is solved using the “fsolve” function available in Python Scipy (Virtanen et al., 2020).

### Estimating convergence to the economic equilibrium

We now describe how the convergence to the economic equilibrium is estimated. Observe that the economic equilibrium is characterized by the last user equating their marginal willingness to pay to the user cost plus price, such that:

$$\bar{h}(x) = \theta(x, V, k) + p \tag{A13}$$

Further, define the demand function as $(p)$ . The marginal willingness to pay equals the inverse of the demand function such that $\bar{h}(x) = \nu^{-1}(p)$ (Varian 1992). By basic mathematics, the equilibrium is defined by $= \nu(\nu^{-1}(p)) = \nu\left(\bar{h}(x)\right) = \nu(\theta(x, V, k) + p)$ . Thus, when $(\theta(x, V, k) + p) = x$ , this implies that $\bar{h}(x) = \theta(x, V, k) + p$ which again defines the economic equilibrium. Consequently, the equilibrium is a fixed point of the demand function. A fixed point of a continuous function is defined as $= f(z)$ . Then $\nu(\theta(x, V, k) + p) = x$ is a fixed point as $\nu$ is a function of $x$ *and* returns .

A fixed point of a function may be found using fixed-point iteration through a recurrence relation (Wood, 1999, p. 144) with $z^n = f(z^{n-1})$ where $n$ denotes the iteration number. We use a simple demand function as our recurrence relation from Balcombe et al. (2004), p. 47:

$$x^{n+1} = \nu(x^n) = x^n * \left(\frac{\theta^{n+1} + p^{n+1}}{\theta^n + p^n}\right)^{\varepsilon_t} \tag{A14}$$

The term inside the parenthesis denote user cost and price respectively. The model is calibrated by estimating the generalized cost elasticity using the formula $\varepsilon_t = \varepsilon_P(\theta_0 + p_0)/p_0$ where $\varepsilon_P$ is the price elasticity of demand. The price elasticity is set to $-0.2$, reflecting the assumption that ferry trips are relatively price-insensitive, as the alternative routes are very long (at least in the cases we examine). The value corresponds to half of the price elasticity for buses found in Balcombe et al. (2004), table 6.1, p. 50. A similar iterative approach as ours was adopted by Li (2002) when estimating optimal congestion prices.

The iteration process solves for optimality in $V, k$ and $p$ by solving the equation system of first-order derivatives, then calculating the new user cost and price, and updating demand. The process converges when the difference in demand between iterations, measured by the vector norm of demand between iteration $n$ and $n-1$ for all periods $\Delta = \sqrt{\sum_t \left(x_t^n - x_t^{n-1}\right)^2}$ falls below a threshold of 0.001. Further, the price is updated by setting $p^{n+1} = 1/2\left(p^n + p^{n-1}\right)$ for numerical reasons to obtain a smoothly convergent series.

## References


Andersen, S.N., Tørset, T., 2019. Waiting time for ferry services: empirical evidence from Norway. Case Stud. Transport Policy 7 (3), 667–676.

Asplund, D., Pyddoke, R., 2020. Optimal fares and frequencies for bus services in a small city. Res. Transp. Econ. 80, 100796. https://doi.org/10.1016/j.retrec.2019.100796.

Baird, A.J., Wilmsmeier, G., 2011. Public tendering of ferry services in Europe. Eur. Trans. 49, 90–111.

Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., White, P., 2004. The demand for public transport: a practical guide.

BCFerries (2020): BCFerries. URL: Connecting the Coast | BC Ferries. Visited 13.05.2021.

Bignon, E., Pojani, D., 2018. River-based public transport: why won't Paris jump on board? Case Stud. Transport Policy 6 (2), 200–205.

Börjesson, M., Fung, C.M., Proost, S., 2017. Optimal prices and frequencies for buses in Stockholm. Econ. Transp. 9, 20–36.

Börjesson, M., Fung, C.M., Proost, S., Yan, Z., 2019. Do small cities need more public transport subsidies than big cities? J. Transp. Econ. Policy (JTEP) 53 (4), 275–298.

City of Toronto (2020): All Ferry Schedules. URL: All Ferry Schedules – City of Toronto. Visited 13.05.2021.

Department of Transportation and Public Affairs (2020): Alaska Marine Highway System. URL: Alaska State Ferry | Alaska Marine Highway System. Visited 13.05.2021.

DNV GL, 2015. *Muligheter og kostnader ved bruk av fornybar energi på Ruters båtsamband* English: Opportunities and costs associated with the use of renewable energy in Ruter's ferry services. Technical report, Det Norske Veritas.

Findley, D.J., Anderson, T.J., Bert, S.A., Nye, T., Letchworth, W., 2018. Evaluation of wait times and queue lengths at ferry terminals. Res. Transp. Econ. 71, 27–33.

Flügel, S., Halse, A. H., Hulleberg, N., Jordbakke, G. N., Veisten, K., Sundfør, H. B., & Kouwenhoven, M., 2018. Verdsetting av reisetid og tidsavhengige faktorer. *Dokumentasjonsrapport til Verdsettingsstudien, 2019*. (English: Technical report of the Norwegian Value of Time Study).

Hanssen, T.E.S., Jørgensen, F., Larsen, B., 2019. Determinants affecting ferry users' waiting time at ferry terminals. Transportation 1–22.

Holmgren, J., 2018. The effects of using different output measures in efficiency analysis of public transport operations. Res. Transp. Bus. Manage. 28, 12–22.

Høyem, H., Odeck, J., 2020. Optimal public transit frequency under stochastic demand and fixed vehicle size: application in the Norwegian car ferry sector. Res. Transp. Econ. 82, 100878. https://doi.org/10.1016/j.retrec.2020.100878.

Jansson, J.O., 1980. A simple bus line model for optimisation of service frequency and bus size. J. Transp. Econ. Policy 53–80.

Jansson J.O., Holmgren, J., Ljungberg, A., 2015. Pricing public transport services, Handbook of research methods and applications in transport economics and policy, 260-308. Edward Elgar Publishing. Cheltenham.

Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2017. Optimal fleet size, frequencies and vehicle capacities considering peak and off-peak periods in public transport. Transp. Res. Part A: Policy Practice 106, 65–74.

Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2020. Strategies for transit fleet design considering peak and off-peak periods using the single-line model. Transp. Res. Part B: Methodological 142, 1–18.

Jara-Díaz, S., Gschwender, A., 2003. Towards a general microeconomic model for the operation of public transport. Transp. Rev. 23 (4), 453–469.

Jara-Díaz, S.R., Gschwender, A., 2009. The effect of financial constraints on the optimal design of public transport services. Transportation 36 (1), 65–75.

Jørgensen, F., Mathisen, T., Solvoll, G., 2007. Kapasitet i fergesektoren (English: Capacity in ferry operations). Technical Report, SIB-rapport 1/2007.

Jørgensen, F., Pedersen, H., Solvoll, G., 2004. Ramsey pricing in practice: the case of the Norwegian ferries. Transp. Policy 11 (3), 205–214.

Jørgensen, F., Mathisen, T.A., Larsen, B., 2011. Evaluating transport user benefits and social surplus in a transport market—the case of the Norwegian ferries. Transp. Policy 18 (1), 76–84.

Jørgensen, F., Solvoll, G., 2018. Determining optimal frequency at ferry crossings. Transp. Policy 63, 200–208.

Li, Michael Z.F., 2002. The role of speed–flow relationship in congestion pricing implementation with an application to Singapore. Transp. Res. Part B: Methodol. 36 (8), 731–754.

Lipsey, R.G., Lancaster, K., 1956. The general theory of second best. Rev. Econ. Stud. 24 (1), 11–32.

Nocedal, J., Wright, S., 2006. Numerical Optimization. Springer Science & Business Media.

North Carolina State Department of Transportation (2020). N.C. Ferry System. URL: https://www.ncdot.gov/travel-maps/ferry-tickets-services/Pages/default.aspx. Visited 13.05.2021.

Norwegian Public Roads Administration, 2015. *Dokumentasjon av beregningsmoduler i EFFEKT 6.6*. SVV-rapport Nr. 358. (English: Documentation of calculation modules in EFFEKT 6.6).

Norwegian Public Roads Administration (2018): Håndbok V712 – Konsekvensanalyser. (English: Handbook for cost benefit analyses).

Nyborg, K., 2014. Project evaluation with democratic decision-making: what does cost–benefit analysis really measure? Ecol. Econ. 106, 124–131.

NYC Ferry (2020): NYC Ferry. URL: New York NYC Ferry Routes & Schedules. Visited 13.05.2021.

Pedersen, P.A., 2003. On the optimal fare policies in urban transportation. Transp. Res. Part B: Methodological 37 (5), 423–435.

Ramsey, F.P., 1927. A Contribution to the Theory of Taxation. Econ. J. 37 (145), 47–61.

Škurić, M., Maraš, V., Davidović, T., & Radonjić, A., 2020. Optimal allocating and sizing of passenger ferry fleet in maritime transport. Research in Transportation Economics, 100868. Article in press.

Slesnick, D.T., 1998. Empirical approaches to the measurement of welfare. J. Econ. Literature 36 (4), 2108–2165.

Statistics Norway (2021): The car fleet. URL: https://www.ssb.no/transport-og-reiseliv/landtransport/statistikk/bilparken. Visited 09.05.2021.

Svendsen, H.J., F. Müller, H. Thune-Larsen, S. Bråthen (2017): Utgiftsbehov til ferjer og hurtigbåter – Utredning av fylkeskommunale forskjeller. Rapport Nr. 1701. Høgskolen I Molde. (English: Cost levels of car and rapid ferry services – analyses of regional county differences).

Tirachini, Alejandro, Antoniou, Constantinos, 2020. The economics of automated public transport: effects on operator cost, travel time, fare and subsidy. Econ. Transp. 21, 100151. https://doi.org/10.1016/j.ecotra.2019.100151.

Tirachini, Alejandro, Hensher, David A., Rose, John M., 2014. Multimodal pricing and optimal design of urban public transport: the interplay between traffic congestion and bus crowding. Transp. Res. Part B: Methodological 61, 33–54.
Transport for London (2020): River. URL: River - Transport for London (tfl.gov.uk). Visited 13.05.2021.
Varian, H.R., 1992. Microeconomic Analysis. WW Norton & Company Inc.
Virtanen, P., Gommers, R., Travis, E.O., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, Weckesser, W., Jonathan, B., Stéfan, J.v.d.W, Matthew, B., Wilson, J., Jarrod, M., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, . E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., Mulbregt P.v., and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.
Washington State Department of Transportation (2020): Washington State Ferries. URL: https://wsdot.wa.gov/ferries. Visited 13.05.2021.
Wood, A., 1999. Introduction to Numerical Analysis. Addison Wesley Longman.

# Paper IV

Economics of Transportation 30 (2022) 100259

Contents lists available at ScienceDirect

Economics of Transportation

journal homepage: www.elsevier.com/locate/ecotra

# Public transport frequency and risk-taking behavior

Harald Høyem [a,b]

[a] Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[b] Asplan Viak, Urbanet Analyse Department, Oslo, Norway

ARTICLE INFO

*Keywords:*
Transport economics
Optimization
Traffic safety
Optimal frequency
Speed selection

When travellers connect to a transit service from a different mode, they must arrive at the connection in a timely manner. If there is uncertainty about the required time to meet the connection, some users might engage in risky behavior by, e.g., increasing their traveling speed. We examine whether the frequency level at a transfer connection may influence the incentive to engage in such risky behavior. We develop an optimization model in which users select an optimal speed in a two-stage process. A simulation study is performed to study the behavior within a wide range of possible preferences and trip characteristics. Our results suggest that increasing the departure frequency may provide a greater incentive for engaging in risky behavior – increasing social costs by increasing the number of accidents. The result is dependent upon average trip length, the initial frequency and the user's perception of scheduling cost. Policy makers should consider the possibly increased accident costs when altering the service level at a transfer connection.

