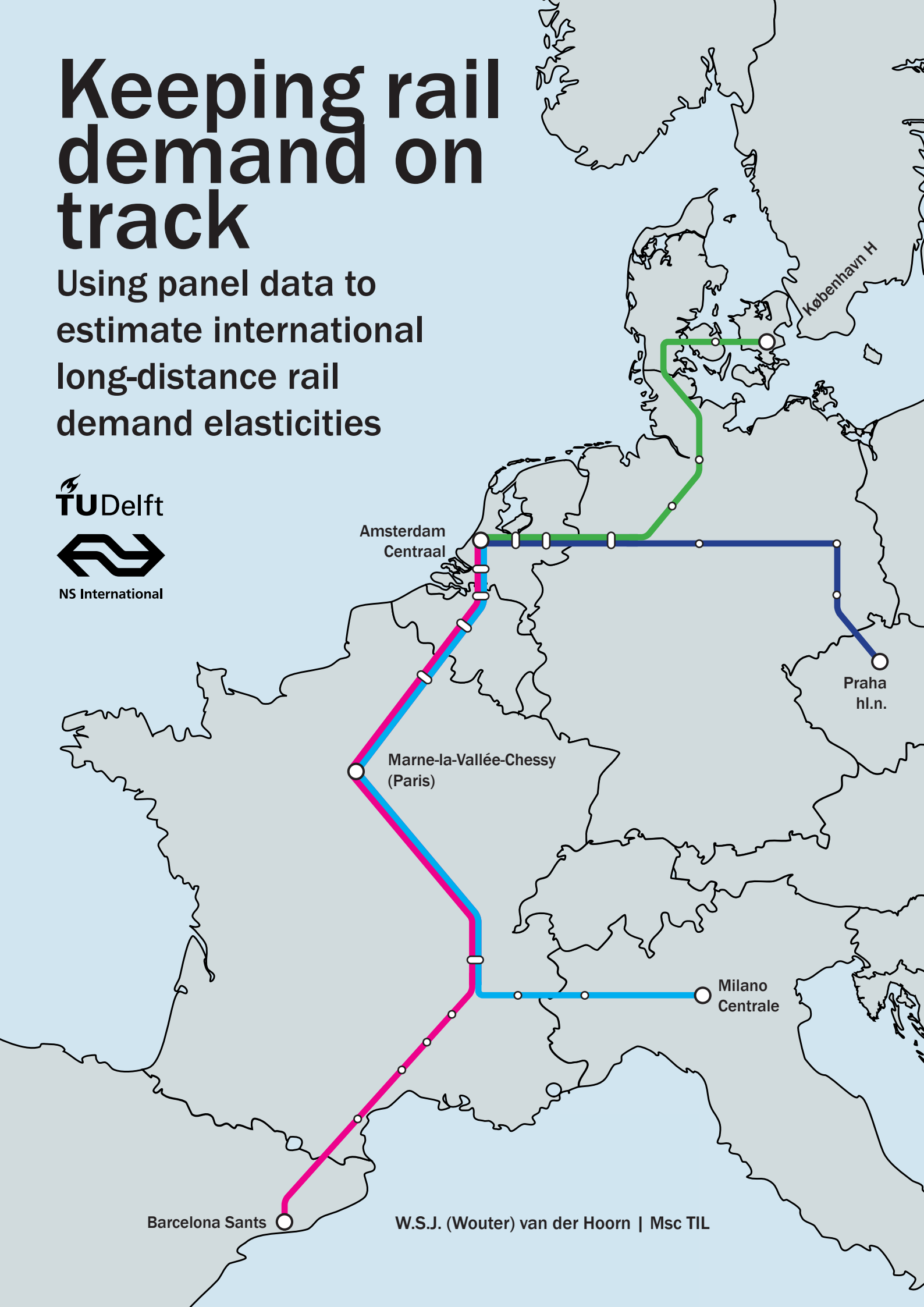
Keeping rail demand on track

Using panel data to estimate international long-distance rail demand elasticities

TU Delft

NS International



W.S.J. (Wouter) van der Hoorn | Msc TIL

Keeping rail demand on track

Using panel data to estimate international long-distance rail demand elasticities

by

W.S.J. (Wouter) van der Hoorn

Public version

Student number:	5370566	
Master program:	Transport, Infrastructure & Logistics	
Thesis committee:	Prof.dr. O. (Oded) Cats	TU Delft, CEG (chair)
	Dr.ir. M. (Maarten) Kroesen	TU Delft, TPM
	L. (Lewis) Shanahan BA	TU Delft, CEG
	Martijn Heufke Kantelaar	NS International
Date	Wednesday 28th October, 2026	
Faculty:	Faculty of Civil engineering and Geosciences	
Cover:	Own design with background map from mapchart.net	

TU Delft

NS International

Preface

Before you lies a piece of paper (or more likely a pdf on your screen) which consumed a big part of my life for the past six months. During my time before my thesis I did not know what I wanted to do my thesis on, and after a first topic that did not work out, I was talking to Oded and a interest from NS International about demand forecasting came up which became the thesis you are reading now.

It was not easy, as I struggled for a long time with defining the topic, scope and methods. And even after officially starting, there was a long time where I felt like I had no idea what I was doing, and just trying to shape the data into something that felt usable to me. But slowly, and mostly nearing the end, my thesis got shape and I am happy that I am at the point where I can write this preface.

I want start by thanking my supervision team. First of all Oded, for supporting and giving me some needed guidance even far before officially starting my thesis. And for connecting me with NS International, which this thesis would not have been possible without. Thank you Maarten, for the regular meetings and keeping me from doubting myself too much. And Lewis, for your approachability and helping me with the writing and methods. I was also want to thank NS International, and especially Martijn for offering this opportunity to do my thesis on a topic which I am very interested in and also seeing some more of what goes on behind the scenes of international train travel.

I also want to thank everyone who supported me during this thesis project. Especially everyone in the thesis room, which made doing the thesis a lot more fun. It was always very nice to have people to talk to and complain with together. I also want to thank my parents and sister, for always supporting me, also when I had doubts or was going through a difficult period. Finally I want to thank my friends I made along the way here in Delft for all the fun times we had and the support.

This thesis is also the last part of completing my life here as a student in Delft, and I am excited and a little bit scared where I will go next. But it would be really cool if one day I can take one the connections I researched today.

Enjoy reading this thesis and I hope to inspire you to consider to travel by long-distance train on your next trip!

W.S.J. (Wouter) van der Hoorn
Delft, October 2026

Summary

Rail travel emits considerably less emissions than air travel, making shifting passengers from air to rail an important element of reducing the negative impacts of passenger transport. However, it can only work if rail offers services that people actually want to use. NS International (NSI) is interested in getting more insight into how demand is formed, and where corridors lie that have demand but do not receive viable long-distance rail services. Forecasting demand for new services is difficult, because demand for a service that does not exist cannot be observed and has to be inferred using existing services. Existing research mostly focusses on single-country or market studies and integrated models covering multiple categories of explanatory variables are limited. While long-distance trips account for a small number of trips they do account for a large share of the travelled distance and emissions, making this market segment relevant. This thesis aims to address these gaps by estimating international long-distance rail demand elasticities and by developing a method that uses these elasticities to forecast demand for new services.

The main research question is:
To what extent do origin-destination characteristics, rail service attributes, and competing air services influence international rail passenger demand, and how can these estimated relationships be used to forecast demand for new international rail services?

The reviewed literature shows that causal models are the most suitable for this research. In addition, research shows that fares, travel time, frequency and transfers are important determinants of demand in combination with market size and income and rail can compete with air if travel time is competitive. Endogeneity between supply and demand complicates the estimation, as operators make decisions about service characteristics with some estimate of demand, which means that these variables are not exogenous.

To estimate the rail demand elasticities, demand between origin-destination pairs is modelled using three types of regression models, namely pooled OLS, origin-destination fixed-effects and origin-destination combined with time fixed-effects. Multiple models are used, because although fixed-effects models absorb any unobserved heterogeneity and therefore generally result in more reliable estimates, some of the variables are time-invariant and can therefore only be estimated in the pooled OLS model. Hypotheses are formulated from the literature to check whether estimates are plausible and to be able to judge the results on a more informed basis than statistical measurements only. The models are estimated using two datasets: passenger volume data from a limited number of services to and from The Netherlands and ticket sales from NSI with origins in The Netherlands and destinations outside of The Netherlands. These two measures of demand are combined with OD characteristics, rail service attributes and air travel characteristics.

The results are mixed, with the passenger volume results giving some implausible estimates. For example, a strongly positive and significant result for rail travel time (3.27 under fixed-effects) is implausible. In addition, many of the variables are insignificant and have very large confidence intervals. The ticket sales models give more plausible estimates. Higher rail fares, longer travel times and more transfers are associated with lower demand (elasticities of about -0.15 for fare under fixed-effects, -0.36 for rail travel time and -0.38 for transfers in the pooled OLS specification), while larger and more economically active destinations attract more demand (a GDP per capita elasticity of 0.37 and population elasticity of 0.08 in pooled OLS). A higher air fare is associated with slightly higher rail demand under fixed-effects, indicating a small substitution effect between air and rail. However, the model fit is low (R^2 of 0.157 for pooled OLS and a within R^2 of 0.022 for fixed- effects), which indicates that the variables explain only a small part of the variation in observed demand.

The results from the ticket sales model (considered more plausible than passenger volume model results) is applied to forecast demand for a number of candidate corridors. The pooled model is used, as it is the only model that can predict absolute demand level for routes that are currently unserved. For every corridor a baseline scenario is estimated, and alternative scenarios are evaluated to examine how the estimated demand responds to changes in rail service attributes. The four candidate corridors all originate in Amsterdam: Barcelona and Milan by high-speed rail, and Copenhagen and Prague by conventional intercity rail. Average demand per OD pair is similar across corridors, at roughly 34 to 44 passengers per day. Milan scores highest per OD pair

and Barcelona has the highest total (about 6,300 passengers per day, total of both directions). This results in the ranking Milan, Barcelona, Prague and Copenhagen. The scenario analysis showed that reducing travel time significantly increases estimated demand (between 8 and 20%), while increasing fare has very little effect (maximum of 1 passenger per day per OD pair) and introducing a transfer has the biggest effect (reducing demand by about 30%)

The ticket sales estimates are the most plausible, as they agree best with the hypotheses and more coefficients are significant, but their magnitudes are generally smaller than those found in the reviewed literature. This may be explained by the data, since ticket sales capture only NSI's share of the market, which is unlikely to be constant across OD pairs. Air fare shows a small positive relation with rail demand within OD pairs, which is the clearest evidence of air-rail substitution in this research. The passenger volume results are less reliable, as they cover few services, are dominated by a few high-volume OD pairs and show little variation within OD pairs, which can inflate the fixed-effects coefficients or give them unexpected signs.

The forecast therefore uses the pooled OLS model, including coefficients that are insignificant or have implausible signs, such as distance and air fare. As a result, the forecast demand is driven mostly by the size and income of the cities served and is only marginally affected by rail service attributes. The method also predicts demand per OD pair, so adding a stop always increases total demand, and it has no awareness of existing services. This leads to under-prediction for established connections and possible over-prediction for small cities far apart. Dynamic pricing, the ease of booking and the time a new service needs to build up demand are not captured either. The method is somewhat generalizable, but the coefficients are specific to NSI ticket sales and should not be transferred to other operators or markets. NSI can use the forecast as a first indication of relative market potential rather than as an absolute demand forecast, and should expect demand to grow once a service is introduced. Future research would benefit from investing in a more complete observed demand dataset which would address the limitations of this research the most. This could be done by data sharing agreements between operators or a central (European) regulatory body. In addition, mitigating endogeneity through the use of instrumental variables (for example track access charges for rail fare) would help with the confidence of the results.

In conclusion, origin-destination characteristics, rail service attributes and competing air services all influence international rail demand, but these relations can only be estimated with limited confidence, and only somewhat plausible for the ticket sales data. They can be used to forecast demand for new services, but given the low model fit, the assumed service characteristics and the extrapolation involved, the results are best read as a relative ranking of corridors and not as absolute demand forecasts.

Contents

Preface i

Nomenclature vi

1 Introduction 1
- 1.1 Contribution and research gap . . . 2
- 1.2 Research questions . . . 2
- 1.3 Report outline . . . 2

2 Literature review 3
- 2.1 Demand forecasting methods for passenger transportation . . . 3
  - 2.1.1 Classic methods . . . 3
  - 2.1.2 Recent developments . . . 4
- 2.2 Determinants of passenger demand . . . 4
  - 2.2.1 Endogeneity in demand modelling . . . 5
- 2.3 Demand forecast accuracy . . . 6
- 2.4 Mode competition on the medium to long-distance market . . . 6
- 2.5 Research gap . . . 6

3 Methodology 8
- 3.1 Conceptual framework . . . 9
  - 3.1.1 Theoretical basis of demand . . . 9
  - 3.1.2 Explanatory variables considered . . . 9
  - 3.1.3 Conceptual model . . . 10
  - 3.1.4 Hypotheses . . . 10
- 3.2 Origin-destination definition . . . 12
- 3.3 Panel data . . . 14
  - 3.3.1 Data requirements . . . 14
  - 3.3.2 Data used . . . 15
  - 3.3.3 Descriptive analysis . . . 15
- 3.4 Estimation strategy . . . 15
- 3.5 Model specification . . . 17
  - 3.5.1 Pooled OLS model . . . 18
  - 3.5.2 OD fixed-effects . . . 19
  - 3.5.3 OD time fixed-effects . . . 19
- 3.6 Model diagnostics . . . 19
  - 3.6.1 Multicollinearity . . . 19
  - 3.6.2 Endogeneity . . . 20
- 3.7 Model evaluation . . . 20
- 3.8 Forecasting approach . . . 21
  - 3.8.1 Model selection . . . 21
  - 3.8.2 Corridor selection . . . 21
  - 3.8.3 Feature construction . . . 22
  - 3.8.4 Demand forecasting . . . 22
  - 3.8.5 Forecast interpretation and uncertainty . . . 23
  - 3.8.6 Corridor ranking . . . 23

4 Data 24
- 4.1 Data collection . . . 24
  - 4.1.1 Historical passenger volume data . . . 24
  - 4.1.2 Timetable data . . . 26

4.1.3 Historical ticket sales data . . . 26
4.1.4 Air travel data . . . 26
4.1.5 origin-destination data . . . 26
4.2 Data processing . . . 26
4.3 OD pair example . . . 29
4.4 Descriptive statistics and visualisations . . . 30
4.4.1 Temporal variation in supply . . . 30
4.4.2 Complete passenger volume dataset . . . 31
4.4.3 Historical ticket sales . . . 32
4.5 Processed datasets . . . 35
4.5.1 Variable construction . . . 36
4.5.2 Data coverage . . . 36
4.5.3 Correlations and multicollinearity . . . 38
4.5.4 Variable distributions . . . 41
4.5.5 Data Limitations . . . 43

5 Results **45**
5.1 Passenger volume . . . 45
5.1.1 Daily results . . . 45
5.1.2 Monthly results . . . 47
5.2 Ticket sales . . . 49
5.3 Comparison of results . . . 50

6 Case study: corridor forecasting **54**
6.1 Forecasting model . . . 54
6.2 Candidate corridors . . . 54
6.2.1 Feature construction . . . 56
6.2.2 Scenarios . . . 56
6.3 Forecast results . . . 57
6.3.1 Corridor 1A: Barcelona HSR . . . 58
6.3.2 Corridor 1B: Milan HSR . . . 60
6.3.3 Corridor 2: Kopenhagen . . . 60
6.3.4 Corridor 3: Prague . . . 61
6.4 Forecasting uncertainty and interpretation . . . 62
6.5 Corridor ranking . . . 63

7 Discussion **65**
7.1 Results discussion . . . 65
7.1.1 Interpretation of the individual coefficients . . . 65
7.1.2 Interpretation of overall models . . . 67
7.1.3 Model specification . . . 68
7.1.4 Comparison to literature . . . 68
7.2 Case study discussion . . . 69
7.3 Limitations and future research . . . 70
7.4 Practical implications . . . 73

8 Conclusion **74**

References **76**

A Data **79**

B Scientific paper **80**

C AI statement **109**

D Elasticity literature overview **110**

E Results of alternative model specifications **114**
E.1 Ticket sales model without air travel variables . . . 114

Nomenclature

Abbreviations

Abbreviation	Definition
FE	Fixed-effects
FUA	Functional Urban Area
GDP	Gross domestic product
HSR	High-speed rail
ICE	Intercity-Express
NSI	NS International
OD	Origin-Destination
OLS	Ordinary least squares
TWFE	Two-way fixed-effects
VIF	Variance inflation factor

1

Introduction

Rail travel emits considerably less emissions than air travel (Givoni, 2007). Shifting passengers from air to rail is therefore an important element of reducing the negative impacts of passenger transport and an important part of the goals of the EU (Kampp et al., 2025). Achieving this shift requires more and better rail services including new high-speed and long-distance rail connections as well as infrastructure (Kampp et al., 2025). NS International can help this shift by identifying new long-distance international rail routes where demand exists but competitive rail options are not available. To achieve this an understanding of how demand for long-distance rail is formed is needed. Once a better understanding of rail demand is formed it can be accommodated by introducing new rail services (Banerjee et al., 2020).

Although many different methods and models exist to forecast passenger demand, demand is often over- or underestimated and there is a lack of standardization (Nicolaisen & Driscoll, 2014). Another aspect which makes forecasting demand for long-distance passenger transport difficult is the focus on specific countries or markets. The insights gained from these models can be useful, but every market is different which makes the comparison of different models difficult (Rich & Mabit, 2012). Complicating demand forecasting is the endogeneity between demand and supply, where supply is created to accommodate demand but demand can also be induced by supply.