## 1. Introduction

The design of optimal service levels for public transport is an important task of decision-makers who oversee transit systems. Because there are significant operational and capital cost requirements associated with operating public transport services, it is important that all the relevant costs and benefits of operating a service are included. Service levels can be influenced in several ways, with the level of frequency being one aspect of importance.

Many public transport systems rely on transfers between different services as well as on transport modes other than public transport, such as walking (O'Sullivan and Morrall, 1996), bicycling (Martens, 2007) and driving (Parkhurst, 2000). Transfers to public transport from other modes require that the users schedule their time appropriately such that they can reach their desired departure on time. However, when there exists uncertainty in the time required to reach the desired departure, or when users experience an unplanned delay, an incentive for engaging in risky behavior (for example, by increasing the traveling speed) can be created because the users are "in a hurry" (Sadia et al., 2018).

An interesting question is whether the service level design of the public transport systems can affect the risk-taking behavior to "catch the bus". If an increased frequency is provided, as an incentive for lowering risk-taking behavior, with all else remaining equal, it could yield an additional benefit to society. On the other hand, if the increased frequency raises the incentive for engaging in risky behavior, it could yield an additional cost to society. Thus, understanding how the frequency of a transfer service and risk-taking for travelers is linked could be important in designing optimal service levels for public transport. To the best of our knowledge, this aspect has not been accounted for in the literature.

Researchers usually agree that risk-taking behavior, such as a person increasing his/her speed, has a significant influence on the number of accidents (Elvik et al., 2019; Aarts and Van Schagen, 2006). Further, governments and individuals usually view accidents as a phenomenon that carries high social costs. For example, the Swedish road authorities estimate the loss to society from a fatal accident to be 46 million 2014 SEK,[1] whereas the Norwegian Public Roads Administration sets the value at 30 million 2016 NOK,[2] with significant figures also for non-fatal, but serious accidents. Thus, traffic safety is an important topic that policymakers account for in their decisions on budget allocations (Odeck, 2010), and it is viewed as a growing public health issue (World Health Organization, 2018).

The purpose of this paper is the development of a theoretical model that links car travelers' incentives for selecting high speeds and the frequency of a transit service. The goal is by no means an attempt to deliver a definitive answer to the question of whether transit frequency could be influence risk-taking behavior. However, we aim to provide some theoretical insights that could be a first step toward understanding

*E-mail address:* harald.hoyem@asplanviak.no.
[1] 1 SEK = 0.09 EUR/0.1 USD.
[2] 1 NOK is approximately equal to 1 SEK in 08.02.2020.

https://doi.org/10.1016/j.ecotra.2022.100259
Received 9 October 2020; Received in revised form 11 February 2022; Accepted 26 February 2022
Available online 13 March 2022

the relevant mechanism and possibly providing a direction for empirical research.

The scientific contributions of this paper should be highlighted. The first contribution of the paper is a widened understanding of the impacts that the service frequency could have on the societal costs in the context of transport economics. The literature on the design of optimal frequency levels has treated different variables that impact the total cost. However, it has mostly focused on intentional effects such as the user waiting time and on-board congestion (Jara-Diaz and Gschwender, 2003; Börjesson et al., 2017). Non-intentional effects of changing the frequency have been studied to a lesser extent, perhaps, except for on-road congestion when many buses are present (Tirachini et al., 2014; Tirachini and Hensher, 2011; Börjesson et al., 2018) and network effects (Fielbaum et al., 2020). To increase the understanding of how the optimal frequency may be influenced, it is important to include all relevant aspects, such as safety. To the best of our knowledge, our model is the first attempt to discuss the relevance of safety aspects in this context.

The paper also contributes by extending the literature on driver speed selection by studying how chosen speeds are influenced by public transit frequency when making a transfer between modes. A two-stage model is developed, where both departure and speed level are chosen, in contrast to the typical one-stage model applied, that studies only the optimal speed level (see, e.g., Jørgensen and Sandberg-Hanssen, 2019).

To make a final policy recommendation, both the costs and benefits of an altered frequency should be accounted for. For example, reducing the frequency also entails a cost for the existing users which need to be considered. Moreover, one should compare different policies to find the most effective one. There is a large literature on measures that seek to reduce risk-taking in road traffic (see, e.g., Elvik et al., 2004), which may also be effective in reducing the incentive to speed.[3] In this paper, we do not seek to develop a complete optimization model which compares the relative merits of different traffic safety policies, but rather investigate if frequency may affect risk-taking in isolation. Such information may be of interest in and of itself, for the following two reasons:

- Firstly, if there are no other safety policies available for implementation, it is interesting to consider frequency.
- Secondly, if a planner considers altering the frequency, information on whether, and possibly in which direction, risk-taking may shift is important. It may be that other policies than altering (or not altering) the frequency is the most effective one, but it is still important to have information on what effects one might observe, and then select the most suitable policy measure. Further, to assess if frequency change is an effective policy, one needs knowledge of its effect.

Our main result indicates that increasing the frequency might yield a larger incentive for engaging in risky behavior by selecting a greater speed level when transferring to a public transit service from another mode. However, the marginal effect of increasing frequency is found to differ depending on the initial level of frequency. For low levels of frequency at the outset, it is more likely that increases will raise speeds. From moderate levels of frequency, it is more likely that they will be reduced. The results are also highly dependent upon trip length and driver's relative perception of different cost components.

These concerns are policy-relevant if the drivers do not internalize the full social costs of their accidents (if drivers do not consider the external social cost of accidents, including their passengers', other travelers' cost or loss to aggregate output of the economy). In the absence of any road-pricing, the drivers will not internalize such costs.[4] Consequently, planners should take possibly altered accident cost into consideration when increasing service levels.

The remainder of this paper is organized as follows: Section 2 provides a literature review. Section 3 gives an exposition of our model. Section 4 provides a simulation study of the model. Section 5 discusses the results, while section 6 gives the conclusions.

## 2. Literature review

Optimization of frequency and speed chosen by drivers (henceforth, speed selection) are topics that separately has received much attention in the transportation literature. We now provide a short overview of the relevant literature.

On the whole, researchers agree that the built environment and road characteristics influence the speed selection (Polus et al., 2000; Poe and Mason, 2000; Fitzpatrick et al., 1997). Speed limits, curvature (horizontal and vertical), grade, traffic volume, the number of accesses and line of sight distance appear to impact the speed levels chosen by drivers. Moreover, Sadia et al. (2018) found that drivers increase their speed when they are at risk of being late for an appointment, and drivers with more experience drive faster, overestimating others' speed and attempting to match the speeds of the other vehicles, which was also observed by Haglund and Åberg (2000). Some studies seek to explain speed selection by means of a rational utility-maximizing or cost-minimizing agent. O'Neill (1977) was the first author to provide an explicit optimization model in which the user balanced the time, accident risk and the cost of receiving a ticket for speeding, and all of the frameworks discussed here are in general related to his initial model. This model has been extended by many authors. Blomquist (1986) considered an optimal driver "safety effort", which in itself decreased utility and accident costs that were reduced if the effort was increased. Jørgensen and Polack (1993) considered how different personal characteristics such as age, experience, preferred free speed, sex and importance of travel time savings affected speed selection. Jørgensen and Pedersen (2005) analysed drivers' perception of being fined or losing their driver's licence, noting that such risk is overestimated. Jørgensen and Wentzel-Larsen (1995) discussed optimal use of warning signs to increase the risk perception of users at accident prone locations. Tarko (2009) investigated how risk, enforcement and subjective time perception affected the optimal speed level chosen by drivers using observed speeds, in addition to different infrastructure related variables such as presence of poles, barriers, etc. Jørgensen and Sandberg Hanssen (2019) developed a model to study how secondary driver engagement, such as talking to passengers or listening to the radio might affect optimal speeds. All of these papers use a utility or cost optimization model, where the optimal speed is a balancing of different cost components. However, none of these studies have investigated the relationship between frequency of a transit service and speed selection by drivers, as with our model.

Optimization of frequency level is an important subject in transportation economics. Safety aspects, however, is lacking form the literature. Mohring (1972) was the first author to provide a mathematical relationship between the optimal service levels, demand and operational cost, where the waiting time cost of travelers are weighted against the operating and capital cost of the policy makers. His models have later been extended and/or used by a number of authors. Jansson (1980) extended the analysis to treat bus size as endogenous.

Furth (1981) considered how resources could be divided optimally between multiple lines when budget constraints are present. Jara-Diaz and Gschwender (2003) extended the model to incorporate crowding; Jara-Diaz and Gschwender (2009) included the effect of a planner facing financial constraints on the optimal frequency as opposed to a case with no constraints. Jara-Diaz and Gschwender (2017) investigated optimization in multiple time periods with both frequency and bus size variable in peak and off peak. Their analysis also extended the one of Jansson (1980), by not assuming an equal bus size in peak and off-peak.

[3] We would like to thank an anonymous reviewer for pointing these factors out to us.

[4] Thune-Larsen et al. (2014) suggests accident externalities amount to about 0.17 NOK/km (2016-prices), in the Norwegian case. 1 NOK = 0.099 EUR.

Tirachini et al. (2014) investigated how increased bus frequency affect both crowding on board and congestion costs for cars sharing the same infrastructure as buses. A similar study was also performed by Börjesson et al. (2017). Börjesson et al. (2018) studied how optimal frequency is affected by the interaction between cyclists and buses, as increased frequency of buses may cause delays for cyclists. Jørgensen and Solvoll (2018) studied optimal frequency for car ferries, while Tirachini and Antoniou (2020) discussed how optimal frequency might change with the introduction of automated vehicles. Fielbaum et al. (2020) considered network effects and its implications for optimal frequency. Jara-Diaz et al. (2020) considered optimization with two different fleets of vehicles in conjunction with bus bunching. Consequently, a wide range of issued have been addressed in the literature. However, none of these studies have incorporated aspects of safety into the costs and/or benefits of increasing the frequency, which is a contribution of our model.

## 3. A model of driver behavior

### 3.1. Structure

We now present our model. We assume that a driver is attempting to connect to a public transit mode, with the case of a car user as an example. It is assumed that the driver is rational and utility maximizing. The driver starts at home, aiming to arrive at the transfer connection in a timely manner. The driver tries to avoid being too late to make a connection, which would incur delay time costs. Importantly, it is assumed that the driver experiences some *uncertainty* as to the time required to meet the connection.

Further, it is important to mention some additional assumptions that underlie the model before proceeding:

1) The time uncertainty is normally distributed. Although a time distribution may be right-skewed, we use this assumption for the sake of model tractability.
2) There is no shortage of capacity when transferring to the connecting mode, i.e., users do not speed up to be “the first in line” to board in case the expected demand for the transit service surpasses its capacity.
3) The time to board or enter a vehicle or vessel is not included

The driver is assumed to select the optimal speed based on the following two-stage decision-making process:

- **Stage 1**: For each departure, calculate the optimal speed that maximizes utility.
- **Stage 2**: Among the possible departures of the connecting service, select the departure that has the largest utility, given its optimal speed.

Thus, users scan all of the departures given, find the speed that yields the highest expected utility for each departure, and select the departure that has the largest utility of all departures, given the optimal speed of each one. This two-stage process accounts for the fact that the users select two variables, one continuous (speed) and one discrete (departure). A similar approach was advocated by Bates et al. (2001) studying reliability in public transport and choice of departure.