Forecasting demand for new services presents an additional challenge, since no historical observations exist for these specific connections. Instead, demand must be inferred from relationships observed on existing services. Two common approaches are: disaggregate demand models and aggregate demand models (Banerjee et al., 2020). Disaggregate demand models, such as discrete choice models, estimate the probability that individual travellers choose a particular transport mode or service. These models use revealed or stated preference data collected at the individual level. Aggregate demand models directly estimate passenger volumes using aggregated observations, relating observed demand to explanatory variables. For this study an aggregate demand modelling approach is used, because this allows the use of observed data, whereas disaggregate models usually rely on stated preference data, which is subject to hypothetical bias. In addition, it allows the model to estimate demand volumes across corridors, rather than the mode choice of individual travellers.

Previous research has shown that explanatory variables such as travel time, fares and service frequency play an important role in explaining passenger demand (Avogadro & Redondi, 2023; Kristoffersson & Liu, 2024; Wardman & Batley, 2022). Demand is also influenced by external factors such as GDP, population and competing transport modes (Wardman, 2006). Econometric demand models can be used to estimate these relationships by linking historical variations in passenger demand to variations in explanatory variables.

Air transport tends to dominate on the medium-distance market (300 to 1000 km) (Fale et al., 2025). Rail can still compete on these routes, with the relative travel time playing an important role. Reductions in rail travel time, for example through high-speed rail or new direct services, have been shown to shift passengers from air to rail (Gundelfinger-Casar & Coto-Millán, 2017). Service frequency also plays an important role in the competition between the two modes (Dobruszkes, 2011). In addition, rail offers city-centre to city-centre connections and generally higher comfort than air travel. Rail demand on these routes therefore depends not only on the rail service itself, but also on what air transport offers on the same connection. Understanding passenger demand for rail services in the medium-distance market is especially relevant, as this market represents an area of competition between air and rail transport (Gundelfinger-Casar & Coto-Millán, 2017).

1.1. Contribution and research gap

The contribution of this study is driven by the company relevance of NS International (NSI) as well as the gap in the existing body of literature.

The rail market is getting more deregulated and competition is stimulated which means that NSI gets more freedom into how and to which destinations they run their international train services and gets more competition from other operators ("Fourth railway package of 2016 - Mobility and Transport", 2016). Each new service is costly and a risk, therefore NSI wants to get more insight into how demand is formed and what corridors are the most promising when considering to operate new medium to long-distance services. This requires estimates of how international rail demand responds to origin-destination characteristics, rail service attributes and competition from air, as well a method to apply the estimates to currently unserved corridors.

The existing literature only partly provides this. First, empirical rail demand elasticities mostly originate from single-country or market studies, and international estimates for long-distance rail are rare, which makes generalization difficult. Second, studies that estimate the effects of multiple demand drivers at the same time, such as fare and travel time, origin-destination characteristics and competition from other modes are limited. Finally, long-distance and international rail demand remains understudied compared to single country, inter-urban and commuting markets. The literature review in chapter 2 provides more detail on these gaps.

This study aims to address these gaps in two ways. First, by exploiting variation between origin-destination pairs and over time, an econometric model is used to estimate how three types of explanatory variables, namely, origin-destination characteristics, rail service attributes and air travel competition, influence international long-distance rail demand by estimating long-distance rail demand elasticities. Secondly, it develops a method to apply the estimated elasticities to forecast demand for currently unserved corridors which can be used to rank the corridors with the most promising demand.

1.2. Research questions

The research is guided by the following main research question and sub-questions.

Main Research question:

To what extent do origin-destination characteristics, rail service attributes, and competing air services influence international rail passenger demand, and how can these estimated relationships be used to forecast demand for new international rail services?

The following sub questions are used to help answer the main research question.

Sub questions:

1. What are the key determinants of passenger demand and how can these variables be quantified?
2. How much variation is there in international rail supply and demand to and from The Netherlands over the study period and is it sufficient to identify demand responses?
3. What are the estimated relations between the explanatory variables and passenger demand?
4. How can the estimated relations be applied to forecast demand for potential new international rail services and to identify the most promising corridors to and from The Netherlands.

1.3. Report outline

The report is structured as follows. In chapter 2 the existing literature on demand forecasting for passenger transport services is reviewed covering different methodologies, problems with demand forecasting, the key determinants of passenger demand found in the literature and mode competition. In chapter 3 the methodology to answer the research questions is described including the development of the case study methodology. Next in chapter 4 the data is introduced, processed and descriptive analysis is performed. Chapter 5 presents the results of the different models and datasets. In chapter 6 the developed methodology is applied to forecast demand for a number of candidate corridors. Next in chapter 7 the results and case study are discussed and suggestions for future research are provided. Finally chapter 8 concludes the research by answering the research questions.

2

Literature review

This literature review provides an overview of the literature relevant to long-distance forecasting passenger demand. Section 2.1 introduces different methods used to forecast passenger transport demand. Section 2.2 reviews what the determinants of passenger demand are and endogeneity in demand modelling. In section 2.3 the accuracy of demand forecasting is discussed. Section 2.4 examines the competition between different modes on the medium- to long-distance passenger market. Finally section 2.5 summarizes the research gap that this research aims to address.

2.1. Demand forecasting methods for passenger transportation

As part of transport planning, there are different levels on which decisions are made. Namely strategic, tactical and operational (Ibarra-Rojas et al., 2015). They differ in terms of time scale, size of decisions and methods. Since the focus of this research lies in estimating long-distance international rail demand elasticities and forecasting demand for new services, the focus lies on forecasting methods on the strategic level. The first subsection gives a brief overview of the types of methods used, while the next describes some more recent developments.

2.1.1. Classic methods

Banerjee et al. (2020) performed an extensive review of the existing literature on passenger demand forecasting for scheduled transportation, with more than 120 papers reviewed. The methods can be categorized into quantitative, qualitative and mixed models. Qualitative methods, such as expert judgement, are typically applied in strategic planning contexts where historical data is unavailable or highly uncertain, but they do not produce any quantitative insight into passenger demand. Only the quantitative and mixed models are therefore discussed further.

Within quantitative methods there are a number of main methods identified by Banerjee et al. (2020), namely time series methods, causal methods, sell up methods and artificial intelligence (AI) methods. Sell up methods are not discussed as they are generally used for low cost air carriers. In this section the main methods are described and where relevant more detail on sub methods are discussed. The AI methods are covered in the next section.

Time series methods

A time series forecast derives its prediction from a series of past observations, typically through extrapolation techniques such as autoregressive integrated moving average (ARIMA) models or exponential time series models (Banerjee et al., 2020). These methods assume that future demand patterns resemble past patterns, and are therefore generally useful for (small changes to) existing services. However, they are not suitable for forecasting demand for new services, since no historical observations exist from which to extrapolate.

Causal methods

A causal method tries to establish the factors that explain past demands and use these to model a forecast (Banerjee et al., 2020). In these models the relation between explanatory variables and passenger demand is estimated, where one or more explanatory variables are used to build a function which describes a demand observation. These vary in form and will be reviewed in more detail since they are more suitable for new services.

In terms of what is predicted, causal models generally fall into three types: models predicting travel volume, modal choice, or route choice. As the goal of this research is to predict passenger volume, the following methods

are discussed with a focus on their application to volume forecasting, although modal choice models remain relevant since they are also a source of elasticity estimates used later in this research.

Regression models identify the relationship between a scalar dependent variable and one or more explanatory variables. Logit models are similar to regression models, but the dependent variable is categorical, and they are used extensively to estimate choices between different travel modes. Logit models are typically applied to disaggregate data, in which individual travel choices are observed directly, usually through revealed preference or stated preference data, whereas regression and econometric models discussed below are usually estimated on aggregate, station-to-station or origin-destination (OD) level data (Ortúzar & Willumsen, 2024). This distinction affects both the type of data required and the interpretation of the resulting elasticities: disaggregate models capture individual-level behavioural responses, while aggregate models reflect the combined effect of these responses across a market or population. However, the accuracy of the logit model depends highly on the model specification and tends to become more inaccurate the more explanatory variables are varied (Börjesson, 2014).

Econometric models model demand as a function of explanatory variables, through linear or log-linear regression analysis to estimate the relationship between demand and the explanatory variables (Tsekeris & Tsekeris, 2011). Typical explanatory variables include travel time, service frequency, ticket price, and measures of market size such as population or GDP. They can also incorporate a gravity model, which estimates demand based on geographical characteristics such as GDP, population, and distance. Econometric models are usually direct demand models, meaning that the model directly deals with variations in aggregate rail demand rather than separate components of trip generation, distribution, and mode choice in classic four step models (Wardman et al., 2007). Cross-sectional models allow for the estimation of demand where the explained and explanatory variables are all related to a single point in time, in contrast to longitudinal models, where the variables are associated with different points in time.

2.1.2. Recent developments

The third main method identified by Banerjee et al. (2020), alongside time series and causal methods, is the use of artificial intelligence (AI) and machine learning (ML) techniques, which have seen growth in recent years. Unlike econometric or regression-based methods, ML models learn patterns directly from data rather than relying on a pre-specified functional form. One of the main machine learning methods used for demand forecasting is a neural network. They can be used for both long- and short-term forecasting and can be relatively accurate due to their ability to capture both linear and non-linear relationships. However, even early applications of neural networks to forecasting showed that a model which fits well in-sample can still produce poor out-of-sample forecasts (Faraway & Chatfield, 1998).

Rocco di Torrepadula et al. (2024) performed a literature review on machine learning for public transport demand forecasting. It showed that, although the focus is mainly on urban transport and time-series short-term forecasting, there are still some useful insights. ML models allow for accurate forecast, but interpretability of the results difficult, which also makes it difficult for transport planners to use them. Furthermore, the complexity, energy use, and computational costs are all unfavourable for ML models. And as is also the case in non-ML methods there is a lack of standardization, availability of datasets, and comparative studies (Banerjee et al., 2020; Rocco di Torrepadula et al., 2024).

These limitations, especially the poor interpretability of ML models, make them less suitable for this research, which aims to produce interpretable elasticity estimates.

2.2. Determinants of passenger demand

In this section demand forecasting literature will be reviewed, specifically to find which variables are predicted, what the determinants of demand are and which limitations are found. In these models, demand is derived from explanatory variables. Many types of variables such as demographics, commercial activities, transport options, land use, demand management and fares influence travel demand (Litman, 2025). The effect of these factors on demand are commonly expressed as elasticities, which is how much demand changes in relation to the change in the explanatory variable.

Demand can be explained by many different types of variables, but the most common variables are travel time and ticket fare. A common way to express time and fare is to use a generalized cost (GC) function combining price, time, and value of time (VoT). The GC typically contains the fare for travel between two stations and the levels of time, service frequency, and interchange after their conversion into equivalent monetary amounts

(Wardman et al., 2007). It does not include the time and cost of accessing and egressing the rail network. Another approach is the generalised journey time and fare (GJT-fare) approach, where time-related variables are combined into a single term and fare remains separate. A third approach is the separate components approach, where the elasticities of all variables are specified separately (Wardman, 2025).

Much of the literature focuses on price elasticity, as it can influence demand directly and ticket fare is easy for an operator to change. However, while a found elasticity gives insight into how demand changes with price, the elasticity greatly depends on the context (Wardman & Batley, 2022). Elasticities vary by travel mode, journey purpose, distance, and even data type. For example, long-distance trips typically show different elasticities than inter urban trips, business travel generally has a lower fare elasticity than leisure travellers and elasticities differ depending on if they are short-run or long-run elasticities (Brons et al., 2002; Wardman & Batley, 2022).

Travel time is also context dependent. A complementary meta-analysis of British time-related demand elasticities finds that elasticities to rail in-vehicle time, headway, and generalised journey time vary significantly by mode, distance, journey purpose, and area type (Wardman, 2022a). Service frequency and journey time changes are also found to affect demand in a revealed preference study of in London, which reports a generalised journey time elasticity of -0.61 and finds that demand responds more elastically to service degradations than to equivalent improvements (Wong & Yap, 2023). Transfers are another relevant service attribute,Allard and Moura (2018) find that the valuation of intercity transfer quality varies considerably between studies and contexts, but that transfer difficulty consistently plays a role in mode choice and negatively influences the travel experience.

Other than fare and travel time, several other factors influence passenger demand. Wardman (2006) showed that (the growth of) GDP is of great importance, while car time, fuel cost, and population also influence demand. This is reinforced by Holmgren (2007), who finds that car ownership, income, and level of service also influence demand, and by Paulley et al. (2006), who similarly identifies fares, quality of service, income, and car ownership as key determinants of public transport demand. Wardman et al. (2007) studied cross-sectional models including access and egress time to a station, showing that these do not exhibit a constant elasticity. A study of Australian urban rail demand found that demand is price inelastic and that car and rail travel are, at best, weak substitutes, with population and service levels the most influential determinants (Wijeweera et al., 2014), showing that determinants and their relative importance can vary considerably by geography and context.

Competition from other modes can also act as an influence on rail demand when multiple modes are feasible on the same corridor. This cross model relationship is covered in section 2.4. An overview of the elasticities from the reviewed literature, which is applied later to formulate hypotheses, is shown in Appendix D.

2.2.1. Endogeneity in demand modelling

When using service supply variables as (part of) determinants for rail passenger demand, it is important to take the endogeneity between demand and supply into account. This is because supply influences demand, but demand also influences supply. For example, a service with high frequency can be the result of high demand, but high frequency can also induce new demand (Tsekeris & Tsekeris, 2011; Wardman et al., 2007). Without accounting for this endogeneity, the model can produce biased and/or inconsistent results and therefore also unrealistic demand estimations (Diaz-Gutierrez & Ranjbari, 2025). Diaz-Gutierrez and Ranjbari (2025) showed that a common approach is to use a fixed-effects approach where an individual specific term is introduced that does not vary over time to explain the time-invariant relationship between the dependent variable and explanatory variables. However, this assumption does not always hold, and therefore another approach is to use an instrumental variables (IV). This is a variable which is correlated with the explanatory variable but not with the dependent variable (Birolini et al., 2020). IV's can be hard to find but do provide a check for endogeneity.

Another aspect to consider regarding the endogeneity between demand and supply is induced demand. By improving a transport service, not only is mode choice influenced, but also new demand can be induced. Yao and Morikawa (2005) model induced travel as a function of service-level changes, finding that improvements to travel time, cost, and access time each results in additional demand. They again also found that business travel is more sensitive to travel time and less sensitive to fare. The elasticities found are substantial (-0.7 to -1.6) and therefore need to be considered when improving existing services or introducing new services. This is reinforced by a study done by Avogadro and Redondi (2023), who found that after introducing Eurostar services between Paris and London, not only did demand shift from other modes, total flows increased, indicating that new travel demand was induced by the new service.

2.3. Demand forecast accuracy

As described in the previous sections there are many different ways to forecast passenger demand for transport services. However, forecast accuracy is an issue, as these forecasts can often under- or over estimate demand. Flyvbjerg et al. (2005) found that in the US, demand for 9 out of 10 rail projects was over-estimated, while demand for road projects was generally under-estimated. Although the US context differs substantially from European long-distance rail, similar patterns of forecast inaccuracy are found across many rail and road projects (Nicolaisen & Driscoll, 2014). There are many reasons for these inaccuracies, but according to Nicolaisen and Driscoll (2014) the most important source of inaccuracies is the auxiliary forecast of other variables, such as inaccurate forecasts of economic growth or car ownership.

Another factor contributing to these inaccuracies is that induced demand is often not taken into account. Model specification also plays an important role, but since there are very few studies using the same dataset, it is hard to compare. Polzin et al. (2025) also emphasized that in more recent years travel behaviour is changing, making traditional models inaccurate. For example, despite rising incomes, trips rates are not increasing any more. Therefore it is crucial to consider the context in which the forecast is created and make use of more data-driven approaches.

In order to check how accurate the forecast is numerous forecast evaluation techniques are used. However, these evaluations are not used to measure performance in absolute terms but to compare different forecasts with each other (Banerjee et al., 2020). The most common methods used are statistical measurements such Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) or Sum of Squared Error (SSE) but many more methods are used. Fildes et al. (2011) made clear that it is important to consider what error measurement is used as this can greatly influence how a forecasting model is evaluated.

2.4. Mode competition on the medium to long-distance market

Medium distance in the passenger transport market is usually defined as trips between 300 and 1000km (Fale et al., 2025). An active body of research on air-rail competition has established that the modes are most competitive/substitutable within this distance category, as high speed rail can effectively compete here on (end-to-end) travel time and is seen as a more sustainable alternative.

Fale et al. (2025) found that for Sweden the break-even point between air and rail is 400km, meaning that for smaller distances high speed rail is dominant and for larger distances air is the dominant mode. Gundelfinger-Casar and Coto-Millán (2017) found that the reduction in travel time due to the introduction of high speed rail influenced the air travel market, indicating that travellers are willing to switch from air to rail when travel time is competitive. Travel time and fare are the most important factors in influencing if rail can replace air travel.