We start by outlining the theoretical structure of the model, before performing some analyses to gain intuition as to what the effect of increasing departure frequency is. Last, we perform some numerical simulations where the model is parametrized, and specific functions are introduced for each component of the model.

We use the .framework of (Small, 1982) as a point of departure in which utility is defined by the following function:

$$U = \alpha T + \beta SDE + \gamma SDL + \theta D_L \tag{1}$$

In Small's original model, $T$ is the expected travel time of a trip, $SDE$ is the scheduling delay of arriving prior to a preferred arrival time (PAT), $SDE$ is the scheduling delay of arriving later than preferred and $D_L$ is dichotomous variable indicating if the trip is delayed or not. Last, $\alpha, \beta, \gamma$ and $\theta$ are cost parameters of the model. This scheduling model has been applied by a number of authors to study trip time choice with uncertain travel time (e.g. Noland and Small, 1995; Fosgerau and Karlström 2010), optimal table tables in public transport considering scheduling costs (de Palma and Lindsey, 2001) and scheduling costs of headway-based services (Fosgerau, 2009) to mention a few.

Our formulation is a modification of Small's model, operationalized by using the following cost of selecting departure $i$ at speed $s$ is given by the following equation:

$$-U(i,s) = C(i,s) = \alpha t(s) + \frac{\omega}{F}\left(i - \frac{1}{2}\right) + \frac{\theta}{F}P_L(s) + \pi A(s) \tag{2}$$

The cost is comprised of the following elements: First, the user experiences a cost given the trip time (access time) to the station, which is the time spent driving, $t(s)$, measured in minutes and dependent on the speed $s$. The cost is valued at $\alpha$, measured in NOK/hour. The time will also depend on the distance to the station, from which we currently abstract. This cost is akin to $\alpha T$ in Small's model.

Second, as with any scheduled transport mode, the user experiences a scheduling time cost, as they are not able to depart exactly at their preferred time.[5] We assume that desired departure times are uniformly distributed within each time period (see, e.g., Osuna and Newell (1972)), which leads to a scheduling time equal to $\frac{1}{F}\left(i - \frac{1}{2}\right)$, where $F$ is the departure frequency per hour. This assumption is reasonable if one considers relatively short time periods. If longer time periods are considered, it may be that desired departure times are more concentrated around a specific time point. We assume the user is planning (by calculating a safety margin). With higher frequencies, some users may not plan their arrival time, such that the cost of a planning user is no longer relevant. When desired departure times are not uniformly distributed, the scheduling cost of a planning and unplanning user can differ (Fosgerau, 2009), such that our cost formulation may become less relevant at higher frequency levels.

The schedule time is evaluated at $\omega$ NOK/hour. The term $\omega = \frac{\beta\gamma}{(\beta+\gamma)}$ is the scheduling cost as given by Fosgerau (2009) and de Palma and Lindsey (2001) for a user planning the arrival time to the station. This cost is akin to the average of $\beta SDE + \gamma SDE$, evaluated over many users, in Small's model, when it is interpreted as a scheduling cost. That is, not as a cost related to the unreliability of travel time, but to the scheduling cost users experience when using a service with fixed time intervals, as done by, e.g., Fosgerau (2009) or de Palma and Lindsey (2001). It is important to underline that this cost is not the same as waiting time cost at the station, but a separate cost that arises since users are restricted in their departure time according to the service schedule. In the framework of Fosgerau (2009), waiting time cost at the station only arises for users who are not planning.

Third, the user has an expected delay cost that arise if they are not able to reach the station in time. The cost has two components: First, if one arrives too late, one has to wait a full headway until the next service arrives ($1/F$), multiplied by the cost per hour of being delayed, $\theta$. It seems reasonable that some users will continue their journey after completing the public transport leg of the trip. In such cases, the delay might be diminished by traveling faster after the public transport leg. In this paper, we only consider the trip to the public transport service.

[5] Some authors, e.g. Fosgerau (2009) used preferred arrival time and setting trip time to zero. In this analysis, we assume the users have a preferred departure time in order to estimate the available time from home to the public transit station.

Second, the probability of arriving too late is given by $P_L(s)$ which is a function if speed. We assume that the probability falls as speed increases.

The literature on travel time variability, in which a traveler faces an uncertain travel time, has highlighted the fact that a safety margin will be added. According to Carrion and Levinson (2012), two different frameworks are typically used: the mean variance approach and the ($\alpha$,$\beta$, $\gamma$)-framework. The latter framework, as introduced by Small (1982), is based on a utility-maximizing consumer and is the one we use in this paper, as it permits utility maximization in the analysis.

For example, Fosgerau and Karlström (2010), showed that, using the so-called ($\alpha$,$\beta$,$\gamma$)-framework, a user will choose a safety margin equal to $\sigma\Phi^{-1}\left(\frac{\gamma}{\gamma+\beta}\right)$ added to the expected trip time, $\mu$, where $\sigma$ is the standard deviation of the travel time distribution and $\Phi$ its cumulative distribution function. Another, but similar approach was taken by Noland and Small (1995) who used a framework to derive safety margins in which they defined a "head start" time which is the "amount one would arrive early if there were no incident-related delays" (Noland and Small, 1995). In our model, the head start time is equal to the difference between the time until departure $i$ (from the users desired departure time) and the expected trip time $t(s)$. This formulation follows the same structure as in Noland and Small (1995):

$$t_{hs}=\frac{1}{F}\left(i-\frac{1}{2}\right)-t(s) \tag{3}$$

The time from the user's desired departure time until the service leaves the station is the same as the scheduling time, $\frac{1}{F}\left(i-\frac{1}{2}\right)$. By selecting a departure $i$ and a speed of $s$, the head start time is adjusted accordingly. As pointed out by Noland and Small (1995), the safety margin is $t_{sm}=t_{hs}-E(t_d)$ where $E(t_d)$ is the expected delay. Some authors have studied travel time variability in the public transport modes as well. Bates et al. (2001) investigated how uncertainty in arrival times at the station for the service itself (i.e. if a bus arrives at the posted time or not) and trip time on-board the service affects user costs. However, they did not consider access time or scheduling costs, which we do in this paper. Moreover, we do not consider the additional uncertainty from variation in travel time onboard the public transport service. However, we do consider the effect of an unreliable arrival time at the station for the service.

Consider the definition of the safety margin $t_{sm}=t_{hs}-E(t_d)$. Let $T_{sm}$ be a stochastic realization from the distribution of $T_{sm}$ given by

$$T_{sm}=t_{hs}-(T_A-T_S) \tag{4}$$

Here, $T_A$ is the delay on the access time to the station and $T_S$ is the schedule deviation from the public transport service. Thus, the safety margin increases whenever $T_S>0$, as more time is available to reach the desired departure.

It is important to underline that departure number is understood as the $i$ th departure after the users desired departure time (from home). That is, with $i=1$, the $i$ th departure is 1/F(i-1/2) hours away from the desired departure time. With increased frequency of one departure per hour, the $i$ th departure is 1/(F+1)(i-1/2) hours away from the desired departure time. Thus, the time posted time of the $i$ th departures, changes when frequency increases. This is important to keep in mind when reading.

Another important point is that the probability of being late, depends on both speed ($s$), departure number ($i$) and frequency through $t_{hs}$. That is, $P_L(t_{hs}(s,F),i)$ is the probability that a stochastic realization of the travel time delay, will be larger than the head start, i.e., $t_{hs}<T_A-T_S$ (that the safety margin is too small) - $P_L(t_{hs}(s,F),i)=P(t_{hs}<T_A-T_S)$. For convenience, we write $P_L(s)$, but it is to be understood as a function of both speed, departure number and frequency (and trip length).

Fourth, accident costs are given as an increasing function of speed $A(s)$, which is not traditionally a part of any scheduling or reliability framework. Accident costs are evaluated at a rate $\pi$ of NOK/incident.

### 3.2. Analysis

We now turn to analyzing the model from a theoretical point of view, before parameterizing it and performing some numerical simulations. As previously mentioned, the traveler faces two principal decisions: (i) what speed to select for each departure and (ii) what departure to select. Thus, there are several questions that may be posed to gain any understanding as to the effect increasing frequency exerts on speed choice. It seems there is an intensive margin, i.e. choice of speed for a given departure, and an extensive margin, i.e., which departure to choose. Consequently, we need to consider what (i) influences the choice of departure, (ii) the choice of speed of a given departure and (iii) how choosing a different departure affects speed when frequency is increased. We will address these questions, starting with the optimal speed for a given departure.

#### 3.2.1. The intensive margin: the optimal speed for a given departure

The optimal speed for a given departure is found by setting $i$ to a fixed value and taking the derivative of equation (2) and setting it equal to zero, leading to the following equation:

$$\frac{\partial C(i,s)}{\partial s}=\alpha\frac{\partial t}{\partial s}+\frac{\theta}{F}\frac{\partial P_L(s)}{\partial s}+\pi\frac{\partial A(s)}{\partial s}=0 \tag{5}$$

The optimal speed for a given departure sets the sum of three components equal to zero: (i) the marginal reduction in travel time savings when speed is increased, which is positive, (ii) the reduction in delay costs stemming from a lower probability of arriving too late at the station and (iii) the increased accident costs. In appendix 1, we show that $\frac{\partial^2 C(i,s)}{\partial s^2}>0$, indicating a minimum.

Some preliminary intuition may be gained from this equation. First is seems reasonable to assume that increased speeds lead to a lower probability of being late, such that $\partial P_L(s)/\partial s<0$. Thus, $\partial t/\partial s<0$ and $\partial A(s)/\partial s>0$, we can infer that the probability of being late leads to choosing a higher speed for a given departure, as compared to the case in which there is no chance of being too late (i.e. $P_L(s)=0$). Moreover, it may seem that increasing the frequency corresponds to a lower speed, as the cost of being too late is lowered ($\theta/F$ falls).

To assess this claim, we may take the derivative of the first-order condition with respect to frequency. If this term is positive, the optimal speed will be reduced (as $\partial t/\partial s$ must become lower, i.e. "more negative").

$$\frac{\partial^2 C(i,s)}{\partial s\partial F}=\frac{\theta}{F}\left(\frac{\partial^2 P_L(s,i)}{\partial s\partial F}-\frac{1}{F}\frac{\partial P_L(s,i)}{\partial s}\right) \tag{6}$$

There are now two effects operating at the same time.[6] Firstly, increasing the frequency lowers the cost of being too late, as the headway is shorter, yielding a lower time until the next service arrives. This is the right-hand side inside the parenthesis. Secondly, the marginal effect of increasing speed is altered. The interpretation being the probability of reaching the first available departure is altered, such that one must maintain a higher speed in order to reach it.[7] Generally, this effect is positive on speed: as frequency is increased, the time until departure $i$ after the users' desired departure time is smaller, while the other effect is negative on speed This is also explained in appendix 2.

In short, one may summarize the intuition from the condition as

[6] Moreover, notice that scheduling costs do not affect the choice if speed at the intensive margin.

[7] As frequency increases from 1 to 2 departures per hour, the average time until the first departure goes from 30 to 15, as viewed from the desired departure time (from home) of the user.

follows: When frequencies are low, the probability of being too late is low (as headways are long), but the cost is high (for the same reason). As frequencies increase, the probability of being late increases (for a given departure - the term $\frac{\partial^2 P_L(s,i)}{\partial s \partial F}$), but the cost is lowered (the term $\frac{1}{F}\frac{\partial P_L(s,i)}{\partial s}$). The effect on speed for a given departure depends on the shape of the probability distribution, frequencies and trip lengths (through the head start $t_{hs}$). If frequency is very large, the effect becomes small. Moreover, the change is directly proportional to delay costs, $\theta$. Thus, a user who is more averse to being late, would increase the speed at any given departure number more, as opposed to one who is less averse, which seems logical.