Beyond showing that substitution between air and rail occurs, several studies provide cross elasticities that quantify the magnitude of air and rail influencing the other mode. Wardman et al. (2018) performed a large scale meta analysis of inter-modal cross-elasticity and reports air-rail (change to air travel influencing rail) cross-elasticities of 0.28 and 0.15 respectively. Also finding that cross-elasticities tend to decrease with distance. Clewlow et al. (2014) also found that improvements in rail travel time reduce short-haul air travel demand, although this effects depends greatly on city and airport characteristics. A few studies research specific corridors, with Román et al. (2014) also showed that improved rail quality decrease air travel demand. Hortelano et al. (2016) researched a new pricing scheme on the Spanish high-speed rail network, finding that it led to substantial growth in demand at the expense of air travel demand.Birolini et al. (2020) and Brons et al. (2002) both found that for air travel the sensitivity of travel determinants differs significantly between trip lengths: for longer trips, sensitivity to travel time, fare, and connections is considerably lower than for shorter trips. Indicating that the distance effects are not specific to rail or air. Not all demand shifted from air to rail is diverted demand, part of it is also induced (see subsection 2.2.1). This was also found again the in Spanish context, as Castillo-Manzano et al. (2015) found that only 13.9 % of high speed rail demand can be attributed to substitution from air travel.

2.5. Research gap

The literature reviewed in this chapter shows that passenger demand for scheduled transportation can be forecast using many different methods. A substantial body of empirical evidence exists on the determinants of rail demand, particularly for price and travel time elasticities, and on the competition between rail and air. However, several gaps remain that this research aims to address.

First, most empirical elasticity estimates for rail demand originate from single-country studies, such as the UK (for example Wardman (2006, 2022b) and Wardman et al. (2007)) or Spain (Castillo-Manzano et al. (2015), Hortelano et al. (2016), and Román et al. (2014)). Meta-analysis and choice studies exist (Rich & Mabit, 2012; Wardman, 2022a; Wardman et al., 2018), but international elasticity estimates for long-distance rail demand are rare, limiting the transferability of existing findings to an international long-distance setting such as in this research.

Second, studies that estimate rail demand as a function of multiple explanatory variables simultaneously, including own price and own travel time elasticities, origin-destination characteristics, and cross-modal elasticities, are uncommon as most studies focus only on one or two of these variables. This research combines these variables within a single regression model.

Third, the influence of competing modes (mostly air travel) on rail demand is well documented for individual corridors and countries (section 2.4), but is rarely incorporated as an explanatory variable within a broader, international rail demand model. This research explicitly includes air travel characteristics as a determinant of rail demand.

Finally, although long-distance trips represent a relatively small share of total trips, they account for a disproportionately large share of total distance travelled and associated emissions (Rich & Mabit, 2012). Despite this relevance, long-distance and international rail demand remains comparatively understudied relative to inter-urban and commuting markets (Kristoffersson & Liu, 2024).

By using recent panel data covering a measure for demand, origin-destination characteristics, rail service supply characteristics and air travel characteristics. This research addresses these gaps by estimating long-distance international rail demand elasticities and developing a methodology to apply these elasticities to forecast demand for currently unserved corridors.

3

Methodology

This chapter describes the methodology used to answer the research questions. The main steps are summarized in Figure 3.1. Every box in the diagram links to the section where the method for this step is described, except for raw data, results and literature review which are all contained in their own chapter. The methodology is applied to data provided by NSI, but can be applied to any other rail operator or ticket vendor as long as sufficient data is available.

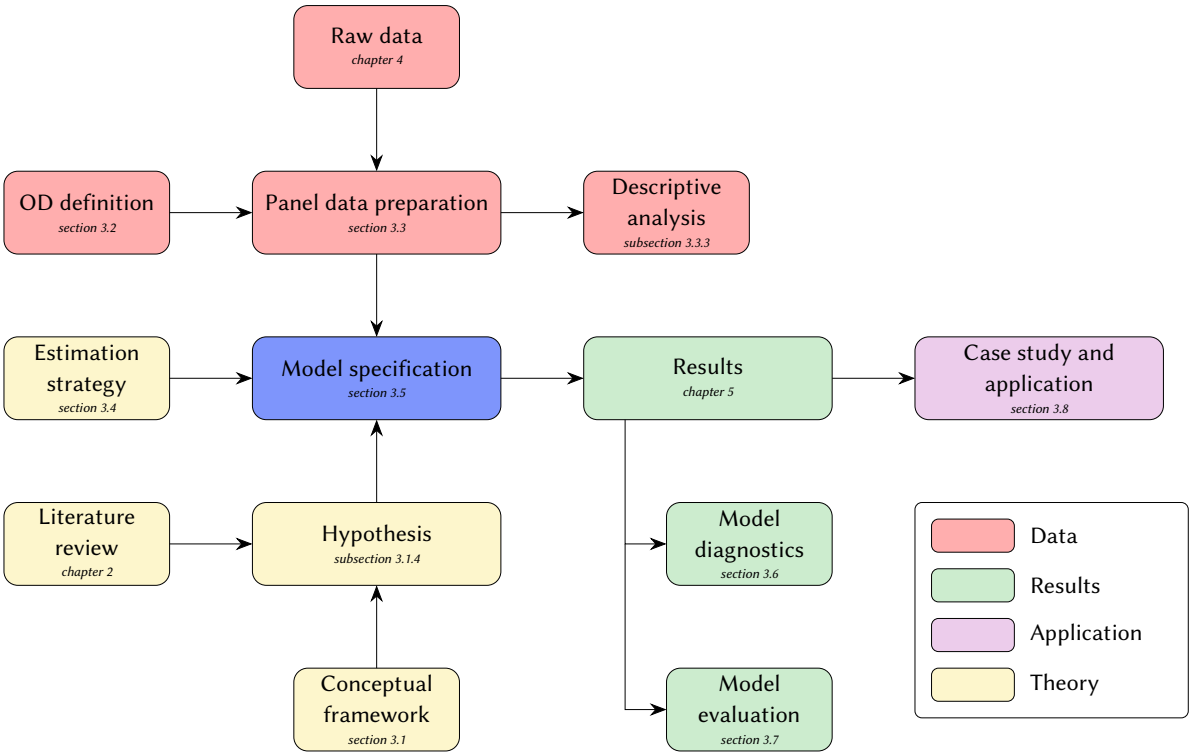


Figure 3.1: Methodology overview

The methodology consists of five main parts, of which data and theory are input for the model, and the results and application are output. For the data input a number of different raw data sources are harmonized into two panel datasets. This is done combining historical passenger volume data, ticket sales data, origin-destination characteristics, rail service supply attributes and air travel competition. All this data is linked to origin and destination train stations by the OD definition. The different data sources result in three analysis panel datasets, at the OD day (and month) level. This means that every entry in the dataset is an origin-destination pair, on a specific date with all the associated attributes. The ticket sales and passenger volume are used as the demand measurement, and the goal is to find the elasticities of the different explanatory variables on this measure for demand. This elasticity value describes how much the demand is expected to change in response to the change in explanatory variable. Before the models are estimated a descriptive analysis on the datasets is done to help with the interpretation of the results and to check whether estimation problems are likely to occur.

The theory feeds into the model specification, and the conceptual framework and existing literature are used to establish hypotheses for the expected signs and magnitude of all the variables. The hypotheses are used to evaluate if the model results are as expected. Based on the estimation strategy different model specifications are used to exploit the variation within OD pairs and between OD pairs. The results are elasticities of the explanatory variables of the different model specifications and are evaluated based on the hypotheses, model fit and statistical significance. The estimated models are also diagnosed on multicollinearity and correlation to find how reliable the estimated coefficients are.

Finally the case study is performed where the results of the models are applied to the developed method to test whether it can be used to forecast demand for potential new train services. This is done by estimating the rail service supply attributes for a few potential new corridors to and from The Netherlands and the demand is estimated for a few different scenarios to see how the demand response to these different scenarios.

3.1. Conceptual framework

This section explains the conceptual framework used in the model. First the way demand is assumed to be formed in this research is explained, next the considered variables are explained and the conceptual model is presented. Finally based on the reviewed literature and theoretical judgement hypotheses are formulated in which the expected relations are shown.

3.1.1. Theoretical basis of demand

This research models rail demand between an origin and destination as a function of rail's own service characteristics, the attractiveness of the origin and destination and air travel characteristics as a competing alternative. The rail service characteristics include fare, travel time, frequency, number of transfers and seat capacity. Fare, travel time and number of transfers are assumed to reduce rail demand when they increase (own price effect), while frequency and capacity are assumed to increase rail demand when they increase by improving the level of service. Correspondingly, an increase in air fare is assumed to decrease rail demand, by assuming that air serves as a substitute for rail in long-distance rail. Thus by making the air alternative less attractive some of the demand is expected to shift to rail (cross price effect). Rail demand is further assumed to depend on the size, economic activity and distance between the origins and destinations. As a larger city or more economic activity is assumed to generate more potential trips independent from a mode's service characteristics.

The conceptual framework implies a specific direction of causality; rail and air service characteristics are assumed to causally influence the volume of rail demand, rather than rail demand causally determining these characteristics. In practice this assumption is not correct, as operators set fares, frequency or capacity with the assumption of some demand on a corridor. This endogeneity between demand and supply and the way it is mitigated is discussed in section 3.6.

3.1.2. Explanatory variables considered

Building on the theoretical basis above and expected relations from the literature, this section lists the considered variables grouped into three categories.

Origin-destination characteristics

Origin and destination characteristics represent the origins and destinations and the attractiveness regardless of mode. The variables used are GDP per capita, population and distance.

Rail service attributes

Rail service attributes are the characteristics of the rail service between OD pairs. Representing the cost and level of service. These include the rail fare, rail in vehicle travel time, rail frequency, number of transfers and seat capacity.

Air travel

As the focus lies on medium to long-distance travel air travel is most prevalent competing mode and is assumed to also influence the demand for rail travel. The corresponding variables for air travel are therefore also included, namely air fare and air frequency.

Passenger volume and ticket sales
Passenger volume and ticket sales are the dependent variable of all analysis, where they represent the total number of travellers or tickets sold between on a OD pair on a specific day or month.

3.1.3. Conceptual model

Figure 3.2 shows the conceptual model summarizing the expected relations between the three categories of explanatory variables and rail passenger demand.

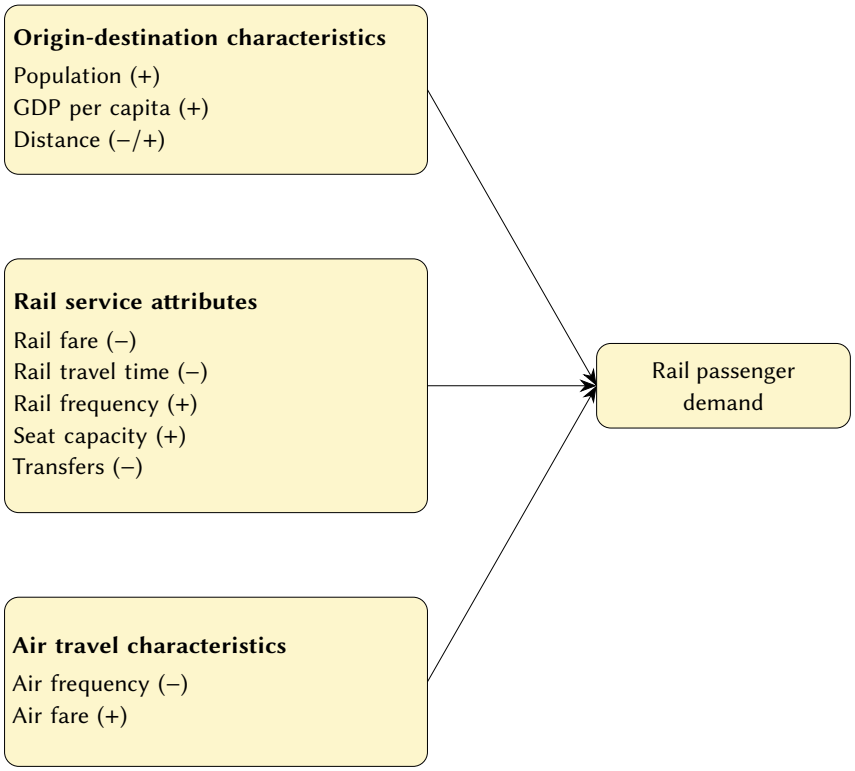


Figure 3.2: Conceptual model

3.1.4. Hypotheses

The results and validity of the models depend heavily on the nature and quality of the data, and may not provide significant or explainable results. Therefore the results need careful interpretation, and for every explanatory variable a hypothesis is formulated. They are based both on literature research and on theoretical judgment. This section lists the hypotheses and they are summarized in Table 3.1. The full table with all the elasticities found in the reviewed literature can be found in Appendix D. It is important to note that some of the studies referenced here are disaggregate choice studies, which means that the interpretation of the elasticities is different and can not be compared directly. The formulated hypotheses will be used to evaluate the plausibility and reliability of the model results.

OD characteristics

- **H1,** Pop_{it}: Population is expected to have a positive elasticity on rail demand, as it indicates the size of the origin and destination, and larger cities naturally attract more travellers. In the literature some studies have estimated elasticity of population, however, multicollinearity problems can lead to unreliable estimates (Wardman, 2006). This is not always the case, as in (Wijeweera et al., 2014) population turned out to be one of the most important determinants of demand. In this research population is included as GDP does not reflect the size of a destination while this is likely to be an important determinant of demand.
- **H2,** GDP_{it}: GDP per capita is expected to have a positive elasticity on rail demand. As higher GDP per capita indicates more economic activity and therefore likely also higher travel demand. In the literature GDP elasticity on rail demand is not often estimated directly, but there are some studies which estimate

income elasticity, which can be interpreted as a proxy for GDP per capita. They usually result in a positive sign, but are not always significant (Wardman & Batley, 2022; Wardman et al., 2018). However, this is not always the same for all modes, as higher income can result in negative elasticity for bus and coach (Börjesson, 2014; Paulley et al., 2006), as rising income drives car ownership and thus substitutes away from these modes. For rail this can also be the case, as (Wijeweera et al., 2014), showed that when the study focusses on commuter rail, higher income led to less rail demand and thus a negative elasticity value.

- **H3,** D_{ij}: Distance is expected to have a negative elasticity on rail demand. Distance is highly collinear with travel time, which could result in an insignificant result. Furthermore, distance elasticity is only identifiable in the pooled OLS specification, since it is time-invariant and absorbed by the OD fixed effect in the model specifications. The literature rarely estimates distance elasticity in isolation, but distance-related heterogeneity is a often researched: rail fare elasticity has been found to decrease with distance (Paulley et al., 2006). The demand response to a fare change has been shown to vary a lot by distance band, from relatively negligible on short corridors to very large on corridors of over 1,000km (Hortelano et al., 2016). Separate from travel time, a negative coefficient on distance itself could reflect residual trip fatigue or planning friction not captured by travel time though the sign is not obvious before estimating the model, since distance could also proxy for unobserved destination attractiveness.

Rail service attributes

- **H4,** $f_{rail,ijt}$: Fare or price elasticity is one of the most researched elasticities in the demand literature, and its expected sign is clear. A higher fare is expected to reduce demand, resulting in a negative coefficient. However, the magnitude of this elasticity is a lot harder to form a hypothesis for. Estimates of own price elasticity in the rail demand literature ranges from -0.38 to -1.12 (Börjesson, 2014; Hortelano et al., 2016; Kristoffersson & Liu, 2024; Román et al., 2014; Wardman, 2006, 2022b; Wardman & Batley, 2022; Wijeweera et al., 2014). This relatively large range is driven by several factors. Long run elasticities tend to result in a larger magnitude than short run elasticities. Furthermore business travellers are usually less price sensitive than private/leisure travellers. Finally elasticities also vary by market segment and route, generally decreasing with longer distance. Given that this research captures the international market over a relatively short time period (2.5 years), the expected magnitude of the own fare elasticity is expected to fall anywhere between -0.4 and -0.8, closer to the short run range.
- **H5,** $TT_{rail,ijt}$: In vehicle rail travel time is expected to have a negative sign, as with fare longer travel time makes rail less attractive. The expected magnitude from the literature ranges from -0.3 to -1.6 (Börjesson, 2014; Paulley et al., 2006; Román et al., 2014; Wardman, 2006, 2022b). Part of this is due to the difference in magnitude found by cross-sectional models compared to elasticities that were found with empirical data, as these models tend to under predict the demand response to large travel time reductions (Börjesson, 2014). As this research relies on observed data, the estimated elasticity is expected to fall towards to lower range, between -0.3 and -0.8
- **H6,** $freq_{rail,ij}$: Higher rail frequency gives travellers more options to travel and thereby makes travelling by rail more attractive. Therefore a positive value is expected for rail frequency elasticity. In the literature frequency is not often estimated directly, but some variables proxy for frequency are estimated. This can be headway, or some other indicator of service quality. These range anywhere between -0.19 for -0.32 (opposite sign because frequency and headway have an inverse relationship) for headway and 0.3 to 1.27 for other service quality indicators (Kristoffersson & Liu, 2024; Paulley et al., 2006; Román et al., 2014; Wijeweera et al., 2014). The expected range for this research is 0.1 to 0.5, as the effect on long-distance travel is expected to be smaller than for urban and commuter rail systems.
- **H7,** cap_{ij}: Seat capacity influences the amount of travellers that can fit on one train, and therefore also how many tickets can be sold for one connection. No directly comparable estimate of seat capacity elasticity was identified in the reviewed literature. This variable is included in this study since data is available before and after a change to seat capacity while all other variables remained the same. It is expected that the elasticity will have a positive sign but since no literature is available there is no hypothesis for the magnitude.
- **H8,** TR_{ijt}: Transfers negatively influence the experience of travellers, as they increase the total travel time but also introduce uncertainty of missing a transfer and make the journey less comfortable. Therefore the transfer elasticity is expected to have a negative sign. In the reviewed literature this is also found although transfers are sometimes expressed in equivalent travel time or as monetary value. Börjesson

(2014), Paulley et al. (2006), and Wardman (2006) all show substantial penalty of transferring, but since they are all not elasticities no quantitative hypothesis is formed for the magnitude.