The equation indicates that an increased frequency, for a given departure number $i$ after the users' desired departure time, will reduce the consequence of being late, but also affect the probability of being late, as there is less time to reach it. Which of these two effects dominate, will determine whether speeds will increase for a given departure number.

### 3.2.2. *The extensive margin: the effect on departure choice of increased frequency*

We now look at the choice of departure, and how this is affected by the frequency. This is interesting as selecting a later departure allows for a greater safety margin until departure. A higher safety margin may lead to lower speeds. Consequently, it is interesting to understand how decisions regarding choice of departure number is determined.

To reiterate: It is assumed that the user first estimates the optimal speeds for each available departure, and then selects the departure which has the lowest overall cost. We further assume that user considers all departures in the set $I = [1, i_{max}]$, where $i_{max}$ is a maximum "search length" regarding departure number one considers. This mean that the optimal departure number $i^*$, satisfies the following equation:

$$i^* = \underset{i \in I}{\operatorname{argmin}} C\left(i, s_i^*\right) \tag{7}$$

Consequently, we note that the optimal departure is a function of the optimal speed of each departure $i$; $s_i^*$. In turn, this speed depends on frequency, such that we may write $s_i^*(F)$. In order to evaluate the effect of increasing frequency on the choice of departure, we may first look at how the cost of a given departure $i$ is affected by an increase in frequency $F$. The cost of departure $i$ may be written as a function of frequency alone, when assuming users select the optimal speed for the given departure. That is, by inserting for $s_i^*(F)$ in $s$ in equation (2):

$$C(i,F) = \alpha * t(s^*(F)) + \frac{\omega}{F}\left(i - \frac{1}{2}\right) + \frac{\theta}{F} P_L(s^*(F), i) + \pi A(s^*(F)) \tag{8}$$

Taking the derivative of this function with respect to $F$ yields the following equation, after rearranging terms:

$$\frac{\partial C(i,F)}{\partial F} = \frac{\partial s^*}{\partial F}\left[\pi\frac{\partial A}{\partial s} + \alpha\frac{\partial t}{\partial s}\right] + \frac{\theta}{F}\left[\frac{\partial P_L}{\partial s}\frac{\partial s^*}{\partial F} - \frac{1}{F}P_L(s^*, i)\right] - \frac{\omega}{F^2}\left(i - \frac{1}{2}\right) \tag{9}$$

Here $\frac{\partial s^*}{\partial F}$ is the change in optimal speed resulting from an increase in frequency $F$ for departure number $i$. The term was discussed in the preceding section, and it was noted it may take either a positive or negative sign, depending on, among other things, the size of $F$. However, we do not need to know the sign of this function in order to gauge the effect of frequency on the cost of a given departure. Using the fact that the first order condition for optimal speed of a given departure holds, we may insert $-\frac{\theta}{F}\frac{\partial P_L}{\partial s} = \alpha\frac{\partial t}{\partial s} + \pi\frac{\partial A(s)}{\partial s}$ into the equation to get the desired derivative written in a more compact manner:

$$\frac{\partial C(i,F)}{\partial F} = \frac{1}{F^2}\left(\theta P_L(s^*, i) - \omega\left(i - \frac{1}{2}\right)\right) \tag{10}$$

We note that a higher frequency only makes the cost of departure $i$ rise when $\theta P_L(s^*) > \omega\left(i - \frac{1}{2}\right)$. That is, when the expected cost of being too late, $\theta P_L(s^*)$, is higher than the schedule time cost $\omega\left(i - \frac{1}{2}\right)$. The economic interpretation is as follows: When frequency increases, the safety margin to departure number $i$ (as viewed from the desired departure time of the user) grows smaller. We note that $t_{sm} = t_{hs} - E(t_d)$ with $t_{hs}$ is smaller (for fixed $i$) when the frequency increases (equation (3)). As $F$ increases, reaching the first service (as viewed from the desired departure time of the user) becomes more difficult. In turn, the increased probability of being too late raises costs. Secondly, an increased frequency means the scheduling cost for departure $i$ is reduced, which lowers its cost. In summation, the effect depends on parameters $(\theta, \omega)$, trip lengths, time variability and so on. However, for simplicity, assuming that $\theta = \omega$, costs increase if

$$P_L(s^*, i) > i - \frac{1}{2} \tag{11}$$

Thus, when the probability of reaching the departure is relatively low at the outset ($P_L(s^*, i)$ large), it is more likely that raising frequency will lead to higher costs. The reason is that safety margin/time to departure $i$ is small at the outset, and diminishing them even further by raising frequency (as viewed towards a specific departure number $i$), increases the cost of being too late more than the reduction from lowered scheduling costs.

However, as the user chooses among a set of available departures, it is not sufficient to only consider how the cost changes for one of them. What is needed, is an understanding of how the relative cost between each departure is affected, to assess whether a change will take place. As the choice of departure number is discrete in nature, a complete marginal analysis, as was performed for the intensive margin (speed) is precluded. However, some intuition may be gained by considering how a user that is indifferent between two consecutive departure numbers is affected when frequency is increased. That is, the cost of departure $i$ and $i+1$ is equal:

$$C(i+n, F) = C(i, F) \tag{12}$$

Consequently, a user will choose the later departure number $i+1$ if its cost is raised less quickly as compared to departure number $i$, when frequency is increased. In turn, this means that the following condition[8] must hold in order for the user to optimally switch to a later departure when frequency is increased:

$$\frac{\partial C(i+n, F)}{\partial F} < \frac{\partial C(i, F)}{\partial F} \tag{13}$$

Inserting for the derivative of $F$ on cost $C$ found earlier and rearranging terms to obtain a simpler expression, we find the following condition:

$$\omega < \theta[P_L(s^*(i)) - P_L(s^*(i+1))] \tag{14}$$

Consequently, if the above condition holds, a later departure is chosen, if not, the same departure number is maintained, when frequency is increased. The interpretation of the condition tells us that a later departure is chosen when frequency is increased if the reduction in expected cost from being too late (the right-hand side) stemming from a lower probability of lateness ($P_L(s^*(i)) > P_L(s^*(i+1))$), is larger than the increase in cost per unit of scheduling cost $\omega$. Thus, the user weights the extra scheduling cost against the reduction in costs of being too late. If the latter is weighted more heavily in relative terms ($\theta/\omega\uparrow$), it becomes more likely that users select a later departure number when frequency increases. Conversely, if scheduling costs are relatively more important than delay costs ($\theta/\omega\downarrow$), it becomes more likely that the departure number is maintained.

[8] Strictly speaking, we must also assume that no other departure is dominating $i+1$ and. $i$ However, for the sake of the argument, the simple condition will suffice.

If a later departure is selected, it seems reasonable that speeds are lowered. To gain some intuition regarding this question, we may first revisit the first order condition of optimal speed for a given departure (equation (5)). In this case, we assume that the frequency is constant. Then, changing to later departure (for a given frequency $F$) always leads to lower speed. That is, if $i\uparrow$ then $\frac{\partial P_L(s,i)}{\partial s}\downarrow$ and $\frac{\partial C(i,s)}{\partial s} > 0$, such that $s\downarrow$ in order for $\frac{\partial C(i,s)}{\partial s} = 0$, as $\frac{\partial t}{\partial s}$ is less than zero. Thus, $s^*(F, i+1) \le s^*(F, i), \forall i$. When the expected cost of arriving too late is low ($\frac{\partial P_L(s,i)}{\partial s} \approx 0$) or very high($\frac{\partial P_L(s,i)}{\partial s} \approx 1$),[9] thus, with a high frequency the speeds tend towards each other, i.e. $\lim_{F\to\infty \text{ or } 0} s^*(F, i+1) = s^*(F, i)$.

However, we cannot assume that frequency remains the same, thus the analysis needs to be extended. We have seen in equation (6) that the effect of frequency on the speed for a given departure is comprised of two opposing effects: The effect on the cost of being too late and the effect on the probability when changing speeds.

In order to assess if increased frequency leads to a lower speed when also selecting a later departure, some additional assumptions will need to be made. For example, we assume that a departure $i+1$ and $i$ has the same optimal speed for a given frequency $F$. As noted above, this happens when frequency is either very low or high. Revisiting the optimality condition for the speed of a given departure, we found that speeds decrease if $\partial C(i,s)/\partial s > 0$. Moreover, we found that, for a given departure number, speeds decrease with an increased frequency if $\partial C(i,s)/\partial F \partial s > 0$. Consequently, for the optimal speed of departure $i+1$ to be lower than $i$ when frequency increases, we must have that the two following conditions hold:

$$\frac{\partial C(i+1,s)}{\partial s \partial F} > \frac{\partial C(i,s)}{\partial s \partial F} \tag{15}$$

$$\frac{\partial C(i+1,s)}{\partial s \partial F} > 0 \tag{16}$$

First, the change in the first-order condition for optimal speed must be larger for departure $i+1$ as compared to departure $i$. This ensures that any adjustments of speeds in departure $i+1$ leads to a lower speed as compared to $i$. Secondly, as shown in equation (6), the marginal effect must be positive in order to lead to a lower speed.

We have seen that if a later departure is not selected, speeds may or may not increase depending on equation (6). Moreover, if a later departure is selected, speeds may or may not be lowered, depending on equation equations 15 and 16. Thus, there is no clear-cut answer as to the effect of changing departure frequency elicits on speeds chosen on access trips to a transit station. The effect depends on weighting of the different cost elements, trip length, variability in trip time and so on. We may only derive some conditions theoretically that provide some indications as to the aggregate effect. Next, we conduct a simulation analysis of the model to better understand the effect of frequency on optimal speeds.

## 4. Simulation

The analytical investigation of the model revealed that there is no clear-cut answer as to how increased frequency influences speeds chosen by drivers. As pointed out previously, the effect is likely dependent on many different factors, such as weighting of the different cost components, trip length, trip time variability and frequency itself. To better understand the effect, we conduct a simulation analysis, where parameters and variables are altered, and optimal speeds are estimated for each set of values. The purpose of the analysis is to gauge the effect within a large set of possible assumptions, such that valid inferences may be drawn, and our results are, hopefully, not artefacts of the specific assumptions made. In the following section, we describe the simulation methodology, assumptions, and results.

### 4.1. Methodology

In the simulations, we use the theoretical model outlined in the previous section to estimate optimal speeds when assumptions are being varied. We now describe the simulation workflow and assumptions regarding variation in model parameters.

The simulation proceeds in the following manner:

1. For $N$ simulations rounds, the following steps are completed:
   a. Draw a "driver" $D_k$ consisting of a tuple of assumptions, $r_k = (\alpha_k, \beta_k, \theta_k, l_k)$
   b. Estimate $\pi$ (this step is explained in more detail below)
   c. For each level of frequency $[1, F_{max}]$ do the following:
      i. For each departure number in $I = [1, i_{max}]$, calculate the optimal speed $s^*$ by minimizing the first-order condition in equation (6)
      ii. Find the optimal departure as $i^* = \underset{i\in I}{\operatorname{argmin}} C(i, s_i^*)$
      ii i.Report optimal speed as $s_{i^*}^*$ and departure number $i^*$ for a given frequency $F$

The simulation outputs, for each driver $D_k$ a list of optimal speeds and departure numbers for each level of frequency $F$. We may then estimate the mean, optimal speed and departure number for each simulated driver, along with its variation. The number $\pi$ is estimated in each simulation round to calibrate the model. That is, for a target speed $s$, $\pi_k$ is estimated such that the optimal speed (in the absence of any risk of being too late) is equal to the target speed (e.g. a speed limit)[10]. In the theoretical model, $\pi$ illustrates the cost of accidents. In reality, there are most likely several other concerns than accidents which regulate the driver's speeds, such as comfort, the risk of being fined by police. When the model is calibrated, $\pi$ will most likely also include several other concerns, other than accidents, which incentivize the driver to choose a given speed. Consequently, in the simulations, it may no longer be viewed purely as an accident cost, but rather as a calibration parameter.