Air travel characteristics

- **H9,** $freq_{air,ij}$: Higher air frequency gives travellers more scheduling flexibility and is therefore expected to make air a more attractive substitute to rail, reducing rail demand. Air frequency is almost never estimated directly in the reviewed literature. Only in Román et al. (2014) the cross elasticity of air headway (inverse of frequency) is estimated with a value between 0.12 and 0.21. Given the small amount of evidence a hypothesis for a magnitude of -0.1 and -0.5 is proposed, so this hypothesis should be considered with more caution.
- **H10,** $fare_{air,ijt}$: Air fare increases make travelling by rail relatively more attractive, and therefore a positive cross elasticity is expected. This is also found in the literature, with multiple studies showing a relatively wide range, between 0.097 to 0.68 (Hortelano et al., 2016; Román et al., 2014; Wardman et al., 2018). Therefore the hypothesis for the air fare cross elasticity is 0.1 to 0.7.

Variable	Notation	Expected sign and magnitude	Hypothesis	References
Origin-destination characteristics				
Population	pop_{it}	+0.1 to 4.0	H1	(Wardman, 2006; Wardman et al., 2007; Wijeweera et al., 2014)
GDP per capita	gdp_{it}	+0.3 to 1.0	H2	(Börjesson, 2014; Paulley et al., 2006; Wardman, 2006; Wardman & Batley, 2022; Wardman et al., 2018)
Distance	D_{ij}	-0.5 to +0.5	H3	(Hortelano et al., 2016; Paulley et al., 2006)
Rail service supply attributes				
Rail fare	$f_{rail,ijt}$	-0.4 to -0.8	H4	(Börjesson, 2014; Hortelano et al., 2016; Kristoffersson & Liu, 2024; Román et al., 2014; Wardman, 2006, 2022b; Wardman & Batley, 2022; Wijeweera et al., 2014)
Rail travel time	$TT_{rail,ijt}$	-0.3 to -0.8	H5	(Börjesson, 2014; Paulley et al., 2006; Román et al., 2014; Wardman, 2006, 2022b)
Rail frequency	$freq_{rail,ijt}$	+0.1 to 0.5	H6	(Kristoffersson & Liu, 2024; Paulley et al., 2006; Román et al., 2014; Wijeweera et al., 2014)
Seat capacity	cap_{ijt}	+ (no magnitude forecast)	H7	*Variable not reported in reviewed literature*
Transfers	TR_{ijt}	- (no magnitude forecast)	H8	(Börjesson, 2014; Paulley et al., 2006; Wardman, 2006)
Air travel characteristics				
Air frequency	$freq_{air,ij}$	-0.1 to -0.5	H9	(Román et al., 2014)
Air fare	$fare_{air,ijt}$	+0.1 to 0.7	H10	(Hortelano et al., 2016; Román et al., 2014; Wardman et al., 2018)

Table 3.1: Overview of variables and their expected relation to demand, reported values are elasticities

3.2. Origin-destination definition

In order to estimate the elasticities of the different variables, the way the datasets are harmonized and origins i and destinations j are defined needs to be determined. The choices are partly driven by the structure of the

data and partly by the modelling choices. This section describes how origins and destinations are defined and how the different datasets are harmonized.

The passenger volume and ticket sales data both use train stations as origin and destination, while the air travel datasets naturally use airports. For the origin and destination characteristics a geographical definition needs to be used. Since the goal is to estimate the relations for rail demand, the train stations are used as leading origins and destinations and the other variables are linked to the train stations.

For the origin and destination characteristics the geographical definition of functional urban area (FUA) is used. A FUA is defined as a city (area with density higher than 1500 people per square kilometer) of at least 50000 inhabitants plus the area that surrounds the city for which at least 15% of that population commutes to the main city ("OECD Definition of Cities and Functional Urban Areas", n.d.). This is used because the focus lies on long-distance travel and therefore it is assumed that the international train services have a relatively large catchment area. The catchment area for functional urban areas larger as is not only includes the city itself, but also the area where people commute to the city and therefore functions as one area. The origin-destination characteristics are linked to the stations, which means that if multiple stations lie within the same urban area, they receive the same values. International train services usually only serve a limited number of stations in a city, which means that the number of OD pairs does not expand massively. In some cases, train services serve relatively small stations which are not large enough to be considered a functional urban area. This can be for many reasons, such as border stations where electrification or train protection systems may need to be switched, stations which are located in small cities but are relevant for transfers, HSR stations which lie outside of cities (common in France) or small cities which are important for the domestic use of an international train. Stations outside of FUAs are eliminated from the analysis, as they only represent a relatively small amount of the total passengers in the dataset (2.10 % for passenger volume data and 7.37% for the ticket sales data, see subsection 4.5.2).

For the matching of air travel variables, it needs to be considered which airports serve as an alternative to a train station. From the literature it is clear that the catchment area of an airport can differ depending on destination (of the flight), the level of service of the flights offered at that airport and alternative airports availability (Lieshout, 2012), with a maximum of 2 hours travel time to the airport. Since the rail OD pairs are all the distance of short-haul flights within Europe the catchment area is expected be less, and therefore a catchment area of 75km is used. For both origins and destinations, multiple airports may lie within the defined catchment area. A maximum of five airports is considered to capture the relevant airport alternatives for a city while limiting the inclusion of airports that are unlikely to influence traveller behaviour. This threshold is chosen based on the observation that even regions with many airports generally have a limited number of commercial relevant airports. For example, the London area represents one of the largest airport systems, with five major passenger airports serving the region. Therefore, using five airports provides a suitable upper bound. For every origin and destination pair and its corresponding airports the air frequency and air fare is determined. The frequency is the sum of all flights between the airports of the origin and destination. The the air fare is a weighted average, where first the weighted mean fare for the different carriers is determined for every airport pair, and afterwards the weighted mean fare between all the airports in the catchment area of the origin and destination train stations.

Figure 3.3 gives an illustration of how the origins and destinations are defined. Airport A falls within the catchment area of station A and B, and both within the same FUA. This means that airport A is linked both to station A and B, and they get the same origin-destination characteristics since they lie in the same FUA. Station C lies outside of any urban area, and is therefore excluded from the analysis. Station D lies in a different FUA and both airport B and C are within the catchment area. Between airport A and B there is one flight per day with a duration of 60 minutes and air fare of €75. Between A and C there are two flights per day with a duration of 75 minutes and fare of €150. This means that for OD pairs AD and BD the flight frequency is 3 flights per day, weighted air travel time is 79 minutes and the air fare is €125. An example of an OD pair with real data is provided in section 4.3.

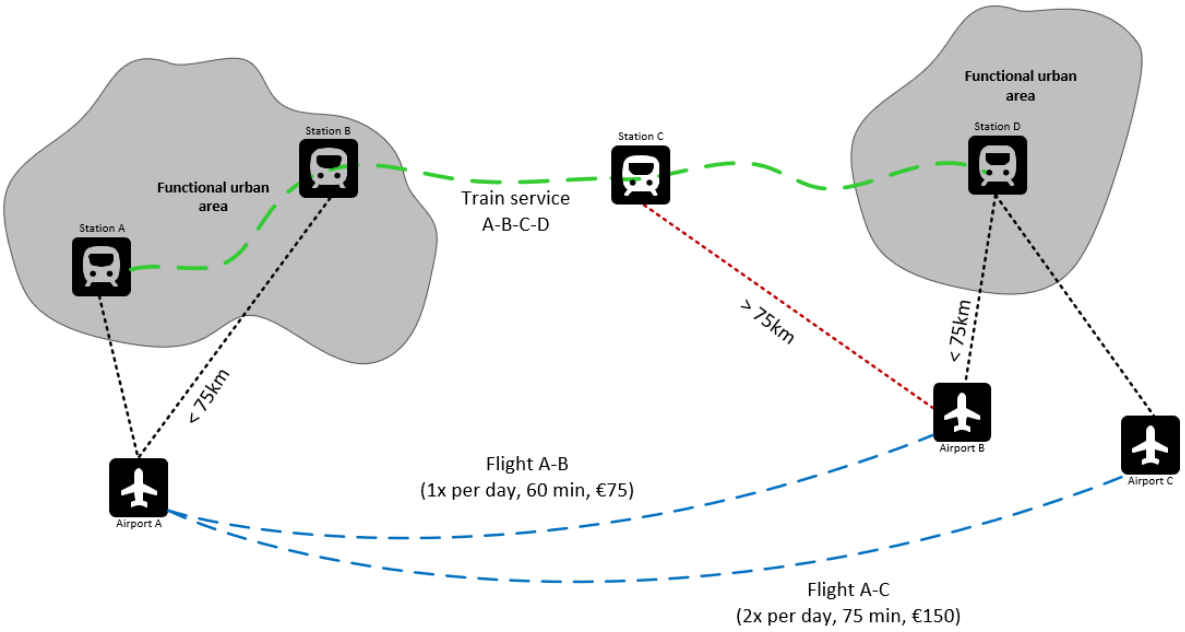


Figure 3.3: Illustration of origin-destination definition

3.3. Panel data

The methodology used for this research described in this chapter is applied to NSI data, but it can be applied to any observed rail demand data as long as a the data requirements are met. This section describes the data requirements, how that data can be harmonized into a dataset suitable for analysis and the steps for a descriptive analysis to understand the data which help to interpret the model results. More detail on the specifics of the data used in this research is provided in chapter 4. It provides details on the characteristics of the data, the processing, a descriptive analysis, extracted variables and data limitations.

3.3.1. Data requirements

In order to apply this methodology there a few data requirements that need to met:

- **Requirement 1**: The model requires some measure of realised passenger demand per OD pair per time period. This can be in the form of counted passenger volumes or ticket sold. Both can be treated as demand indicators. Ideally the temporal resolution is high (daily) as this makes it easier to reach statistical significance. Therefore the data should cover enough time range, with a minimum of 1 year to account for seasonal fluctuation.
- **Requirement 2**: Secondly, each observation of demand must be paired with the corresponding rail service attributes whose demand response is of research interest. Fare and travel time should be included as they enter the model as explanatory variables. Any other remaining attributes (such as frequency, capacity or number of transfers) should be included where available.
- **Requirement 3**: Thirdly, a measure of spatial attributes is required to capture demand drivers that are external to the rail service itself, such as the socioeconomic characteristics of the origin and destination. As these attributes are not observed for train stations, a geographical classification is needed that matches train stations to spatial units for which the data is available. Each station can be assigned to the corresponding spatial unit, allowing the relevant attributes to be matched to the origin and destination of every observation using the OD definition.
- **Requirement 4**: Finally some measure of the availability and competitiveness of competing modes should be available. Depending on the distance scope relevant modes can be chosen such as car, long-distance bus or air.

3.3.2. Data used

For every requirements listed above, this section shows which data was used in this research. The data sources are described in more detail in section 4.1.

Dataset	Source	Open	Requirement	Reference(s)
Historical passenger volume	NSI	No	1 & 2	
Historical ticket sales	NSI	No	1 & 2	
Timetable data	NSI	No*	2	
Train station database	Kaggle	Yes	3	("Train Stations in Europe", n.d.)
Functional urban areas defintion	OECD	Yes	3	("OECD Definition of Cities and Functional Urban Areas", n.d.)
GDP and population	OECD	Yes	3	(OECD, n.d.-b; "OECD Data Explorer • Economy - FUAs", n.d.)
Air travel data	TUDelft**	No	4	

Table 3.2: Used datasets for requirements, *: this particular source not publicly available but data can be found on public sources, **: original source is OAG, processed by TU Delft

3.3.3. Descriptive analysis

To help with interpreting the results and understanding the data, a descriptive analysis of the data is performed on the processed panel datasets. A number of analysis steps are taken to understand the nature of the data. All analysis is performed on the processed panel datasets, where all other variables are assigned to the demand measure. The results of the descriptive analysis can be found in section 4.4.

The analysis starts with a check the size of the datasets, the number of observations, the number of origins and destinations are reported together with the time span and temporal resolution. Also the share of observations excluded due to missing values for required variables is reported. As the fixed-effects models rely on within OD-pair variation to estimate the relations between the explanatory variables and observed demand, the temporal variation in the rail service supply variables is explored. Next the distribution of the demand measure and the explanatory variables is explored, to identify any skewness and outliers. This is done to verify any problems and to help with the interpretation of the coefficients in the estimated models.

The relative contribution of individual OD pairs to the total observed demand is examined, to determine whether the panel data is dominated by a small number of high-volume OD pairs or whether it is more evenly distributed. Finally the pairwise correlations between the explanatory variables and variance inflation factor is assessed to identify multicollinearity.

3.4. Estimation strategy

The primary objective is to estimate the relations between passenger demand and a range of explanatory variables. The data consists of two panel datasets where each entry is a daily observation of demand (ticket sales or passenger volume) between origins and destination pairs, over a study period of 2.5 to 3 years (passenger volume and ticket sales data respectively).

Panel data combines both cross-sectional and time-series dimensions by observing the same entities repeatedly over time. Compared with purely cross-sectional data, panel data allow the analysis to exploit variation both between OD pairs and within OD pairs over time. This is useful to when there is limited variation of the variables over time (Wardman, 1997). Another advantage is that it allows for control of unobserved contributing factors between entities as well as the ability to identify temporal variation of travel behaviour (Baltagi, 2021; Kitamura, 1990).

The econometric framework is based on the general linear regression model form in Equation 3.1, where V_{ijt} represents the passenger demand for OD pair ij at time t, and $X_{k,ijt}$ represent the set of explanatory variables. β_k represents the estimated coefficient for these variables, α is the constant and ϵ_{ijt} is the error term. The error term captures both the time invariant specific effect and the time effect not captured by the variables (Baltagi,

2008).

$$V_{ijt} = \alpha + \sum_{k} \beta_k X_{k,ijt} + \epsilon_{ijt} \tag{3.1}$$

The estimation of linear regression models relies on several assumptions. First of all, the model assumes a linear relationship between the explanatory variable and the dependent variable. Second, the explanatory variables are assumed to be independent of the unobserved factors affecting passenger demand (Wooldridge, 2016). This shown in Equation 3.2.

$$E(\epsilon_{ijt}|X_{ijt}) = E(\epsilon_{ijt}) \tag{3.2}$$

This assumption is particularly relevant for transport demand models and likely to be violated. For example, OD pairs with high demand will likely receive high frequency services, while the underlying attractiveness of these OD pairs may not be fully captured by the explanatory variables. Then if the unobserved characteristics are correlated with the explanatory variables, the estimates of the pooled OLS model may be biased. For this example this could mean that more of the demand is attributed to the high frequency while this is not the reason for the high demand.

To mitigate this issue, the panel structure of the data is exploited by estimating fixed-effects models. Fixed-effects allows unobserved components to be correlated with the observable explanatory variables (Wooldridge, 1995). This is done by controlling for all the time-invariant characteristics of the OD pairs by allowing each OD pair to have its own constant, this means that the model estimates relationships based on changes within the same OD pair over time rather than only relying on differences between OD pairs. This reduces the influence of omitted variables that remain constant throughout the study period. However, another part of endogeneity is the result of reverse causality. The model assumes only one-way causality, where the explanatory variables influence demand. In practice, this is not the case as operators set some of the variables in anticipation of a certain amount of demand, such as the fare or frequency. This is not accounted for by fixed-effects.

Another assumption in standard OLS models is that regression errors are homoskedastic. With panel data the cross-sectional units are often of varying size, and as a result exhibit different variations (Baltagi, 2021). This is the case, as can be seen in subsection 4.4.2: some OD pairs have much larger volumes than others. Consequently, the variance of the error term is unlikely to be constant, resulting in heteroscedasticity. A second concern specific to panel data correlation within entities: since each OD pair is observed repeatedly over time, unobserved shocks affecting a given corridor are likely to remain across consecutive periods, violating the assumption that error terms are independent across observations (Wooldridge, 2016). Ignoring this within-entity correlation typically results in understated standard errors and overstated statistical significance. Entity-clustered standard errors address both concerns simultaneously: they are robust to heteroscedasticity across OD pairs and account for arbitrary correlation of error terms within each OD pair over time, resulting in more reliable statistical inference (Wooldridge, 2016). All reported standard errors and confidence intervals in this study use entity-clustered standard errors.

Based on this framework, three models for both panel datasets are estimated. Resulting in six models. First a pooled ordinary least squares (OLS) model is used as the baseline model. This model treats all observations as independent and estimates relations using the available variation within and between OD pairs. However, the pooled OLS model does not account for unobserved differences between OD pairs and therefore relies on the assumption (Equation 3.2) that these characteristics are not correlated with the explanatory variables. Second, fixed-effects models are estimated to control for the unobserved OD-specific characteristics, and finally time fixed-effects are also added to account for factors that affect all OD pairs at the same time, such as seasonal demand patterns or broader changes in travel. The comparison between pooled OLS and fixed-effects specifications provides insight into the importance of controlling for unobserved heterogeneity and the robustness of the estimated relationships, while also making use of the differences between OD pairs. In section 3.5 the mathematical formulation of the models is specified. Table 3.3 summarizes the advantages and disadvantages of the different models.

Model type	Advantages	Disadvantages
Pooled OLS	Uses full dataset by using both between and within-OD variation, estimates effects of time-invariant variables	Biased estimates due to not accounting for unobserved OD differences, assumes no correlation between regressors and error terms. Endogeneity of omitted variables and reverse causality not addressed.
OD fixed-effects	Controls for all time-invariant OD characteristics, also when correlated with regressors. Estimates relations from changes within the same OD pair	Only identifies relationships from within-OD variation which can be limited. Time invariant variables cannot be estimated. Reverse causality not addressed.
OD + time fixed-effects	In addition to OD fixed-effects also accounts for factors affecting all OD-pairs at the same time (seasonal changes)	In addition to OD fixed-effects further reduces available variation. Variables that move similarly over time partly absorbed.