In the optimization of speeds given a departure, bounds on the minimum and maximum allowed speed [km/h] is given as $[LB, 120]$, where $LB$ is defined by $l/\frac{1}{F}\left(i-\frac{1}{2}\right)$. This restriction ensures that no driver selects a speed such that the "head start" margin (equation (3)) is negative, i.e., one is expected to be late on average. Last, we set $i_{max} = 12$, i.e. one maximally considers the twelfth departures after the desired departure time.

### 4.2. Assumptions

The assumptions regarding parameter values made in the simulations are detailed below. Table 1 shows the variables that are altered in each simulation run, their unit and from what kind of distribution they are drawn.

To estimate the value of the time component, a base value of time is multiplied by a set of weights, which indicate the value of a minute of a given time component, relative to the in-vehicle time. The value of the access time (AC) weight is given a lognormal distribution with mean 1.0 based on the assumption that all other costs are measured relative to this, with the standard deviation set as an own assumption. The mean value of delay time (DT) weight is set at 2.5 based in the Norwegian value of time study (Flügel et al., 2020), with a standard deviation of 0.5, based on an own assumption.

[9] If it is highly likely one reaches the departure, irrespective of speed. That is, the headway is very long.

[10] $\pi_k$ is estimated by using the following equation. $\pi_k = \alpha_k \frac{l_k}{s^2} / \frac{\partial A(s)}{\partial s}$

**Table 1**
Simulation parameters and assumptions.

| Variable | Symbol | Unit | Distribution | Mean | Standard deviation | Source |
|---|---|---|---|---|---|---|
| Value of access time | α | NOK/hour | Lognormal | 1.0 * VOT | 0.5 | Flügel et al. (2020) |
| Value of early schedule time | β | NOK/hour | Lognormal | 0.5 * VOT | 0.15 | Fosgerau (2009) |
| Value of early versus late schedule time | a | Ratio | Fixed | 0.25 | 0 | Fosgerau (2009) |
| Value of delayed time | θ | NOK/hour | Lognormal | 2.5 * VOT | 0.5 | Flügel et al. (2020) |
| Trip length | l | km/trip | Emprical distribution | 8.4 | 6.5 | National travel survey 2018/2019 |
| Value of time | VOT | NOK/hour | Fixed | 167 | 0 | Flügel et al. (2020) |

We use the relationship between the cost of being early ($\beta$) and late ($\gamma$) as given by $\beta = a\gamma$. The value of $a$ time is based upon Fosgerau (2009) who uses[11] $\beta = 0.5$ and $\gamma = 2$, yielding $a = 0.25$. Inserting $\beta = a\gamma$ into the expression for $\omega$, to find such that $\omega = \beta/(1 + a)$. The underlying assumption is then a value of $a$ that is constant in the simulations. According to Fosgerau et al. (2008), $a < 1$ and relatively low, such that this is an innocuous assumption. Table 2 shows the values of uncertainty parameters.

It is important to note that the model is parametrized by using assumptions from different studies. The result of such studies may themselves be subject to variations depending on the contexts in which they are performed. Consequently, even though each parameter makes sense in isolation, they may not do so when used in conjunction with one another.[12] To reduce such risks, we have tried to (i) include as few parameters as is possible, while keeping the model relatively realistic and (ii) not assuming any correlation between the values draw in each simulation round. That is, we do not place any assumption of whether a high value of delayed time also corresponds to a high value of schedule time.[13]

Trip lengths are drawn from the empirical distribution of the Norwegian national travel survey for 2018/2019 (Epinion, 2019). We restricted the sample to all trips completed as a car driver between 5 and 70 km of length, giving at total of 68 168 trips to draw from. We study access trips to a public transport station. As the trips sampled were undertaken with car in its entirety, we set the mean distance in the simulation to half of the mean, to reflect the assumption that access trips are shorter on average than trips wholly undertaken with car. In the simulation, a random number, $n$, between 1 and 68 168 is drawn, and the $n$-th trip is extracted from the empirical distribution. The mean trip distance is 8.4 km with a standard deviation of 6.5 km. The value of time (NOK/hour) is fixed as 167 2018 NOK/hour from Flügel et al. (2020). Changing the value of time per hour affects all cost components proportionally, such that effects are not altered by changing it.

The speed limit is set to 70 km/h. Changing this assumption will only shift the results to a higher base speed, as the model is calibrated to the speed limit.

The access/travel time variability distribution is assumed to be normal, such that $P_L$ is the cumulative distribution function of the normal distribution. The distribution is parameterized with mean delay ($\mu_A$) and standard deviation ($\sigma_A$), such that $\mu_A = \sigma_A = b_0 + b_1 {*} l$. The constant $b_0$ is set to 5 min, whereas the parameter $b_1$ is varied between 0.1 and 0.3. The lowest value of ($b_1 = 0.1$) yields a mean delay of 7 min on a 20-km-long trip and ($b_1 = 0.3$) yields a mean delay of 11 min. The parameters reflect the assumption of increasing variability with distance.

**Table 2**
Assumptions used in the simulation runs.

| Simulation # | Name | | | | |
|---|---|---|---|---|---|
| 1 | Low uncertainty | 0.1 | 1 | 0.1 | 0.5 |
| 2 | Medium uncertainty | 0.5 | 1 | 0.3 | 0.5 |
| 3 | High uncertainty - High correlation | 0.9 | 5 | 0.3 | 0.5 |
| 4 | Low uncertainty - High schedule cost | 0.1 | 1 | 0.1 | 1.5 |

The uncertainty in the public transport mode is given by the variation in departure time from the station. That is, the public transport mode sometimes arrives a bit early, sometimes a bit late, and does not "hold" until the posted departure time. We assume a mean variation of $\mu_P$ and standard deviation of $\sigma_P$, where $\mu_P = \sigma_P$ is varied at 1 and 5.

As given in equation (4), the safety margin given uncertainty in both access ($T_A$) and departure time of the public transport ($T_S$) is given by:

$$T_{sm} = t_{hs} - (T_A - T_S) \tag{17}$$

Thus, the mean delay is $E(t_d) = E(T_A - T_S) = \mu_A - \mu_S = b_0 + b_1 {*} l - \mu_S$ and its standard deviation is $SD = \sqrt{\sigma_A^2 + \sigma_S^2 + 2\rho\sigma_A\sigma_S}$. In the numerical simulations we vary the correlation between low ($\rho = 0.1$) and high ($\rho = 0.9$).

The probability of accidents $A(s)$, is found by using the concept of a survival function.[14] If $h(s)$ is the unit rate of accidents per kilometer, assuming an exponential survival function, the probability of having an accident can be written as

$$A(s) = 1 - \exp(-h(s) \times l) \tag{18}$$

where $A(s)$ is the probability of having an accident within $l$ kilometers, and $\exp(-h(s) \times l)$ is the cumulative density function (cdf) of the exponential distribution.

A popular model of the accident rate per kilometer driven is the so-called power model (Nillsson, 2004; Cameron and Elvik, 2010; Elvik et al., 2019), which relates the accident rate for per kilometer to the speed level:

$$h(s) = a_0 (s/s_0)^{a_1} \tag{19}$$

where $a_{0,k}$ is the accident base rate per kilometer, and $s_0$ is a «base speed». As a simplification, we set the base speed equal to the speed limit in our model. Thus, we assume that the user will always vary changes in the risk with respect to the given speed limit. Further, because one parameter is used for all speed limits, we effectively assume that the base rate is equal over a certain range of speeds. The base rate of accidents per kilometer driven, $a_0 = 2.1 {*} 10^{-6}$, is estimated using data provided by Nilsson (2004) for the category "serious accidents". To keep the model parsimonious, we use only a single model of accidents, whereas there in reality may be other relevant categories of seriousness. As we aim to describe the dynamics, we contend this approach is sufficient, as the model is calibrated to match observed speed levels. The elasticity of the accidents with respect to the speed level, $a_1$, is set and 2.0 for serious accidents, as gathered from Cameron and Elvik (2010).

The trip time function $t(s)$ is defined as $t(s) = l/s$.

[11] Supressing the scaling of VOT.

[12] We would like to thank an anonymous reviewer for pointing this out to us.

[13] An exception is here $\gamma$ and $\beta$ which are assumed to be perfectly correlated through $a\gamma = \beta$. That is, only assumptions on the level, and not ratio between them is made to keep the model somewhat parsimonious.

[14] See Kleinbaum and Klein (2010) for reference.

## 4.3. Scenarios

We perform a total of 4 different simulation runs using different parameter values as given below.

The purpose is to assess how different levels of uncertainty and perceptions of scheduling versus delay cost affect the results. Low uncertainty sets the variability in the public transport mode to 1 min and the length-dependent coefficient of uncertainty in the car mode to 0.1, meaning a 10-km-long trip will have mean variability of 5 + 0.1*10 - 1 = 5 min. Medium uncertainty sets the correlation between variability in travel time and public transport to 0.5, which increases the total variance of the variability (i.e., the delays are more variable). Then we increase the correlation between variability of the access time and departure time of the public transport mode to a high level of 0.9. Last, we increase the value of early schedule delay cost ($\beta$) relative to its base value. The purpose of this exercise is to evaluate the case were users experience deviances from their desired departure times as more costly. For example, during the morning peak, one may have a strong preference for not starting at work after a given time point. At the same time, the preference might be quite strong for not being too early as well. The assumption underlying scenario 4, is that both the lateness and earliness penalty is increased, as the parameter $\alpha$ is kept constant. At the same time, one could argue that $\theta$ should be raised as well. However, there is a difference between planning that one arrives a bit later than preferred, and arriving later than planned, which is the essential difference between $\theta$ and $\omega$.

# 5. Simulation results

We now present the simulation results of the four scenarios. We start by discussing the optimal speed in all four scenarios, before presenting it subdivided by trip length. Last, we discuss how the "head-start" margin and departure number is affected by frequency.

## 5.1. Optimal speed

The optimal speed for in the four scenarios is given in Fig. 1. For each level of frequency, the distribution of optimal speeds is given. The main result is that speeds tend to increase when moving from a low level of frequency, before being reduced after a certain level. The results also reveal that the magnitude of the effect is highly dependent upon the assumed scheduling cost, and less dependent upon the variability of time. For scenario 1–3, we assume the lowest schedule cost with $\beta = 0.5$, whereas in scenario 4, we assume a higher value of $\beta = 1.5$. When assuming a low schedule cost, there is only a moderate incentive to increase speed, when frequency increases from 1 to 2 departures per hour. When assuming a higher level of schedule cost, increased speeds are more pervasive, as one is less willing to choose a later departure, as this would increase the schedule costs. Moreover, it is worth pointing out that the marginal effect on speed is dependent upon the frequency level. Increasing frequency from a moderate level (i.e., 2–4 departures per hour, depending on assumptions) may reduce the incentive to speed. Increasing frequency from a low level (i.e., 1 departure per hour), may induce a portion of drivers to increase in speeds observed. This portion heavily depends on the schedule cost parameters of drivers (e.g. observe the difference in distributions between scenario 1 and 4).