Table 3.3: Advantages and disadvantages of different model types

3.5. Model specification

This section presents the mathematical specification of the three regression models estimated in this research: a pooled OLS model, an OD fixed-effects model, and an OD time fixed-effects (two-way fixed-effects) model. Each specification is estimated separately on two panel datasets; the ticket sales panel and the passenger volume panel. They only differ in their dependent variable (tickets sold versus passenger volume). Observations are indexed by origin-destination pair ij, where i and j denote the origin and destination train stations, and by time period t. The time index is set at the daily level (travel day) wherever possible, to retain as many observations as possible. However, part of the passenger volume data (see chapter 4) is only available aggregated at the monthly level; for this data, the models are instead estimated with t defined at the monthly level. The pooled OLS model is included to exploit cross OD variation: because some of the OD and air travel characteristics are time-invariant, and rail service supply attributes vary only marginally over time (see subsection 4.4.1), a fixed-effects model would absorb most of this variation.

A log transformation is applied to all variables except the number of transfers. This is done for three main reasons. First it allows the coefficients to be interpreted directly as elasticities, where the β value represents the percentage change in demand in response to a one percent change in the explanatory variables. Second the demand data is right skewed, where a few OD pairs represent a large amount of the total volume. Log transformation helps reduce the disproportionate influence of the high volume OD pairs on the coefficients. The number of transfers (TR_{ij}) is retained in levels because it takes small integer values including zero. Therefore the coefficient is interpreted as semi elasticity.

For the passenger volume dataset, dummy variables are included for each specific train service, since the data originates from a limited number of distinct services. Line dummies capture unobserved, service-specific characteristics such as rolling stock quality, onboard services, or brand reputation, that vary across lines but are not otherwise captured in the model. Omitting these dummies in the pooled OLS specification risks omitted variable bias, as line-specific characteristics may be correlated with observed regressors such as fare or frequency (for example, premium service may combine higher fares and high demand for reasons that are not related to the ticket fare itself). The dummies are included only in the pooled specification: since the train service is time-invariant within each OD pair, its influence is already absorbed by the OD fixed-effects in the remaining two specifications.

Variable	Notation	Pooled OLS	OD fixed-effects	OD time fixed-effects (two way fixed-effects)
Origin-destination characteristics				
Population	pop_{it}	Controlled	Controlled	Controlled
GDP per capita	gdp_{it}	Controlled	Controlled	Controlled
Distance	D_{ij}	Controlled	Absorbed	Absorbed
Rail service supply attributes				
Rail fare	$f_{rail,ijt}$	Controlled	Controlled	Controlled
Rail travel time	$TT_{rail,ijt}$	Controlled	Controlled	Controlled
Rail frequency	$freq_{rail,ijt}$*	Controlled	Controlled	Controlled
Seat capacity	cap_{ijt}*	Controlled	Controlled	Controlled
Transfers	TR_{ijt}**	Controlled	Controlled	Controlled
Air travel characteristics				
Air frequency	$freq_{air,ij}$	Controlled	Absorbed	Absorbed
Air fare	$f_{air,ijt}$	Controlled	Controlled	Controlled
Other influences				
Service dummy variables	δ_s*	Controlled	Absorbed	Absorbed
Origin/destination time-invariant specific characteristics (other competing modes, cultural influences, attitudes, weather etc)	α_{ij}	Unobserved, Unabsorbed	Absorbed	Absorbed
Overall time varying fluctuation (holidays, seasonal swings)	λ_t	Unobserved, Unabsorbed	Unobserved, Unabsorbed	Absorbed
Origin-destination specific time-varying shocks	N/A	Unobserved, Unabsorbed	Unobserved, Unabsorbed	Unobserved, Unabsorbed

Table 3.4: Overview of explanatory variables and their relation in the model specifications. *: only in passenger volume models, **: only in ticket sales models

Equation 3.3 shows the general form of the model, where V_{ijt} is the passenger volume on OD pair ij on day t, $X_{k,ijt}$ is the vector of explanatory variables, β_k is the coefficient of explanatory variable k, α is the model constant and ϵ_{ijt} is the error term. Below the specific versions of the model are shown.

$$\ln(V_{ijt}) = \alpha + \sum_k \beta_k \ln(X_{k,ijt}) + \epsilon_{ijt} \tag{3.3}$$

3.5.1. Pooled OLS model

Equation 3.4 shows the pooled OLS model for the passenger volume model, where each observation is treated as independent. This used to exploit the cross OD variation.

$$\begin{aligned}\ln(V_{ijt}) = {} & \beta_{pop,i} \ln(pop_{it}) + \beta_{pop,j} \ln(pop_{jt}) \\ & + \beta_{gdp,i} \ln(gdp_{it}) + \beta_{gdp,j} \ln(gdp_{jt}) + \beta_D \ln(D_{ij}) \\ & + \beta_{f,rail} \ln(f_{rail,ijt}) + \beta_{TT,rail} \ln(TT_{rail,ijt}) \\ & + \beta_{freq,rail} \ln(freq_{rail,ijt}) + \beta_{cap} \ln(cap_{ijt}) \\ & + \beta_{freq,air} \ln(freq_{air,ij}) + \beta_{f,air} \ln(f_{air,ijt}) \\ & + \alpha + \sum_s \delta_s servicedummy_{s,ij} + \epsilon_{ijt}\end{aligned} \tag{3.4}$$

Equation 3.5 shows the model specification where the dependent variables is ticket sales instead of passenger volume. The main difference is the dependent variable, which the number of tickets sold. The explanatory variables do not include frequency, capacity and service dummies as they are not known in the ticket sales data but instead includes transfers.

$$\begin{aligned}\ln(T_{ijt}) = {} & \beta_{pop,i}\ln(pop_{it}) + \beta_{pop,j}\ln(pop_{jt}) \\ & + \beta_{gdp,i}\ln(gdp_{it}) + \beta_{gdp,j}\ln(gdp_{jt}) + \beta_D\ln(D_{ij}) \\ & + \beta_{f,rail}\ln(f_{rail,ijt}) + \beta_{TT,rail}\ln(TT_{rail,ijt}) \\ & + \beta_{TR}TR_{ijt} \\ & + \beta_{freq,air}\ln(freq_{air,ij}) + \beta_{f,air}\ln(f_{air,ijt}) \\ & + \alpha + \epsilon_{ijt}\end{aligned} \tag{3.5}$$

3.5.2. OD fixed-effects

Equation 3.6 shows what the model looks like for origin and destination fixed-effects. A separate constant is added α_{ij} is added for every OD pair instead of one constant α. However, as mentioned before, fixed-effects absorbs any time-invariant characteristics of the origins and destinations. This means that any variable that does not vary over time, is absorbed. This results in the model getting reduced to Equation 3.6, where only the variables that vary over time remain.

$$\begin{aligned}\ln(V_{ijt}) = {} & \beta_{pop,i}\ln(pop_{it}) + \beta_{pop,j}\ln(pop_{jt}) \\ & + \beta_{gdp,i}\ln(gdp_{it}) + \beta_{gdp,j}\ln(gdp_{jt}) \\ & + \beta_{f,rail}\ln(f_{rail,ijt}) + \beta_{TT,rail}\ln(TT_{rail,ijt}) \\ & + \beta_{freq,rail}\ln(freq_{rail,ijt}) + \beta_{cap}\ln(cap_{ijt}) \\ & + \beta_{f,air}\ln(f_{air,ijt}) \\ & + \alpha_{ij} + \epsilon_{ijt}\end{aligned} \tag{3.6}$$

3.5.3. OD time fixed-effects

Equation 3.7 shows the model with time and OD fixed-effects. Just like in Equation 3.6 only the variables with variation over time remain. The main difference is the introduction of the time constant λ_t which captures any time effects. These are effects that vary over time but are the same for all OD pairs.

$$\begin{aligned}\ln(V_{ijt}) = {} & \beta_{pop,i}\ln(pop_{it}) + \beta_{pop,j}\ln(pop_{jt}) \\ & + \beta_{gdp,i}\ln(gdp_{it}) + \beta_{gdp,j}\ln(gdp_{jt}) \\ & + \beta_{f,rail}\ln(f_{rail,ijt}) + \beta_{TT,rail}\ln(TT_{rail,ijt}) \\ & + \beta_{freq,rail}\ln(freq_{rail,ijt}) + \beta_{cap}\ln(cap_{ijt}) \\ & + \beta_{f,air}\ln(f_{air,ijt}) \\ & + \alpha_{ij} + \lambda_t + \epsilon_{ijt}\end{aligned} \tag{3.7}$$

3.6. Model diagnostics

In order to assess the reliability of the results a number of model diagnostics checks are performed.

3.6.1. Multicollinearity

Multicollinearity occurs when independent variables are correlated with each other, however, only checking the pairwise correlation as described in subsection 3.3.3 is not enough. As one variable might not be highly correlated with another variable but actually with the combination of two variables. In order to check this, the variance inflation factor (VIF) is used, which quantifies how much the variance of a regression coefficient is inflated due to correlation among multiple explanatory variables. When this is the case it does not necessarily

reduce the explanatory power of the model, but can make the coefficient of one variable unstable. This could result in the incorrect sign, magnitude or statistical significance (Wooldridge, 2016).

As the interpretation of each individual variable is relevant in this research, it is important to assess whether there is strong correlation among the explanatory variables. The VIF for the explanatory variable X_k is shown in Equation 3.8, where R_k^2 is calculated by regressing X_k on all other explanatory variables. A higher VIF means that more of the variation in the variable X_k can be explained by the other explanatory variables. This does not mean the variable should be removed from the model, but it does mean that careful interpretation is needed. A threshold for when the VIF is problematic is hard to determine, but a common value is 10 (Wooldridge, 2016). Here any value above 5 is reported. A higher VIF does not mean the variable should be removed from the model, but it does mean that it needs careful interpretation.

$$VIF_k = \frac{1}{1 - R_k^2} \tag{3.8}$$

3.6.2. Endogeneity

As discussed in section 3.1. Rail fare, frequency and capacity are set by the rail operator, and travel time is also influenced by choices made by the operator. This means that these variables are not necessarily exogenous to demand. Operators can set these characteristics with some expectation of demand on a route. This can bias the coefficients of these variables. OD fixed-effects provide some protection against this problem by absorbing the time invariant characteristics of a route which also drive demand. However, any time simultaneity (for example an operator increasing service levels with an expectation of increased demand) is not addressed by fixed-effects. Testing for endogeneity is possible with the use of an instrumental variable (Wooldridge, 2016). This is a variable which is correlated with a potential endogenous regressor (such as fare) but has no effect on demand. For this research no instrumental variable was available to formally test for endogeneity. This limitation means that the coefficients for fare, frequency, capacity and in some ways also travel time should be interpreted with this issue in mind.

3.7. Model evaluation

This section describes how the models are evaluated. A single criterion is not good enough to judge the result quality. Statistical measurements can favour a model that fits the model well without showing meaningful relationships, and says nothing about if the results are consistent with prior literature or if it is theoretically plausible. Therefore the models are evaluated in three ways. Namely statistical measurements, the hypotheses formulated in subsection 3.1.4 and theoretical interpretation taking into account the characteristics of the data.

The statistical measurements are based on measures of goodness of fit, statistical significance and precision of the estimated coefficients. These measures are used to compare the different model specifications and assess how well they can explain the influence of the explanatory variables on passenger demand for both datasets. The goodness of fit of each model is evaluated using the coefficient of determination (R^2), which is the most common way regression models are evaluated (Wooldridge, 2016). It indicates the proportion of variation in the dependent variables that is explained by the model. Also the within R^2 is reported, which explains how much of variation within specific OD pairs is explained by the model. Although a higher value indicates a better fit of the observed data, it is not only used as the main criterion, as this does not mean that the estimates for the explanatory variables are better.

The statistical significance of the estimated coefficients is evaluated using p values, where a *p-value* below 0.05 is considered statistically significant. This means that the explanatory variables has a statistically significant relationship with the dependent variables at the 95% confidence level. For every parameter the 95% confidence interval is also visualized, showing the precision of the estimated parameter coefficients. This is used to evaluate how well each model explains passenger demand. Finally for every model also the number of observations is reported and the unit of analysis to give a understanding of the context of the model.

Every estimated coefficient is tested against the corresponding hypotheses formulated in subsection 3.1.4, to assess whether the sign and magnitude are as expected and whether the estimate is comparable to elasticities reported in the reviewed literature.

Finally the results are also evaluated based on theoretical interpretation. As the results are highly influenced by the characteristics of the data they may not provide statistically significant results or match the formulated

hypotheses. Therefore the results are also interpreted while considering all the aspects of the data described in chapter 4.

3.8. Forecasting approach

This section describes the developed method for the application of the model, which is to forecast demand for new potential train services. As the model results contain a lot of uncertainty, the main goal is to rank a number of corridors by the expected demand to see which is the most promising. This uncertainty is amplified by the fact that the ticket sales represent only part of the total volume on the OD pairs and therefore will under predict total passenger volume. A correction is applied based on the known share of ticket sales for the train services with passenger volume data but this also amplifies the uncertainty. The prediction of passenger volume is done by estimating the demand for a few specific corridors for a two scenarios where the rail service attributes of these services are varied. Figure 3.4 shows an overview of the forecasting method. The methodology is applied in chapter 6.

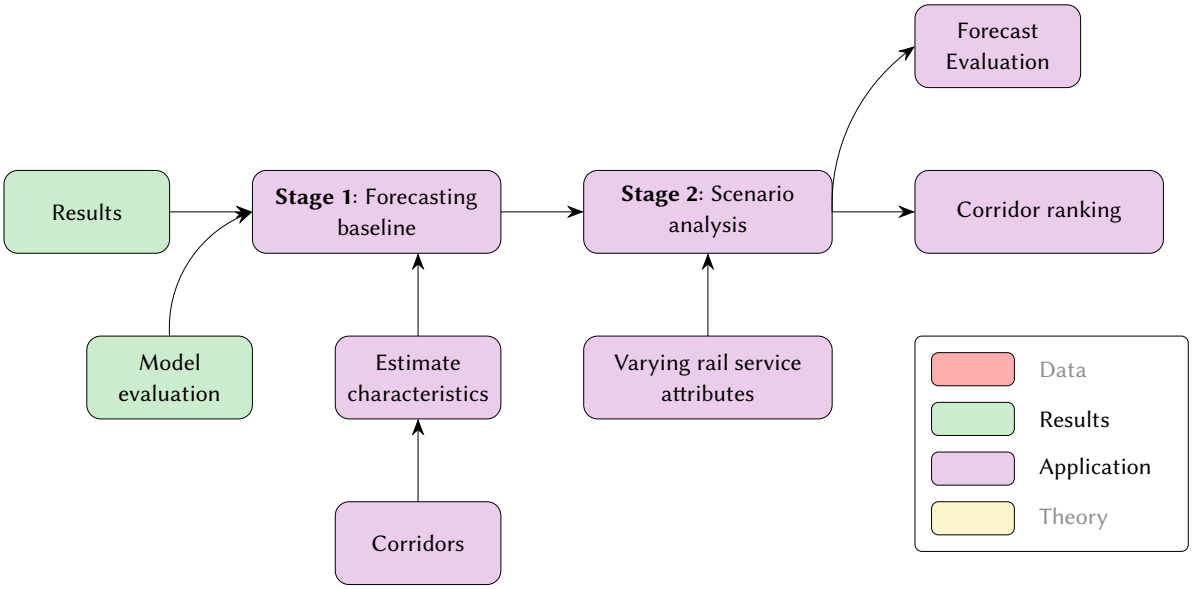


Figure 3.4: Application methodology overview

3.8.1. Model selection

The forecasting application relies on the model specifications introduced in section 3.5. The pooled OLS model is the only model capable of producing an absolute demand level for OD pairs that do not appear in the estimation sample, as the fixed-effects model absorb all cross sectional variation from the panel structure and therefore can not be used for an unobserved OD pair, as the model constants are OD specific in the fixed models. Therefore the pooled OLS model is used to estimate demand for each corridor.

For every corridor a baseline scenario is created based on expert knowledge from NSI for which the demand is estimated. In the second stage an alternative scenario is created where a rail service attribute is varied to examine how the estimated demand responds.

The full pooled OLS model including insignificant coefficients are used for prediction, because dropping insignificant terms after estimating biases the estimated demand. Whether the results for the ticket sales data or passenger volume data is used depends on how well the models perform. As described above (section 3.7), this is based both on statistical measurements, the formulated hypotheses and theoretical judgment.

3.8.2. Corridor selection

A corridor is defined as an ordered sequence of train stations, served by a single train service, connecting an origin and a destination via a route that does not deviate substantially from the most direct feasible path and is technically possible given the existing railway infrastructure. The corridors to evaluate are selected based on four criteria:

1. Sufficient distance to fall within the long-distance market: the corridor should serve sufficient distance to fall within the long-distance market (>300 km) with a maximum travel time of 12 hours. The travel time should not exceed 12 hours as the journey should be possible within daytime and longer travel time requires extrapolation outside of the range of which the model is estimated. Furthermore as only air is taken into account as the competing mode, the rail service should also service destinations that are also feasible by air.
2. The corridor serves destinations not currently served by direct services to and from The Netherlands. The new corridor should serve destinations that are currently not served by a direct train service, require (multiple) transfers to reach and are time intensive.
3. Relevant for NSI: the corridor should be relevant for NSI to operate. This means that the corridor should originate in The Netherlands and have enough market potential to be viable to operate.
4. Data availability: origins and destinations must lie within a FUA, so that GDP and population data is available and as well as proximity to airports. This is necessary because these characteristics are required inputs for constructing the predictor variables used to explain demand in this research.
5. Only one stop per FUA. All stations within the same FUA share the same origin-destination characteristics, so including more than one would serve the same population twice and inflate estimated demand.