This result is driven by two factors, as explained in the theoretical section. Starting with equation (6), firstly, increasing the frequency lowers the consequence of being too late, as the headway is smaller. All else equal, this should result in smaller speeds. Secondly, the probability of reaching the first available departure is altered, such that one has to maintain a higher speed in order to reach it (as frequency is increased, smaller headway also means smaller, possible safety margins for each departure number). As frequency is increased from a low level, one may save schedule time costs by trying to take benefit of the smaller headway, given that the ratio of waiting versus delay costs ($\theta/\omega$) is so that the increase in probability of being too late ($P_L(s^*(i)) - P_L(s^*(i+1))$) is not too large, evaluated at the optimal speed of each departure ($s^*(i)$ $s^*(i+1)$). At lower levels of frequency, the reduction in headway is the largest, offering the largest advantage of maintaining a smaller safety margin. The change in safety margin depends strongly on trip length, which renders the effect sensitive to that parameter.

As frequency increases, the marginal reduction in headway is lower. Thus, the gain from exploiting the reduced waiting time by trying to reach a departure closer in time (with higher speed) is lower. Consequently, a later departure is more likely to be selected. For example, with one departure per hour, the time until first departure (TUFD) is 30 min (on average, as viewed from the desired departure times of users). With 2 departure per hour TUFD = 15, with 3 TUFD = 10, and with 6 TUFD = 5. Consequently, if trying to reach the first departure available (to always take advantage of the lower schedule time, by not postponing) becomes increasingly more difficult. At the same time, this means the condition for selecting a later departure, is more likely to hold as $P_L(s^*(i)) - P_L(s^*(i+n))$ will become larger with $n$. That is, at TUFD = 15, the difference between the first and second departure $P_L(s^*(1)) - P_L(s^*(2))$, will likely be smaller than with TUFD = 5 (a difference of (15 + 30) – 30 = 15 min versus 5 + 10–10 = 5 min by choosing a later departure). With a higher frequency, the same gain in schedule cost reduction may be achieved by selecting a later departure, and a lower speed as compared to when frequency is low. As a later departure is chosen, speeds may or may not become lower, depending on the conditions on equations (15) and (16), which was not determinable from the theoretical model alone. Thus, it seems that choosing a later departure reduces speeds at higher levels of frequency. As predicted by equation (6), the effect of frequency on speed diminishes (the derivative $\frac{\partial^2 C(i,s)}{\partial F \partial s}$ tends to zero for increased $F$) with increased frequency.

The scenarios (1–3) were subject to an increasing level of variation on the travel and public transport arrival time. The effect indicates that increasing uncertainty does not have a large impact on our results. However, at the highest level of uncertainty (scenario 3), there is a slight increase in the number of drivers speeding for low levels of frequency. A higher level of uncertainty will increase the expected costs of being too late. That is, a higher standard deviation will yield a higher $P_L$, for all levels of $s$. For any given departure, the expected cost of being too late will increase, and a higher speed will have to be, maintained in order for the first order condition to hold, given each departure, increasing total cost, such that switching to a later departure becomes more likely.

Moreover, the simulations show the response at lower levels of frequency is heterogenous. Some likely maintain a lower speed, while other engage in a significantly higher speed. If some users value waiting time costs more than delay costs ($\theta/\omega\uparrow$) they may be willing to select a higher speed, or if the trip length is not too long, such that a given departure may be reached with a sufficiently high probability. At longer trips, the probability of reaching a point in time will be lower for the same level of speed, as compared to a shorter one.

In Fig. 2, we have subdivided the results from scenario 1 into trips longer and shorter than 10 km.[15] For trips longer than 10 km, the distributions are slightly more skewed to left (towards lower speeds) as compared to trips shorter than 10 km for higher levels of frequency. As the probability of being too late will be higher for longer trips, fewer drivers may aim for the earliest departure, as the probability of reaching it is too low as compared to the cost. At the same time, speeds at 1 departure per hour is higher for longer trips as compared to shorter ones. With more distance to cover, some trips will find it optimal to engage in higher speeds necessary to avoid the longer waiting time of selecting a later departure. As the risk of accidents is proportional to trip lengths, long trips to low headway services are likely to benefit the most from an increase in frequency. With shorter trips, the distributions are more

[15] 10 km correspond to the 3 quantile of the length distribution we used.

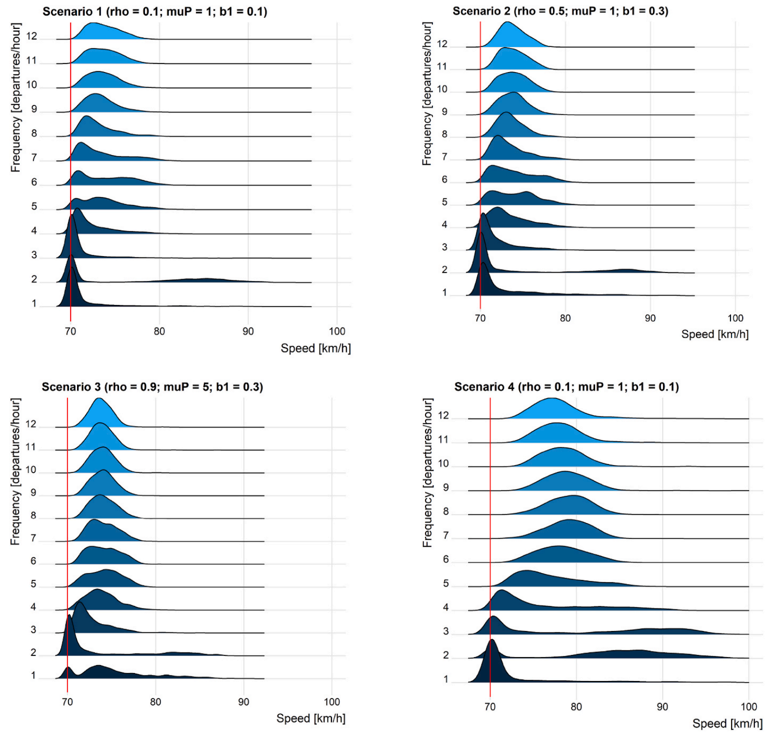



**Fig. 1.** Simulated optimal speeds [km/hour] at different levels of frequency in the four simulation scenarios.

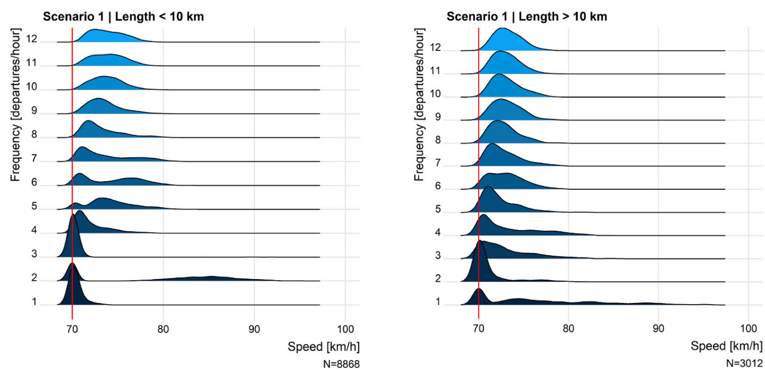



**Fig. 2.** Simulated optimal speeds [km/hour] at different levels of frequency for trips shorter and longer than 10 km.

right-skewed (higher speeds) as the probability of being too late is lower. However, the effects depend on frequency level.

### *5.2. Variability*

The variability of the simulated speeds may be assessed by observing the variation in the distribution of speeds for each of the scenario and frequency level in Fig. 1. One can observe two major results.

At low levels of frequency, the variability is the highest, and the distribution tends towards bimodalilty. In practical terms, this indicates that there are two groups: One which does not speed and one which speed by a quite large margin. At lower levels of frequency, there may be two sources of risk. First of all, the drivers who speed will have a higher risk of experiencing an accident. Secondly, the ones who do not speed, may also have a higher accident risk as the speeds chosen are quite heterogenous. This may increase the risk of vehicle-to-vehicle accidents.

At higher levels of frequency, the variability is reduced, and the distribution becomes unimodal. This indicates that a higher level of frequency reduces the heterogeneity in speeding behavior. Consequently, at higher levels of frequency, one does not observe the extreme speeds, and the variation among users is lower. Both of these factors may reduce the overall accident risk.

### 5.3. Head start and departure number

Fig. 3 shows the distribution of optimal time to departure or "head start", $t_{hs}$, as given by equation (3) and optimal departure number, $i^*$, for scenario 1, given the level of frequency. At the lowest level of frequency, the average margin is the highest, at 34.56 min, converging to 19.48 min at 12 departures per hour. Consequently, the margin approaches a value in which the cost of selecting an earlier departure is higher than the savings in waiting time costs. The higher the frequency, the more likely it becomes that a later departure is chosen, with the average approaching the fourth departure at a frequency of 12 departures per hour. The effect on optimal departure number is strongly dependent on the total trip length. Fig. 4 shows the optimal departure number for trips shorter and longer than 10 km. For trips shorter than 10 km, the average departure number is lower than for longer trips. It is important here to remind the reader that no "boarding time" is included in the analysis. Trips longer than 10 km, almost all select at least the second departure after the three or more departures are available per hour.

In Fig. 5, we compare the optimal departure number in scenario 1 with "standard" scheduling cost to the increased cost level in scenario 4 ($\beta\uparrow$). A higher scheduling cost has the effect that drivers tend to not select a later departure. The condition for when it is optimal to select a later departure is given by $\omega < \theta(P_L(s^*(i)) - P_L(s^*(i+n)))$. Thus, if $\omega$ is increased relative to the cost of being too late, $\theta$, it becomes more likely that one does not select later departures when frequency increases. In turn, this affects the speed as a higher one needs to be maintained to reduce the probability of being too late. Fig. 5 reveals that an increase in the scheduling costs, relative to the delay cost, do in fact prompt drivers to select earlier departures, on average, as compared to the case when scheduling costs are relatively low as compared to delay cost. That is, when one is less willing to accept a longer schedule cost to reduce the probability of being too late, earlier departures are selected.

### 5.4. Discussion

The simulations performed indicates that frequency may influence the optimal speed chosen by travelers connecting to a public transport mode. As frequency increases, two effects are present. Firstly, the consequence of being too late is lowered, which isolated leads to lower speeds. Secondly, one may obtain a lower waiting time, by trying to reach the first available departure. When frequencies are low, the probability of being too late is low (as headways are long), but the cost is high (for the same reason). Thus, some users speed to avoid being too late, but also to potentially lower their schedule cost by reaching an earlier departure.

As frequency increase, the marginal reduction in waiting time is reduced, which again lowers the incentive for trying to reach the first available departure, as the schedule time saved becomes smaller. Moreover, the probability of being too late for the first departure is higher, but the consequence is lower (as headways are shorter). In turn, this leads to a lower effect on speed at higher frequencies, but still larger than at the "base level" of one departure per hour, but possibly smaller than a "mid-range" level of departures per hour (2–4).

The actual effect observed among road users is likely to vary significantly depending on trip length and scheduling costs. The simulation has revealed that, using standard assumption on the weighting of scheduling costs, only increases in frequency from lower levels may induce drivers to significantly alter their speed, and perhaps, only a minor fraction. However, for users who perceive scheduling costs as high compared to the cost of being delayed, a significant speed incentive may be present at several levels of frequency. Thus, it is possible that economic costs are affected by an increase in frequency through accidents, but the extent to which this occurs is uncertain, as it depends on parameters expected to vary from case to case.