3.8.3. Feature construction

For every candidate corridor, the predictor vector required by the pooled OLS equation is constructed from the same data sources as are also used for the panel data. The only variables that are unknown are the rail service supply attributes. Based on expert knowledge a baseline fare, frequency, travel time and capacity is set. Some of the attributes are varied in the scenarios to evaluate how demand responds to these changes. A detailed description of the corridors and corresponding features is described in chapter 6.

3.8.4. Demand forecasting

Forecasting is done in two stages. In the first stage the constructed features of the candidate corridor under the baseline scenario are evaluated using the pooled OLS coefficients. The general model (Equation 3.9) is rewritten to predict the volume of demand for one general day.

$$\ln(V_{ijt}) = \alpha + \sum_k \beta_k \ln(X_{k,ijt}) + \epsilon_{ijt} \tag{3.9}$$

Equation 3.10 shows the estimated model, the *hat* indicates the coefficients are estimated. The error term drops out is it has an expectation of zero.

$$\widehat{\ln(V_{ijt})} = \hat{\alpha} + \sum_k \hat{\beta}_k \ln(X_{k,ijt}) \tag{3.10}$$

To get the estimated demand level the equation is exponentiated back (Equation 3.11). Because exponentiation is a nonlinear transformation, this back transformation results in the median of V_{ijt} rather than the mean. This means that $\widehat{V_{ijt}}$ systematically underestimates the demand (Duan, 1983).

$$\hat{V}_{ijt} = \exp\left(\hat{\alpha} + \sum_k \hat{\beta}_k \ln(X_{k,ijt})\right) \tag{3.11}$$

To correct for this retransformation bias, Duan's smearing estimator is applied (Duan, 1983). The smearing factor $\hat{\phi}$ is computed from the model's own residuals $\hat{\epsilon}_{ijt}$ in the estimation sample (Equation 3.12), avoiding the need for a specific error distribution.

$$\hat{\phi} = \frac{1}{n} \sum_{ijt} \exp(\hat{\epsilon}_{ijt}) \tag{3.12}$$

The corrected demand estimate, representing the mean instead of the median of V_{ijt}, is then obtained by scaling the exponentiated prediction by $\hat{\phi}$ (Equation 3.13).

$$\hat{V}_{ijt} = \hat{\phi} \cdot \hat{V}_{ijt}^{median} \tag{3.13}$$

When the ticket sales model is used to generate the baseline forecast, the result will systematically underpredict actual demand, since NSI sells only a share of total tickets on a given OD pair, meaning the model effectively predicts NSI ticket sales rather than total OD demand. To correct for this, the share of tickets sold by NSI, δ, is calculated from the corridors for which passenger volume data is available. For connections longer than 300 km, the average of these observed shares is applied as a correction factor to scale ticket sales forecast up to an estimate of total demand (Equation 3.14). The correction factor is based on the share of ticket sold by NSI presented in subsection 4.5.2

$$\hat{V}_{ijt}^{total} = \frac{\hat{V}_{ijt}}{\delta} \tag{3.14}$$

Combining these steps into one formulation the model for forecasting demand becomes:

$$\hat{V}_{ijt} = \frac{\exp\left(\hat{\alpha} + \sum_k \hat{\beta}_k \ln(X_{k,ijt})\right) \cdot \hat{\phi}}{\delta} \tag{3.15}$$

This correction extrapolates the additional demand not captured by ticket sales, but since it is itself an estimate derived from a limited set of OD pairs, it introduces an additional source of uncertainty into the final forecast.

3.8.5. Forecast interpretation and uncertainty

The forecast is subject to many different sources of uncertainty, and therefore should be interpreted carefully, and not be seen as the actual expected demand when running a train service under the estimated service characteristics. Instead, the forecast should be seen as a way to apply the methodology and see how the model responds to different corridors and their characteristics. For this research no way to quantify the uncertainty was found, but the model fit and significance presented in chapter 5 can serve as an indication of the forecast quality.

3.8.6. Corridor ranking

Based on the forecast the corridors are ranked to find which is the most promising to operate a new service on. To do this 3 metrics are used, namely the total demand, average demand per OD pair and demand between the endpoints.

4

Data

In this chapter, the data used for the analysis is described and the data processing steps are shown. It provides descriptive statistics about the data, and the processing choices made. The chapter start by describing how the different raw data sources are collected (section 4.1), next the data processing steps are described (section 4.2), an example observation is shown using the OD pair definition (section 4.3), descriptive statistics are shown (section 4.4) and finally the data limitation imposed by the data are described (subsection 4.5.5).

4.1. Data collection

This section describes which raw data was collected and what it contains. Five main types of data were collected, which are shown below.

4.1.1. Historical passenger volume data

Provided by NSI is the historical passenger volume data. It contains how many passengers travelled on a number of train services during the past years. Table 4.1 and Figure 4.1 gives an overview of the different train services and the corresponding data. For the ICE (Intercity-Express, to and from Germany) services data is available for a period from January 2020 until April 2026. It is provided on a daily, origin-destination, per train level. This means that for each specific train, it is known how many passengers travelled between each origin and destination that the train service serves. It also includes whether these passengers travelled in first or second class and where their journey started and ended.

For the Eurostar, it serves Paris with a frequent (nearly hourly) service as well as Disneyland Paris with a direct service in the weekend. Next to the Paris service data is also available for the Eurostar Sun and Eurostar Snow, which run in the summer and winter as a seasonal service towards the south of France. These services run only once a week for a few months per year. For all Eurostar services, the passenger volume data is available aggregated on a monthly origin-destination level from January of 2024 until December of 2025. However, the Eurostar Sun only ran during 2024, and thus the corresponding data is also only available for 2024. For the Eurostar services, also the average ticket fare is available. This is also aggregated on a monthly OD level and does not differentiate between the different classes within the Eurostar service.

Train service	Origin	Destination	Frequency	Data resolution	Data period
IC(E) Berlin	Amsterdam CS	Berlin Hbf	7x per day	OD daily level	Jan 2020 - April 2026
ICE Frankfurt	Amsterdam CS	Frankfurt (main) Hbf	6x per day	OD daily level	Dec 2023 - April 2026
ICE Munich	Amsterdam CS	München Hbf	1x per day	OD daily level	Dec 2024 - Apr 202
ICE Basel	Amsterdam CS	Basel SBB	1x per day	OD daily level	Dec 2023 - Jun 2024
Eurostar Paris	Amsterdam CS	Paris Gare du Nord	10x per day (+ 4x per day to Brussels)	OD month level	Jan 2024 - Dec 2025
Eurostar Disneyland	Amsterdam CS	Disneyland Paris (+ CDG airport)	3x per week	OD month level	Jan 2024 - Dec 2025
Eurostar Snow	Amsterdam CS	Bourg St Maurice	1x per week (winter)	OD month level	Jan 2024 - Mar 2024, Dec 2024 - Mar 2025
Eurostar Sun	Amsterdam CS	Marseille St Charles	1x per week (summer)	OD month level	July 2024 - August 2024

Table 4.1: Train services with historical passenger volume data

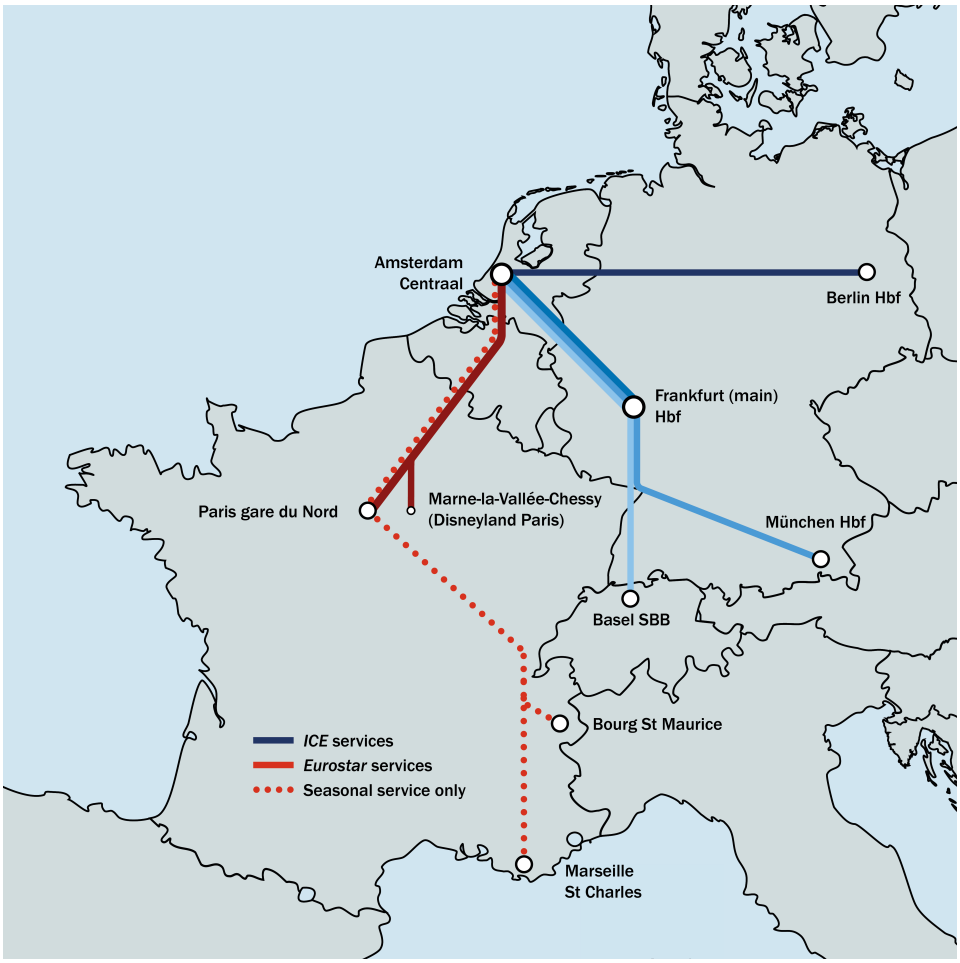


Figure 4.1: Map of train services with passenger volume data (own design)

4.1.2. Timetable data
For the services with historical passenger volume data, also the timetable is provided. This contains the arrival and departure time at every station served by the train service as well as the frequency (trains per day). Rolling stock type is also provided from which the seat capacity is derived. When the timetable was not provided by NSI, it was derived from the 'Train Archive' by using the train number and the date ("Treinarchief", n.d.).

4.1.3. Historical ticket sales data
In addition to historical passenger volume data, ticket sales data from NSI is available. This dataset contains information on tickets sold by NSI for services operated by NSI and other rail operators. The data includes all tickets originating in the Netherlands with a destination outside of The Netherlands. This results in 395 origins (nearly all train stations in The Netherlands) and 5660 different destinations outside of The Netherlands. The dataset covers the period from July 2023 to April 2026 and is observed at a daily origin-destination level, where each observation represents the number of tickets sold for a specific OD pair on a specific departure date, along with the corresponding average fare paid without differentiating between different fare classes. In addition it also includes the travel time and number of transfers between the origin and destination. The rolling stock type was not included so seat capacity could not be derived from the data, and frequency was also not available.

It should be noted that the data does not represent total passenger volume or demand, as tickets purchased through other booking channels or operators are not included. As a result, the dataset reflects the sales within the NSI booking platform rather than complete passenger flows. This means that the ticket sales are likely to be biased by Dutch passengers as they are more familiar with NS (International). In addition the further away a destination is from The Netherlands, the less likely it is that passenger booked through NSI. This is also reflected in Figure 4.5 and Figure 4.7.

4.1.4. Air travel data
Other than rail also data about air travel is available. For all major cities in Europe (defined by the air travel dataset) airports within 100km of that city were identified, resulting in 284 airports. For these airports pairs, per carrier the mean flight duration and daily flight frequency was determined for the period from March 2022 until December 2025. The frequency is aggregated and the travel time is averaged based on origin and destination to get the travel time and frequency combined for all carriers for every airport pair. In addition the for the same set of airports data about the air fare was available. It includes the fare per carrier per OD airport pair per month.

4.1.5. origin-destination data
For the origins and destinations characteristics the definition of functional urban area is used (FUA). More detail on how the origins and destinations are defined is described in section 3.2. For every FUA within Europe the GDP and population data is gathered, for the years between 2019 and 2025 (OECD, n.d.-b; "OECD Data Explorer • Economy - FUAs", n.d.). The GDP is only available for FUAs that are larger than 250000 inhabitants, and more recent years are incomplete. Therefore this means that the GDP and population data is not always available for the same time frame as the passenger volume or ticket sales data.

For geographical analysis the boundaries of the FUAs were used to determine which origins and destinations stations lie in each FUA. For this the FUAs cities and boundaries were used ("OECD Definition of Cities and Functional Urban Areas", n.d.) and a dataset on all the train stations within Europe ("Train Stations in Europe", n.d.). This contains the name, coordinates and basic properties of 36000 train stations in Europe.

4.2. Data processing
In order to derive variables from the different data sources and create a suitable data structure for analysis a number of data processing steps need to be performed. The main goal of the data processing is to harmonize the different data structures into one consistent structure. Figure 4.2 shows the steps performed which are explained in more detail below. In the end two panel datasets are created: one using passenger volume as measurement of demand, and the other using ticket sales. The two measurements of demand indicators in separate datasets because the interpretation of the explanatory variables differs between passenger volume and ticket sales, as explained in more detail in chapter 3. In this section, the data processing is described: first the steps used for each data type, and afterwards the way they were integrated.

Origins and destinations
As origins and destinations for the analysis, train stations are used (see section 3.2 for more detail). Therefore the train station dataset was used as the leading dataset where different attributes can be matched to. The names were stripped to all lowercase without any special characters for easier matching with other datasets and only countries which could theoretically be reached within 12 hours of travel time were kept. Based on the train station coordinates and the FUA boundaries, for every station it was identified in which functional urban area it is located and the corresponding GDP per capita and population of that FUA was matched to the train station. In addition, based on the coordinates of the train stations and the airports, the closest airports within the catchment area of 75 km (see section 3.2 for reasoning), were identified, and a maximum of five airports were matched to the train station. This results in every station (origin or destination) having zero up to five possible airports.

Passenger volume
For the historical passenger volume data processing needs differed depending on the source. For the ICE services, the data originated from Deutsche Bahn. This meant that for every data entry with passengers to or from The Netherlands, the corresponding boarding or alighting station was set to the last station within the Netherlands. For example, for the IC(E) Berlin this was Hengelo. However, the start or end station was also known (where the trip ended or originated but not a station served by the train service) which was used to reconstruct the actual boarding or alighting station. By using the coordinates of the start/end station of the trip and the known actual served stations, the passengers were mapped to their actual boarding or alighting station based on the geographical distance. In addition, another issue was for some entries, that there were incorrect boarding or alighting stations, such as stations in Poland or Austria for the ICE services from The Netherlands to Germany. These entries were filtered out. Finally the entries were aggregated based on boarding, alighting station and travel date, which results in every entry being the number of passengers between two stations on a specific date, rather than every separate booking. The Eurostar data required very little processing, only renaming columns data to match the other datasets, and again having the same structure except for the temporal resolution. Where every observation is the number of passengers between two stations in a specific month.

Timetable
Based on the timetable of the services the travel time, frequency and seat capacity is matched to the corresponding OD pairs of the passenger volume data. In addition the number of transfers is also part of the ticket sales data, but no matching to the volume data is needed because all services are direct.

Ticket sales
For the ticket sales data, some filtering was needed. Mainly to remove some false entries, such as negative number of travellers or negative travel time. Afterwards the data was aggregated based on travel date, boarding and alighting station to get the number of ticket sold for a specific travel date and the corresponding average ticket fare. Again this meant that every entry in the ticket sales dataset was a number of tickets sold between an origin and destination station on a specific date with the corresponding fare, travel time and number of transfers.

Air travel
The air travel data contains the frequency of flights, travel time and fare per airport pair and carrier, however due to the time zone differences, the travel time needed to be corrected. This was done by mapping the time zone to every airport, and when the time zones between an airport pair was not the same, the travel time was checked and an hour was added or subtracted to get the actual travel time by air. Afterwards the travel time and frequency was aggregated based on the origin and destination airport to get the total frequency of flights and mean travel time between every airport pair in the dataset. In addition, as the air fare data also includes the time dimension this was processed separately. For every carrier and airport pair the weighted average fare was calculated using the frequency of this carrier and corresponding air fare data. This resulted in every airport pair getting the corresponding average fare per month.

Integration
For both the ticket sales and the passenger volume data the station names were mapped to the train station database, to get all the corresponding attributes. To do this fuzzy matching was used to find the matching station for every station in the passenger volume and ticket sales data. When the match was exact or >90% the station was mapped automatically. This was sufficient for roughly 95% of the stations in the dataset. For the other stations with a worse match the matching was done manually by reporting the best five matches and

manually picking the correct one or specifying another station. After matching this meant that for all stations in the passenger volume and ticket sales data they had corresponding characteristics such as the coordinates, which were used to calculate the geographical distance between each OD pair.

Based on the air travel data and the airports matched to the stations, the flight frequency and travel time between each OD pair was calculated. This was done by summing the frequency of the flights between the different airports at every origin and destination, and computing the weighted average air travel time. In addition the weighted average fare was also calculated using the frequency of every carrier and the fare for those carriers for every OD airport pair.