The marginal effect of increasing frequency was found to differ depending on the initial level of frequency. For low levels of frequency at the outset, it is more likely that increases will raise speeds. From moderate levels of frequency, it is more likely that they will be reduced. Consequently, the effect is also expected to vary according to current service levels at a service. This highlights the fact that the net effect of frequency on speeding behavior may be highly uncertain, but present in one way or another (i.e., *an* effect is likely not absent), and may vary from case to case.

## 6. Conclusion

Accidents are costly to society and their causes are of interest such that effective interventions may be designed to limit their extent. The relationship between accidents and speed is thoroughly established in the literature, and many governments take action to reduce the speed of road users. Many studies focus on interventions in the road-user system, but there is less research on the relationship between public transport services and accidents. In this paper, we have studied a specific aspect of public transport; whether departure frequency may influence the speed chosen by drivers trying to connect to a public transport service via a transfer. Such a topic is of interest to policy makers, as changing the frequency in public transport services may change the costs that society incurs through accidents. Thus, these findings are relevant for policy makers engaged in the field of public transportation.

To address the present question, we have extended the existing literature on speed selection, by developing a two-stage model where users select both departure and speed, in contrast to the one-level models typically applied (see, e.g., Jørgensen and Sandberg-Hanssen, 2019). We used this model in two ways. First, we performed a

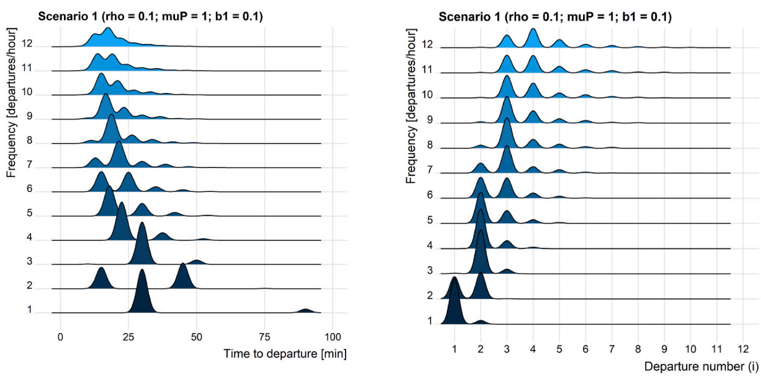



**Fig. 3.** *Simulated optimal time to departure ($t_{hs}$) [min] and optimal departure number ($i^*$) in scenario 1.*

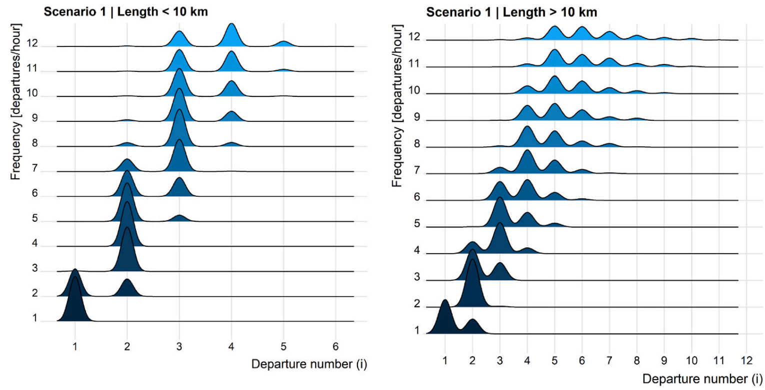



**Fig. 4.** *Simulated optimal departure number ($i^*$) in scenario 1 for trips shorter and longer than 10 km.*

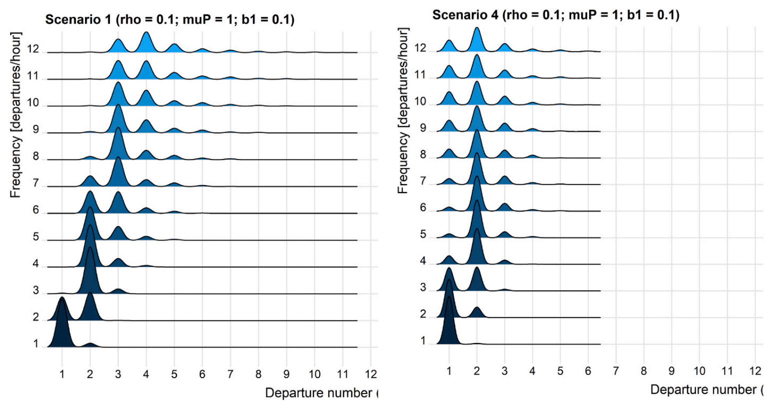



**Fig. 5.** *Simulated optimal departure number ($i^*$) in scenario 1 and 4. Scheduling costs are lower in scenario 1 ($\beta = 1/2$) as compared to scenario 4 ($\beta = 3/2$).*

theoretical analysis of the model's mechanics. We found that several factors influence the speed selection, such as the relative weighing of schedule cost versus delay time, trip length and the frequency itself. Moreover, we established conditions for when, and in what direction, speeds are likely to change.

As the analysis showed the effect to be highly dependent upon several factors that are expected to vary between individuals (such as weighting of different cost components) and between separate trips (such as trip length), a simulation study was undertaken, where assumptions were varied to obtain a range of possible outcomes. It was shown that increasing frequency from a low level may induce higher speeds for some users, as they aim to save schedule time by catching the first available departure. At higher level of frequencies, there is incrementally less waiting time to save, reducing the effect. Consequently, the paper has established a theoretical case for why frequency may affect speeds, and shown that (i) the extent to which this occurs is expected to vary greatly and (ii) the marginal effect is dependent upon the initial level of frequency. Further, with lower levels of frequency, the variation in speed is expected to be larger, as compared to higher levels of frequency. Thus, at lower levels of frequency, the risk of vehicle-to-vehicle accidents may be higher due to heterogenous speed levels.

Our results indicate that increasing the frequency of a public transport service may induce drivers to engage in more risky behavior by selecting higher speed when increasing frequency from a lower level. Policy makers concerned with designing appropriate service levels, should consider the possibly increased accident costs when upgrading the service level at a transfer connection and seek to improve the risk perceptions of the users – in the case where accident externalities are present, that are not internalized by the drivers. Moreover, they should also consider whether accident costs may be reduced if frequency is increased from moderate levels. Consequently, our study provides theoretical arguments for why safety considerations may be relevant for the optimization of service levels from an economics perspective. However, recommendations on changing, or not changing, the frequency cannot be made solely on the basis of our results. Policy makers should include the cost of increased frequency and valuation of a reduction in accidents, and then compare this to other possible safety policies, before making a recommendation. However, our results indicate that such analyses may be relevant to perform when the frequency is changed, which is important.

As with any study, ours remains a simplification of reality with intrinsic limitations that are vital to discuss in a scientific context. Some of the possible limitations of our study are as follows:

- **Uniformly distributed departure times**: The desired departure times could peak at a certain time point within a period. For example, if all users start work at 08:00 and they require on average half an hour to reach it, the desired departure times would be quite different. If the assumption is violated, our cost formulation may become less relevant at higher levels of frequency.

- **The aggregate effects are sensitive to assumptions**: Analyzing the model results by splitting the sample into long and short trips revealed trip lengths to be of significance for the results. The motivation to speed was higher for short trips, as it is then more likely that increasing speeds actually reduce the probability of being too late. Thus, the effects observed will vary with the average trip length and may very well be quite small (or nil) if one considers longer trips. Moreover, the valuation of scheduling versus delay costs plays an important role in determining the actual effect. When the ratio $\theta/\omega$ is low, it is more likely that speeds will increase from a low level of frequency. When it is high, it is not as likely. Thus, the actual effect observed will also depend on the ratio $\theta/\omega$ of road users at different times and locations.
- **Generalizability**: We have used a combination of theoretical arguments and simulation in this paper. Such methods require assumptions to provide a conclusion, which may or may not be justified. First, we assume a sufficient description of the driver's decision-making progress has been established. Secondly, using different parameters, especially behavioral ones, in a model simulation does not guarantee than their values make sense when taken together, even though they do so individually.[16]

Consequently, our results should not be viewed, in any way, as a definite answer to the questions posed but rather a note of consideration to which mechanisms could be in play under reasonable assumptions on behavior and the physical environment that a user might face.

Finally, some possible avenues of further research should be mentioned. Transport planners may use several different tools to limit the incentive for risk taking behavior. For example, increasing the user's risk-perception or employing a price mechanism that makes the users internalize any external accident costs they impose on others from speeding may be an interesting extension. Moreover, our results may be of use if empirical studies on the subject are to be designed. With the effect expected to vary, one may obtain different answers from different studies depending upon the context. Our paper possibly enables a framework, in which to analyze such results.

## Conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments & funding sources

We would like to thank James Odeck, Kristine Wika Haraldsen and Kristoffer Berg for constructive feedback on this article and Ingunn O. Ellis for data extraction from the National Travel Survey. We would like to thank the Norwegian Public Roads Administration (grant 19/145723-2), Oslo Norway and Urbanet analyse, Oslo, Norway for funding this study.

## Appendix 1. Second order condition

The first order condition of optimal speed for a given departure reads as follows:

$$\frac{\partial C(i,s)}{\partial s} = \alpha\frac{\partial t}{\partial s} + \frac{\theta}{F}\frac{\partial P_L(s)}{\partial s} + \pi\frac{\partial A(s)}{\partial s} = 0 \tag{A1}$$

To establish a minimum, we must have that $\partial^2 C(i,s)/\partial s^2 > 0$. It is reasonable to assume that $\partial^2 t/\partial s^2 > 0$ and that, at least, $\partial^2 A(s)/\partial s^2 \geq 0$ (accident probability does not decrease on the margin with increased speed). Substituting in head start time for a given departure, $t_{hs}(s)$), we have that $P_L(s) = 1 - \Phi(t_{hs}(s))$, where $\Phi$ is the normal, cumulative distribution function. Thus, $\frac{\partial P_L(s)}{\partial s} = -\varphi(t_{hs}(s))\frac{\partial t_{hs}(s)}{\partial s}$. The first factor in this product is negative and the second is positive ($\frac{\partial t_{hs}(s)}{\partial s} = l/s^2$). The second derivative is given by the following equation:

$$\frac{\partial^2 P_L(s)}{\partial s^2} = \frac{\partial}{\partial s}\left[-\varphi(t_{hs}(s))\frac{\partial t_{hs}(s)}{\partial s}\right] \tag{A2}$$

Which turns into:

$$\frac{\partial^2 P_L(s)}{\partial s^2} = -\left[\frac{\partial\varphi(t_{hs}(s))}{\partial t_{hs}}\left(\frac{\partial t_{hs}(s)}{\partial s}\right)^2 + \varphi(t_{hs}(s))\frac{\partial^2 t_{hs}(s)}{\partial s^2}\right] \tag{A3}$$

Generally, the products of the terms inside the parenthesis are negative, given that the user does not choose $t_{hs} < 0$ (which we have assumed to be true in the model). That is, $\frac{\partial\varphi(t_{hs}(s))}{\partial t_{hs}} < 0$ (for a normal probability distribution) and $\frac{\partial^2 t_{hs}(s)}{\partial s^2} < 0$, while the other product terms are trivially positive. Thus, $\frac{\partial^2 P_L(s)}{\partial s^2} > 0$, and the sign of the second derivative for the optimization problem ensures a local minimum.