As described in section 4.1 the passenger volume data does not contain ticket fare except for the Eurostar services to Paris. Therefore the ticket sales data was used to find the appropriate ticket fare for the other train services. For every entry in the passenger volume data, the nearest date was found were for the same origin-destination pair a ticket was sold. This ticket fare was mapped to the passenger volume data. For the Eurostar services both the ticket fare for all tickets as well as the ticket fare for the tickets sold by NSI was provided. This was used to verify that the ticket fare sold by NSI is representative for the overall ticket fare which turned out to be correct. Since the ticket sales data only contains ticket sold with origins in The Netherlands and destination outside of The Netherlands it was assumed that the ticket fare for A-B is the same as for B-A, so that the analysis could be done for both directions.

Since the Eurostar passenger volume data is only available aggregated per month, it cannot be combined with the other services at a daily level. Therefore, passenger volume datasets were created at both a daily and monthly resolution. The daily dataset excludes Eurostar, whereas the monthly dataset includes Eurostar services.

Figure 4.2 shows the data processing steps performed in a diagram.

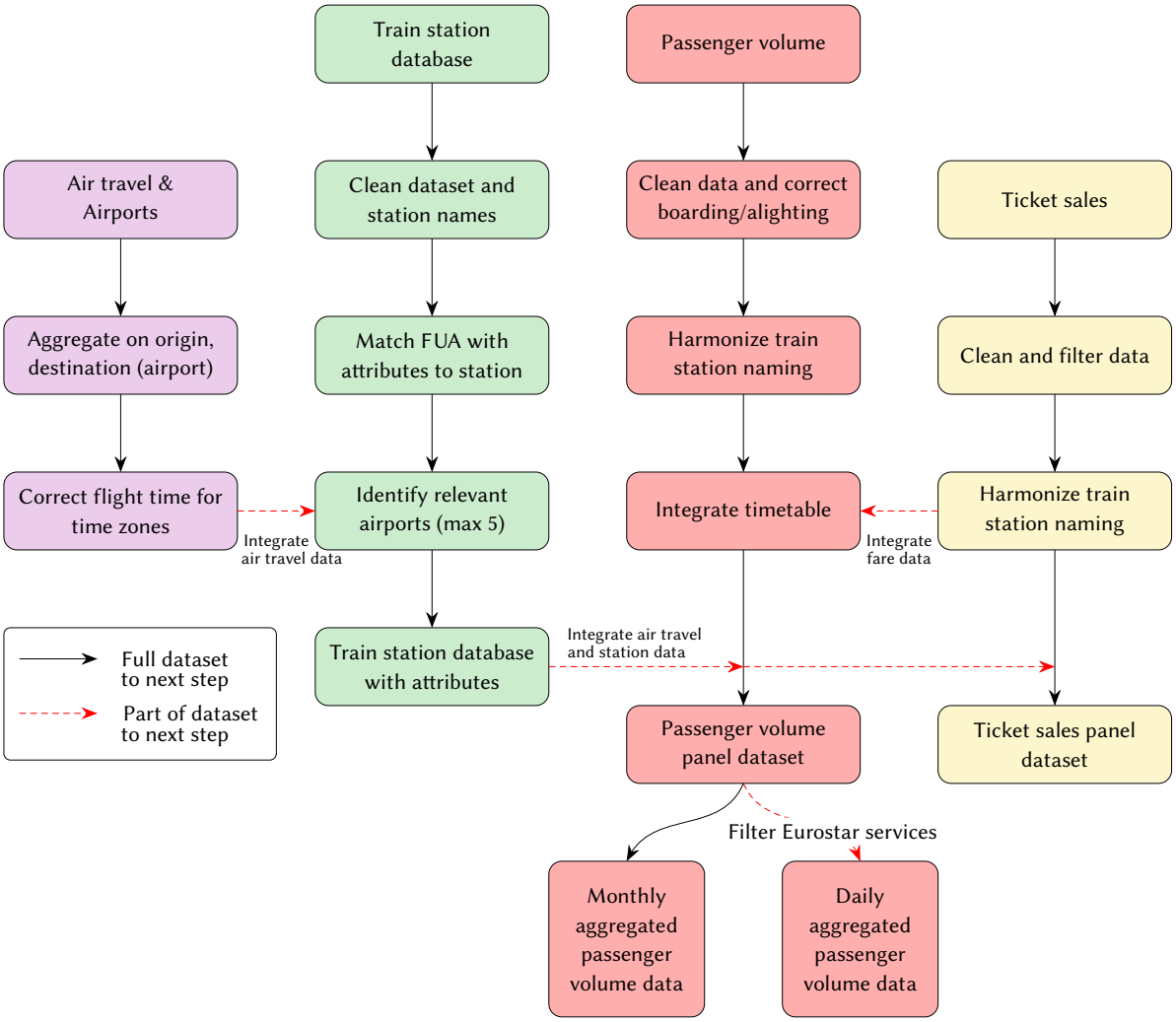


Figure 4.2: Data processing steps

4.3. OD pair example

This section shows an example of what one observation from the processed dataset looks like based on the OD definition defined in section 3.2. This is an observation from the daily passenger volume dataset. Table 4.2 shows an example of the OD pair Amsterdam Centraal and Berlin Hbf with all the corresponding attributes. In this example there are 2 airports close to Amsterdam Centraal and 1 close to Berlin Hbf. For this example the number of travellers is from the ICE Berlin dataset, but it is important to note that this journey (and many more OD combinations) is also possible by using regional services and transferring. This means that the actual number of travellers between this origin and destination is likely higher.

Category	Variable	Origin	Destination
OD characteristics			
	Station	Amsterdam Centraal	Berlin Hbf
	Longitude, Latitude	52.38, 4.90	52.53, 13.37
	FUA code	NL001F (Amsterdam)	DE001F (Berlin)
	Population	2,988,878	5,106,083
	GDP per capita (€)	79195	53876
Airports			
	Airport 1	AMS	BER
	Airport 2	RTM	N/A
	Airport 3	N/A	N/A
	Airport 4	N/A	N/A
	Airport 5	N/A	N/A
Travel characteristics			
	Travel date		2024-03-17
	Rail travel time (minutes)		341
	Number of travellers		383
	Distance (km)		574
	Rail fare (€)		44.99
	Transfers		0
Rail service attributes			
	Rail frequency		6
	Seat capacity		587
	Train service		ICE Berlin
Air travel characteristics			
	Air travel frequency		8.2
	Air fare (€)		108

Table 4.2: Example of an origin-destination observation from the passenger volume dataset

4.4. Descriptive statistics and visualisations

This section shows the descriptive statistics about the historical passenger volume data as well as the ticket sales data. This is used to get a better understanding of the properties of the data and can help by interpreting the results of the model.

4.4.1. Temporal variation in supply

This section describes the variation of rail service supply over time for the train services for which historical passenger data is known (see Table 4.1), and only within the study period. The variation is summarized in Table 4.3 and Figure 4.3.

- IC(E) Berlin

From the timetable of 2024 (starting in December 2023) some changes were made to the IC Berlin which resulted in a reduction of travel time of 33 minutes between Amsterdam and Berlin. This was partly achieved by removing one stop in The Netherlands and three stops in Germany ("Binnen 6 uur van Amsterdam naar Berlijn | NS International", n.d.). In 2025 the rolling stock was gradually changed from intercity to a high speed trainset which from September 2025 fulfilled all services. This meant the capacity decreased from 587 to 439 seats. In addition the service type was changed from IC (intercity) to ICE (intercity express) which resulted in a fare structure change but during this change the timetable remained the same.

- Travel time variation due to detour ICE Frankfurt

For four months in 2025, the ICE to Frankfurt (as well continuing on to Munich), had to detour due to construction works resulting in increased travel time ("Improvements to German railways tracks | NS International", n.d.) as well serving different stops in The Netherlands. This increased the travel time between The Netherlands and

Germany by 24 minutes. This also lead to the departure times being shifted throughout the day, although not all departures were shifted by the same amount.

- ICE destination change

Until July 2024, NSI ran a direct ICE service to Basel. In the timetable of 2025 (starting December 2024) a new ICE service was introduced replacing the ICE to Basel with the destination of Munich.

- Eurostar service changes

As shown in Table 4.1, Eurostar ran a seasonal train to Marseille in the summer. However from 2025 this service does not run anymore ("Eurostar Sun | Thalys Zon treinen naar Zuid-Frankrijk", n.d.).

Train service	Variable change	Time period	Change
IC(E) Berlin	Travel time reduction & stops removed	From December 2023	33 minutes reduction
IC(E) Berlin	Capacity decrease & rolling stock change	From September 2025	Seat capacity reduction of 148 seats
ICE Frankfurt	Travel time increase & stops changed	April, July, August and October 2025	24 minutes increase and routing via Venlo
ICE Basel/Munich	Destination change	Until July 2024 and from December 2024	Destination change
Eurostar Sun	Termination of service	Until 2024	Service terminated

Table 4.3: Temporal variation of rail service attributes in the passenger volume dataset

Figure 4.3 shows a timeline on when the changes in supply characteristics occur in the dataset.

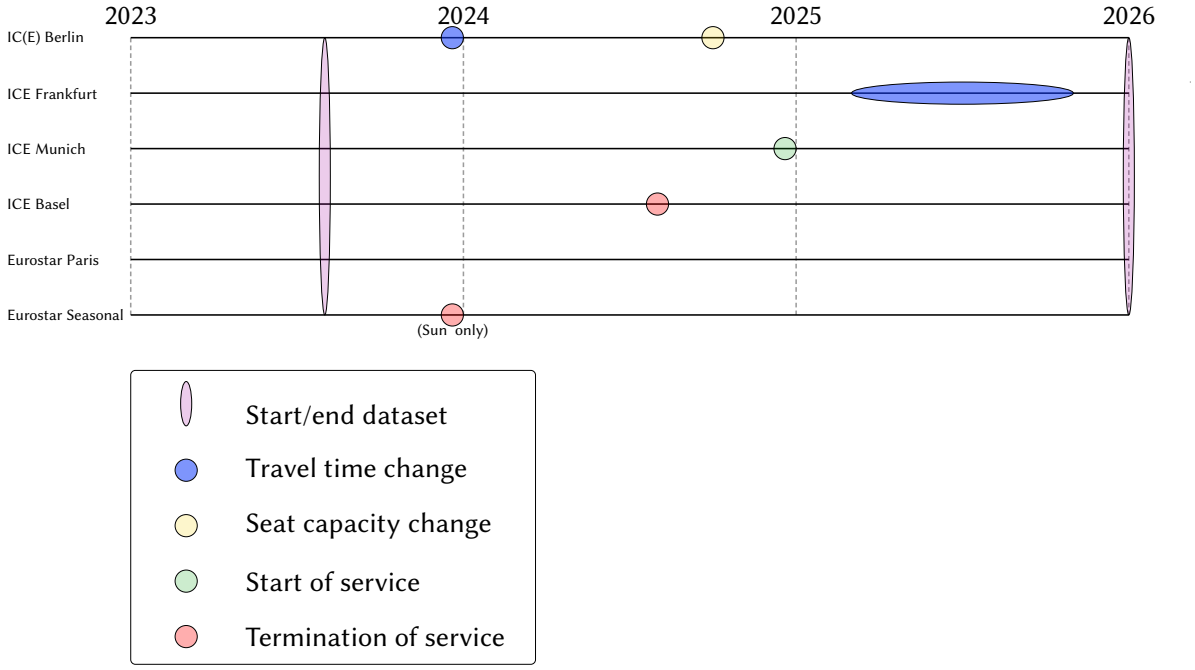


Figure 4.3: Timeline of temporal variation of rail service attributes in the passenger volume dataset

4.4.2. Complete passenger volume dataset

This section shows some visualisations of the combined (monthly) passenger volume dataset. This contains the passenger volume data of all the ICE and Eurostar services. Figure 4.4 shows the distribution of the total volume of the different OD pairs. The volumes are dominated by the Eurostar services, which are represented

by the six largest OD volumes, after which comes the ICE Berlin and ICE Frankfurt. This is expected, as the Eurostar to Paris is also the most frequent service in the dataset.

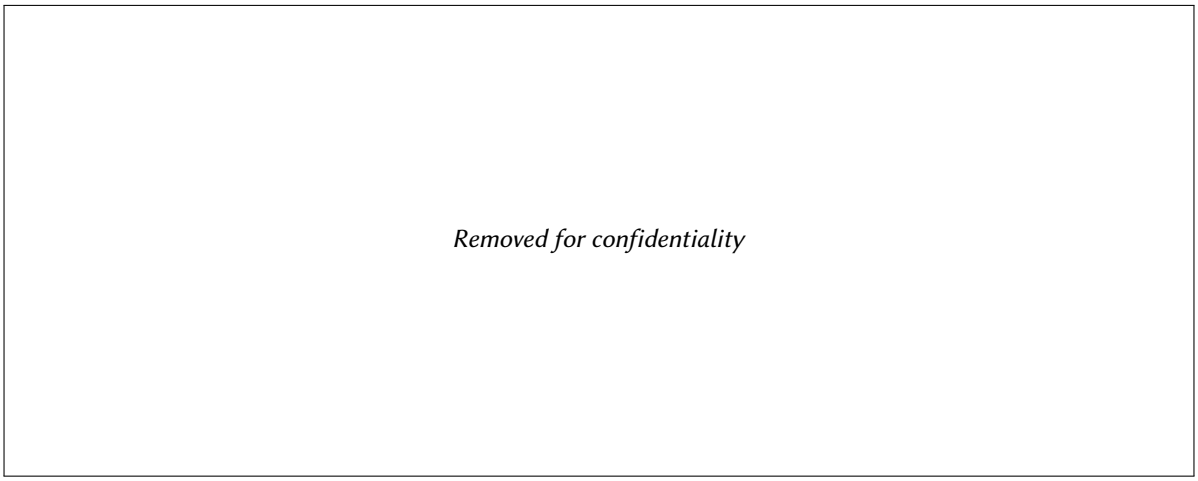


Figure 4.4: Passenger volume by OD pair for the passenger volume dataset

Table 4.4 show the date range for which the data is available as well as the number of observations, train services, origins and destinations. It is clear that the total volume roughly twice as large as for the ticket sales dataset, while the total number of observations and number of OD pairs is much smaller (see Table 4.6).

Observation period	July 2023 - December 2025
Number of observations	106393
Total passenger volume	*Removed for confidentiality*
Number of train services	6
Number of origins	52
Number of destinations	52

Table 4.4: Passenger volume data characteristics

Table 4.5 shows for every train service in the passenger volume dataset, the number of observations and total volume and the corresponding percentages compared to the whole dataset. It shows that the dataset is uneven. The Eurostar Paris corresponds to more than half the total volume with only 0.8 % of the observations, although it is important to note that the small number of observations is mostly driven by the monthly aggregation.

Train service	Observations	% Observations	Total Volume	% Volume
Removed for confidentiality				

Table 4.5: Train services and corresponding number of observations and volume

4.4.3. Historical ticket sales

Figure 4.5 shows from which origins (green) and to which destinations (red) tickets have been sold. This shows that the destinations are dominated by Belgium, Germany, France, Switzerland and Austria. While some other countries (such as Spain, Italy or Poland) do receive booking to a limited number of stations, the amount of stations where a ticket has been booked to is very limited. This is party due to NSI not being able to sell tickets to all possible stations in those countries, as well as them being further away from The Netherlands, and thus

being less likely to be booked through NSI. Table 4.6 shows for which date range the data is available as well as the number of observations, total number of tickets sold and origins and destinations.

Observation period	July 2023 - April 2026
Number of observations	1,309,729
Total number of tickets	*Removed for confidentiality*
Number of origins	395
Number of destinations	5660

Table 4.6: Ticket sales data characteristics

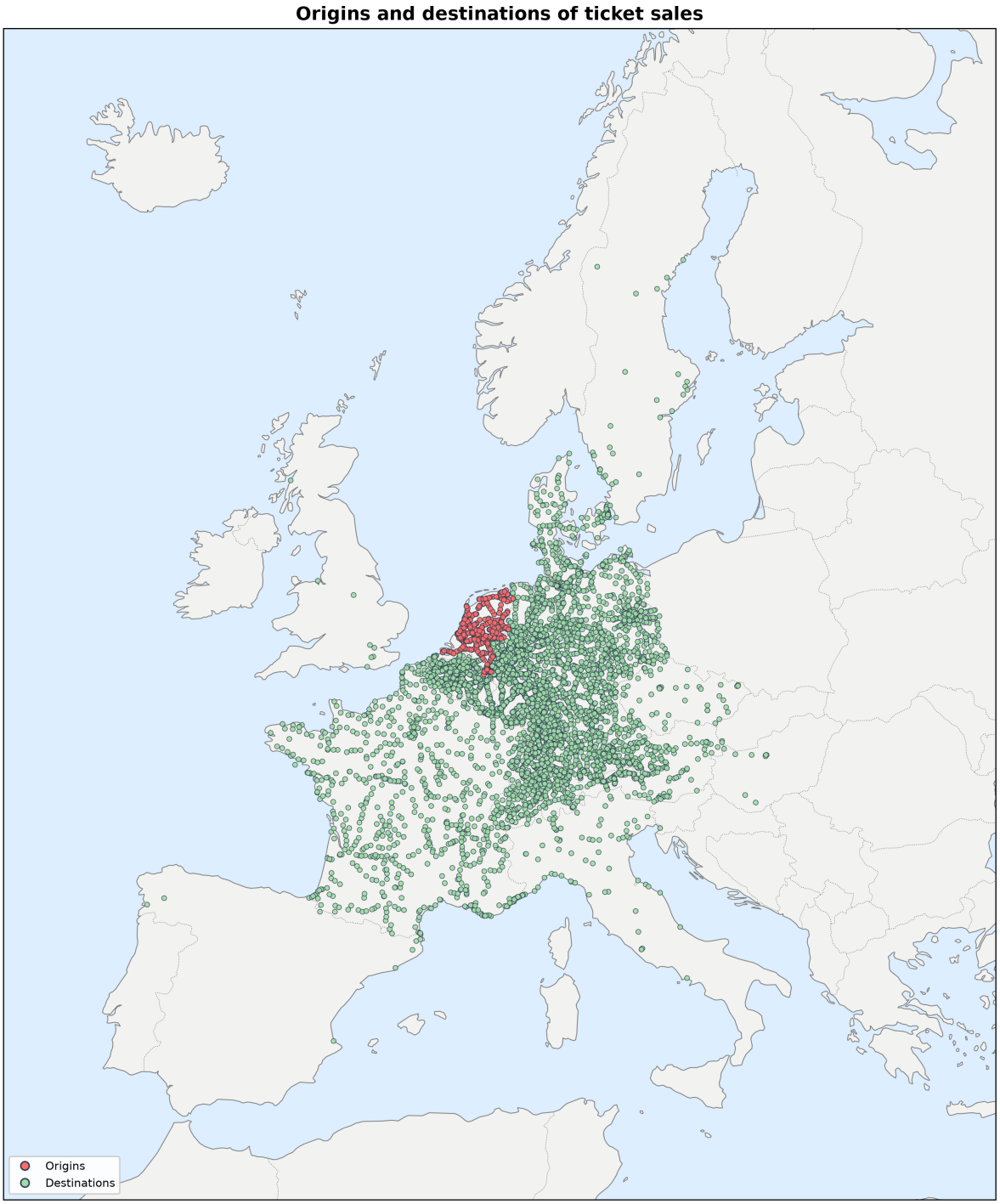


Figure 4.5: Map of origins and destinations for tickets sold by NS International

Figure 4.6 shows which OD pairs are sold the most by NSI. As expected it is dominated by the direct services from The Netherlands, and corresponds to the OD pairs seen in the passenger volume data. OD pairs such as Rotterdam Centraal to Antwerpen Centraal rank even higher here in then in the passenger volume data. This is due to the passenger volume data only containing Eurostar services for this route while this route is also served by very frequent Eurocity and Eurocity direct services, which results in many tickets being sold for the OD pairs which both services serve. The most popular destination which does not receive direct service from The Netherlands is Brugge in Belgium.

Removed for confidentiality

Figure 4.6: Passenger volume by OD pair for the ticket sales dataset

Figure 4.7 shows the number of tickets sold per OD pair against the geographical (straight-line) distance, together with the total number of tickets sold per distance bin (set at 20km). Both the individual observations and the totals decrease with distance: the totals fall drastically between the peak at roughly 130 to 170 km and 1250 km, beyond which only a few tickets are sold. This is in line with the expectation of diminishing amount of tickets sold as both travelling gets less attractive and travellers are less likely to book (their whole trip) through NSI. The decline is not smooth, and several spikes correspond to a small number of popular destinations. For example, the peaks around 170 and 130 km correspond to Brussels (from Amsterdam and from Rotterdam respectively), and the peak around 430 km corresponds to Paris.

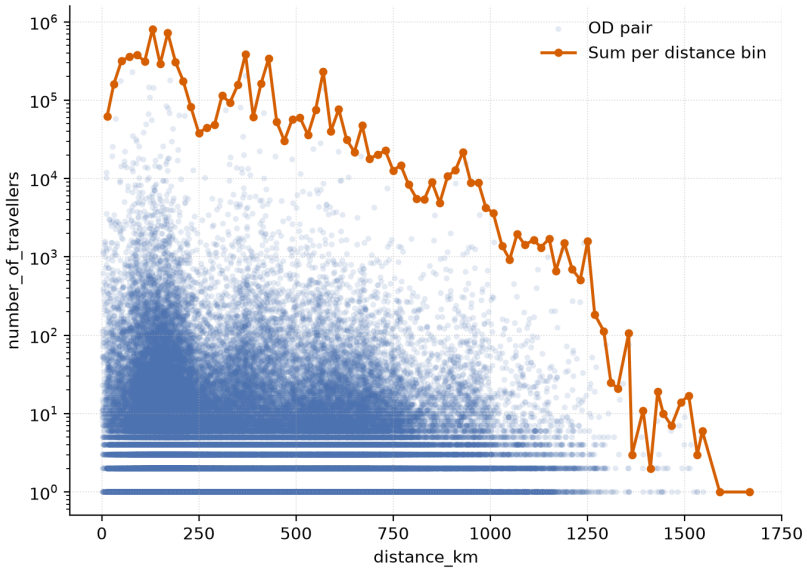


Figure 4.7: Scatterplot of tickets sold vs distance

4.5. Processed datasets

As mentioned in section 4.2, two datasets are created which are used for analysis. One where ticket sales is the measurement for demand, and the other passenger volume. Other than demand measurements the other variables remain the same, although not all of them are available in both datasets. In this section the two datasets that were created by merging the different data sources are described. First the variables that were derived from the data are described, the share of NSI ticket sales is shown, correlations and multicollinearity are discussed, distributions are shown and finally the data limitations are discussed.

4.5.1. Variable construction

Table 4.7 shows which variables were derived from the data. They are based on the explanatory variables introduced in section 3.1.

Variable	Description	Resolution	Notation	Unit
Demand measurements				
Passenger volume	Number of passengers between OD pair during a certain time period	OD daily/monthly	V_{ijt}	passengers
Ticket sales volume	Number of tickets sold between specific OD pair on a specific date	OD daily	T_{ijt}	tickets
Origin-destination characteristics				
Population	Number of inhabitants per FUA in a specific year	Yearly per FUA	pop_{it}	inhabitants
GDP per capita	Gross domestic product per capita per FUA in a specific year	Yearly per FUA	gdp_{it}	€
Distance	Geographical distance between every station pair	OD pair	D_{ij}	km
Rail service attributes				
Rail fare	Average price paid for train ticket between an OD pair on a specific date sold	OD daily	$f_{rail,ijt}$	€
Rail travel time	Travel time between OD pair including transfers (if applicable) on a specific date	OD pair	$TT_{rail,ijt}$	minutes
Rail frequency	Number of direct trains between an OD pair on a specific day	OD pair	$freq_{rail,ijt}$*	trains/day
Seat capacity	Number of seats available (all fare classes combined) on a specific train between OD pair	OD daily	cap_{ijt}*	seats/train
Transfers	Number of transfers in itinerary on ticket sold on specific OD pair and day	OD daily	TR_{ijt}**	Number of transfers
Air travel characteristics				
Air frequency	Average number of flights between OD airports	OD pair	$freq_{air,ij}$	flights/day
Air fare	Average monthly fare between OD airport pairs, for all included carriers	OD monthly	$f_{air,ijt}$	€

Table 4.7: Overview of available variables. *: only in passenger volume models, **: only in ticket sales models

4.5.2. Data coverage

This section shows what the share of ticket sold by NSI is, as well as how much of the ODs and corresponding data lies within FUAs and/or has air travel alternatives available.

Ticket sales share

Figure 4.8 shows the OD pairs with the highest volume and for which both passenger volume and ticket sales data are available. Belgium is not shown as this does not give useful insight into what the share of tickets sold by NSI is because there are other competing services to Belgium (Eurocity and Eurocity direct) and thus the share of Eurostar tickets sold by NSI is not known. Figure 4.9 shows the normalized version of the same graph where instead of the volume the shares are shown. Like for Belgium for some other OD pairs other services can also be available which could result in a share that is higher than actual. For example, for the OD pair Arnhem centraal - Düsseldorf Hbf it is possible to take the ICE or a regional train and NSI can sell tickets for both although only the volume for the ICE is known and thus resulting in the figure showing a higher share than actual.

The forecasting methodology (see section 3.8) requires the share of tickets sold by NSI in order to retransform the predicted number of tickets sold into absolute demand. This share is therefore determined for OD pairs longer than 300km. For these OD pairs NSI accounts for 25.4% of tickets sold. The 300 km threshold is used

because it excludes OD pairs between the Netherlands and Belgium (where NSI share is relatively high), and it fits the distances used in the case study. The resulting share is likely still an overestimate. As all lines in the passenger volume data are direct services from The Netherlands, which are also the most sold services by NSI, whereas for OD pairs in general, and especially for itineraries with transfers, the NSI share is likely lower.

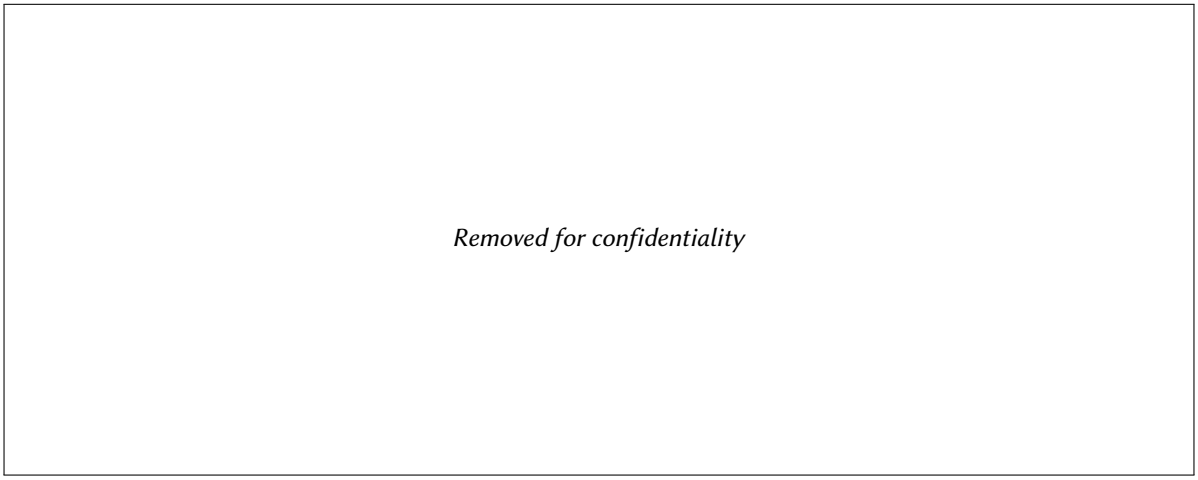


Figure 4.8: Coverage of ticket sold by NS International compared to total passenger volume

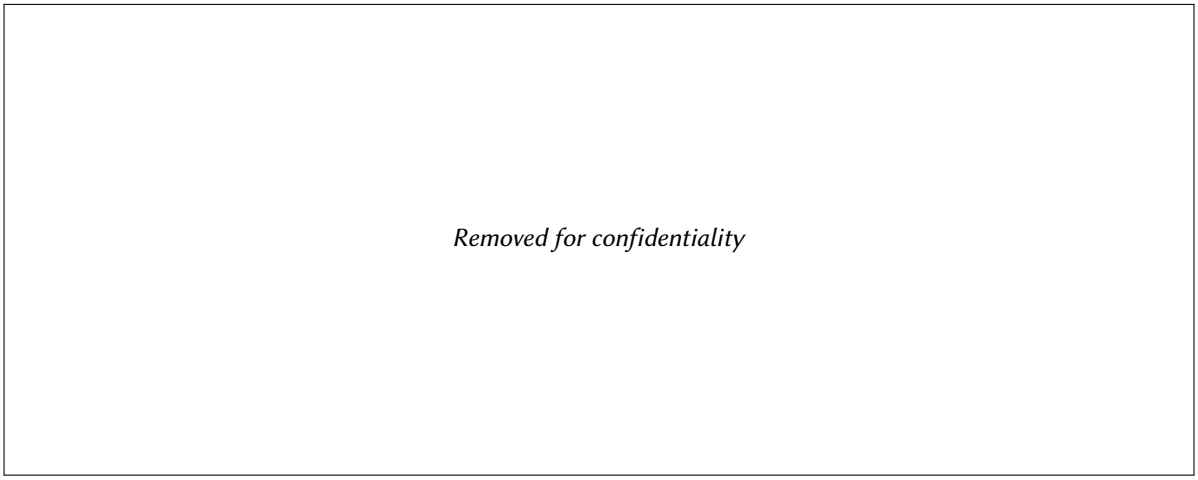


Figure 4.9: Normalized coverage of ticket sold by NS International compared to total passenger volume

FUA and air travel coverage

Table 4.8 show how much of the passenger volume data lies outside of FUAs or has no air travel alternative available (0 frequency). Although 15% of the OD observations lies outside of FUAs, only 2 % of the passengers make use of these stations showing that only a small amount of the data lies outside of FUAs. The same can be seen for the amount of data without air travel alternatives. Since most passengers use the larger stations, which generally have an airport nearby, only 7 % of the passengers have no air travel alternative even though 45% of the OD observations does not have any air travel alternative available.

(a) OD pairs outside Functional Urban Areas (FUAs)			
Category	**Outside FUA**	**Total**	**Share**
OD observations	*Removed for confidentiality*		15.37%
Passengers	*Removed for confidentiality*		2.10%
Stations	*Removed for confidentiality*		15.38%
(b) OD pairs without an air travel alternative			
Category	**No air alternative**	**Total**	**Share**
OD observations	*Removed for confidentiality*		44.97%
Passengers	*Removed for confidentiality*		7.07%

Table 4.8: Coverage of passenger volume observations by Functional Urban Area (FUA) coverage and air travel availability

Table 4.9 show how much of the ticket sales data lie outside of functional urban areas or has no air travel alternative (0 frequency). The same pattern as for the passenger volume data occurs, where more observations lie outside of FUAs than the share of passengers. For the ticket sales data a larger share of the data lies outside of FUAs and a larger share has no air travel alternative. This is due to the the ticket sales data containing a lot more stations, and the passenger volume data only containing stations which serve long-distance train services, which are more likely to lie within a FUA and have air travel alternatives available.

(a) OD pairs outside Functional Urban Areas (FUAs)			
Category	**Outside FUA**	**Total**	**Share**
OD observations	*Removed for confidentiality*		17.96%
Passengers	*Removed for confidentiality*		7.37%
Stations	*Removed for confidentiality*		41.95%
(b) OD pairs without an air travel alternative			
Category	**No air alternative**	**Total**	**Share**
OD observations	*Removed for confidentiality*		42.59%
Passengers	*Removed for confidentiality*		28.77%

Table 4.9: Coverage of ticket sales observations by Functional Urban Area (FUA) coverage and air travel availability

4.5.3. Correlations and multicollinearity

This section shows the correlation matrices which indicate pairwise correlations and the variance inflation factor. They are used to check if certain variables should be interpreted with caution and if there are no problematic correlations.

Correlation matrices

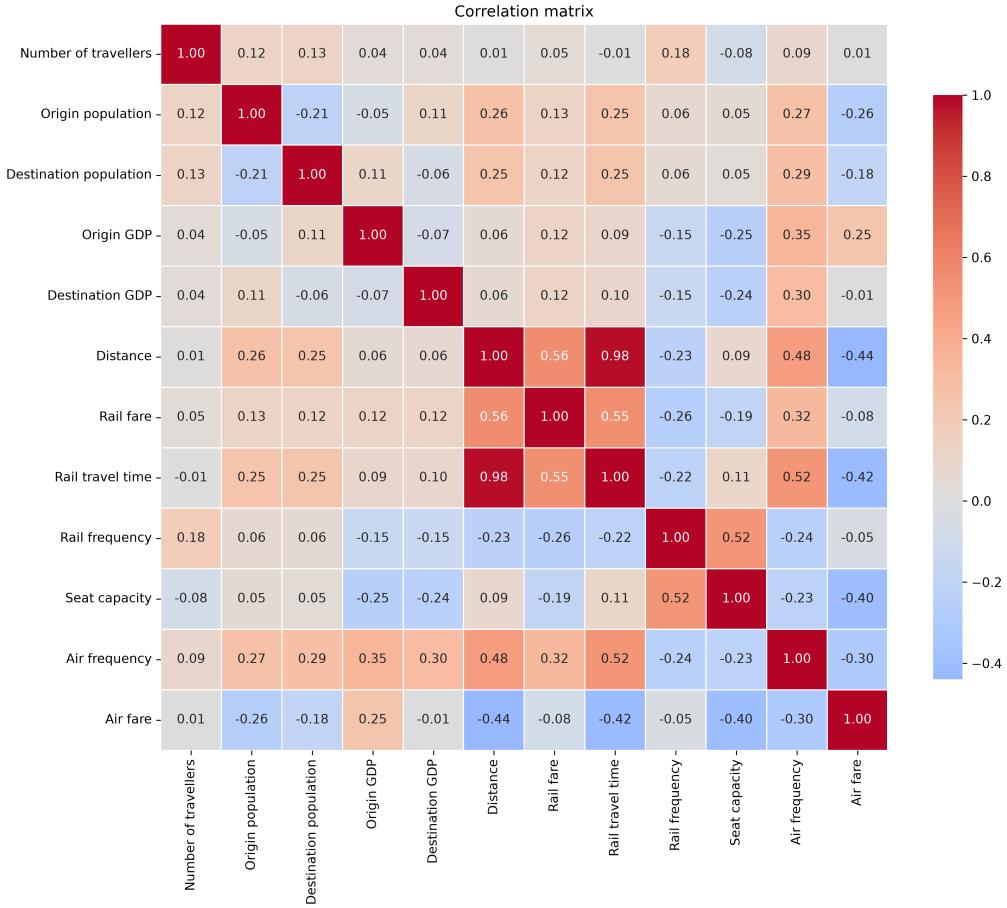


Figure 4.10: Correlation matrix of the passenger volume data