## Appendix 2. The sign of equation (6)

The sign of the condition in equation (6) is found by taking the following derivative (substituting in head start time for a given departure, $t_{hs}(s)$):

$$\frac{\partial^2 P_L(t_{hs}(s))}{\partial s\partial F} = \frac{\partial}{\partial s}\left[-\varphi(t_{hs}(s))\frac{\partial t_{hs}(s)}{\partial s}\right] \tag{A4}$$

$$= -\left[\frac{\partial\varphi(t_{hs}(s))}{\partial t_{hs}} * \frac{\partial t_{hs}(s)}{\partial s} * \frac{\partial t_{hs}(s)}{\partial F} + -\varphi(t_{hs}(s)) * \frac{\partial^2 t_{hs}(s)}{\partial s\partial F}\right] < 0 \tag{A5}$$

[16] We would like to thank an anonymous reviewer for pointing this out to us.

By definition $\frac{\partial^2 t_{hs}(s)}{\partial s \partial F} = 0$, (see equation (3) to verify this). Further, we have that $\frac{\partial \varphi(t_{hs}(s))}{\partial t_{hs}} < 0$, $\frac{\partial t_{hs}(s)}{\partial s} > 0$ and $\frac{\partial t_{hs}(s)}{\partial F} < 0$, such that $\frac{\partial^2 P_L(t_{hs}(s))}{\partial s \partial F} < 0$ (as $-\ [- * + * -] = -$). Further, $-\frac{1}{F}\frac{\partial P_L(s)}{\partial s} = \frac{1}{F}\varphi(t_{hs}(s))\frac{\partial t_{hs}(s)}{\partial s} > 0$ (see appendix 1).

## Appendix 3. Derivatives of the first order condition for optimal speed of a given departure

In this appendix, we provide the explicit functional forms that constitute the first order condition of optimal speeds for a given departure number. The sum of these functions is minimized numerically using Brent's method (Brent, 1971) implemented in the Scipy Library (Virtanen et al., 2020).

Accidents:

$$\frac{\partial A(s)}{\partial s} = \exp[-h(s) \times l] * h(s) \times l * a_0 * a_1 * \frac{1}{s_0} * \left(\frac{s}{s_0}\right)^{a_1 - 1} \tag{A6}$$

Time:

$$\frac{\partial t(s)}{\partial s} = -\frac{t}{s^2} \tag{A7}$$

Cost of being too late:

$$\frac{\partial P_L(s)}{\partial s} = -\varphi(t_{hs}, \mu, \sigma) * \frac{t}{s^2} \tag{A8}$$

## References


Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: a review. Accid. Anal. Prev. 38 (2), 215–224.
Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. Transport. Res. E Logist. Transport. Rev. 37 (2–3), 191–229.
Blomquist, G., 1986. A utility maximization model of driver traffic safety behavior. Accid. Anal. Prev. 18 (5), 371–375.
Brent, R.P., 1971. An algorithm with guaranteed convergence for finding a zero of a function. Comput. J. 14 (4), 422–425.
Börjesson, M., Fung, C.M., Proost, S., 2017. Optimal prices and frequencies for buses in Stockholm. Econ. Transport. 9, 20–36.
Börjesson, M., Fung, C.M., Proost, S., Yan, Z., 2018. Do buses hinder cyclists or is it the other way around? Optimal bus fares, bus stops and cycling tolls. Transport. Res. Pol. Pract. 111, 326–346.
Cameron, M.H., Elvik, R., 2010. Nilsson's Power Model connecting speed and road trauma: Applicability by road type and alternative models for urban roads. Accid. Anal. Prev. 42 (6), 1908–1915.
Carrion, C., Levinson, D., 2012. Value of travel time reliability: a review of current evidence. Transport. Res. Pol. Pract. 46 (4), 720–741.
de Palma, A., Lindsey, R., 2001. Optimal timetables for public transportation. Transp. Res. Part B Methodol. 35 (8), 789–813.
Epinion, 2019. Dokumentasjonsrapport: Den Nasjonale Reisevaneundersøkelsen 2016-2019. English: Documentation Report: the National Norwegian Travel Survey 2016-2019.
Elvik, R., Vadeby, A., Hels, T., van Schagen, I., 2019. Updated estimates of the relationship between speed and road safety at the aggregate and individual levels. Accid. Anal. Prev. 123, 114–122.
Elvik, R., Vaa, T., Hoye, A., Sorensen, M. (Eds.), 2009. The Handbook of Road Safety Measures. Emerald Group Publishing.
Flügel, S., Halse, A.H., Hulleberg, N., Jordbakke, G.N., Veisten, K., Sundfør, H.B., Kouwenhoven, M., 2020. Verdsetting Av Reisetid Og Tidsavhengige Faktorer. Dokumentasjonsrapport Til Verdsettingsstudien 2018-2020. English: Valuation of Travel Time and Time Dependent Factors. Documentation of the Value of Time Study 2018-2020. Institute of Transport Economics. Technical report 1762/2020.
Fosgerau, M., Hjorth, K., Brems, C., Fukuda, D., 2008. Travel Time Variability: Definition and Valuation. DTU Transport, Denmark.
Fosgerau, M., 2009. The marginal social cost of headway for a scheduled service. Transp. Res. Part B Methodol. 43 (8–9), 813–820.
Fosgerau, M., Karlström, A., 2010. The value of reliability. Transp. Res. Part B Methodol. 44 (1), 38–49.
Fielbaum, A., Jara-Diaz, S., Gschwender, A., 2020. Beyond the Mohring effect: scale economies induced by transit lines structures design. Econ. Transport. 22, 100163.
Fitzpatrick, K., Shamburger, C.B., Krammes, R.A., Fambro, D.B., 1997. Operating speed on suburban arterial curves. Transport. Res. Rec. 1579 (1), 89–96.
Furth, P.G., Wilson, N.H., 1981. Setting frequencies on bus routes: theory and practice. Transport. Res. Rec. 818, 1–7, 1981.
Haglund, M., Åberg, L., 2000. Speed choice in relation to speed limit and influences from other drivers. Transport. Res. F Traffic Psychol. Behav. 3 (1), 39–51.
Jansson, J.O., 1980. A simple bus line model for optimisation of service frequency and bus size. J. Transport Econ. Pol. 53–80.
Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2020. Strategies for transit fleet design considering peak and off-peak periods using the single-line model. Transp. Res. Part B Methodol. 142, 1–18.
Jara-Díaz, S., Fielbaum, A., Gschwender, A., 2017. Optimal fleet size, frequencies and vehicle capacities considering peak and off-peak periods in public transport. Transport. Res. Pol. Pract. 106, 65–74.
Jara-Díaz, S., Gschwender, A., 2003. Towards a general microeconomic model for the operation of public transport. Transport Rev. 23 (4), 453–469.
Jara-Díaz, S.R., Gschwender, A., 2009. The effect of financial constraints on the optimal design of public transport services. Transportation 36 (1), 65–75.
Jørgensen, F., Hanssen, T.E.S., 2019. Implications for traffic safety from car drivers' secondary task engagement–An economist's view. Econ. Transport. 20, 100136.
Jørgensen, F., Pedersen, H., 2005. Enforcement of speed limits—actual policy and drivers' knowledge. Accid. Anal. Prev. 37 (1), 53–62.
Jørgensen, F., Polak, J., 1993. The effect of personal characteristics on drivers' speed selection: an economic approach. J. Transport Econ. Pol. 237–252.
Jørgensen, F., Solvoll, G., 2018. Determining optimal frequency at ferry crossings. Transport Pol. 63, 200–208.
Kleinbaum, D.G., Klein, M., 2010. Survival Analysis, 3. Springer, New York.
Martens, K., 2007. Promoting bike-and-ride: the Dutch experience. Transport. Res. Pol. Pract. 41 (4), 326–338.
Mohring, H., 1972. Optimization and scale economies in urban bus transportation. Am. Econ. Rev. 62 (4), 591–604.
Nilsson, G., 2004. Traffic Safety Dimensions and the Power Model to Describe the Effect of Speed on Safety. Doctoral dissertation, Univ.
Noland, R.B., Small, K.A., 1995. Travel-time uncertainty, departure time choice, and the cost of morning commutes. Transport. Res. Rec. (1493), 150–158.
Odeck, J., 2010. What determines decision-makers' preferences for road investments? Evidence from the Norwegian road sector. Transport Rev. 30 (4), 473–494.
Osuna, E.E., Newell, G.F., 1972. Control strategies for an idealized public transportation system. Transport. Sci. 6 (1), 52–72.
O'Neill, B., 1977. A decision-theory model of danger compensation. Accid. Anal. Prev. 9 (3), 157–165.
O'Sullivan, S., Morrall, J., 1996. Walking distances to and from light-rail transit stations. Transport. Res. Rec. 1538 (1), 19–26.
Parkhurst, G., 2000. Influence of bus-based park and ride facilities on users' car traffic. Transport Pol. 7 (2), 159–172.
Poe, C.M., Mason Jr., J.M., 2000. Analyzing influence of geometric design on operating speeds along low-speed urban streets: mixed-model approach. Transport. Res. Rec. 1737 (1), 18–25.
Polus, A., Fitzpatrick, K., Fambro, D.B., 2000. Predicting operating speeds on tangent sections of two-lane rural highways. Transport. Res. Rec. 1737 (1), 50–57.
Sadia, R., Bekhor, S., Polus, A., 2018. Structural equations modelling of drivers' speed selection using environmental, driver, and risk factors. Accid. Anal. Prev. 116, 21–29.
Small, K.A., 1982. The scheduling of consumer activities: work trips. Am. Econ. Rev. 72 (3), 467–479.
Tarko, A.P., 2009. Modeling drivers' speed selection as a trade-off behavior. Accid. Anal. Prev. 41 (3), 608–616.
Tirachini, A., Antoniou, C., 2020. The economics of automated public transport: effects on operator cost, travel time, fare and subsidy. Econ. Transport. 21, 100151.
Tirachini, A., Hensher, D.A., Rose, J.M., 2014. Multimodal pricing and optimal design of urban public transport: the interplay between traffic congestion and bus crowding. Transp. Res. Part B Methodol. 61, 33–54.
Tirachini, A., Hensher, D.A., 2011. Bus congestion, optimal infrastructure investment and the choice of a fare collection system in dedicated bus corridors. Transp. Res. Part B Methodol. 45 (5), 828–844.
Thune-Larsen, H., K. Veisten, K. L. Rødseth, R. Klæboe: Marginale Eksterne Kostnader Ved Vegtrafikk Med Korrigerte Ulykkeskostnader (Marginal External Costs

Associated with Road Traffic with Corrected Accident Costs). Institute of Transport Economics, Technical Report 1307/2014.
Virtanen, Pauli, Gommers, Ralf, Oliphant, Travis E., Haberland, Matt, Reddy, Tyler, Cournapeau, David, Burovski, Evgeni, Peterson, Pearu, Warren, Weckesser, Bright, Jonathan, Stéfan, J., van der Walt, Brett, Matthew, Wilson, Joshua, Jarrod Millman, K., Mayorov, Nikolay, Andrew, R., Nelson, J., Jones, Eric, Kern, Robert, Larson, Eric, Carey, C.J., Polat, İlhan, Yu, Feng, Moore, Eric W., VanderPlas, Jake, Laxalde, Denis, Perktold, Josef, Cimrman, Robert, Henriksen, Ian, Quintero, E.A., Harris, Charles R., Archibald, Anne M., Ribeiro, Antônio H., Pedregosa, Fabian, Mulbregt, Paul van, SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.
World Health Organization, 2018. Global Status Report on Road Safety 2018. World Health Organization, Geneva, 2018. Licence: CC BYNC-SA 3.0 IGO.

ISBN 978-82-326-6776-5 (printed ver.)
ISBN 978-82-326-6913-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology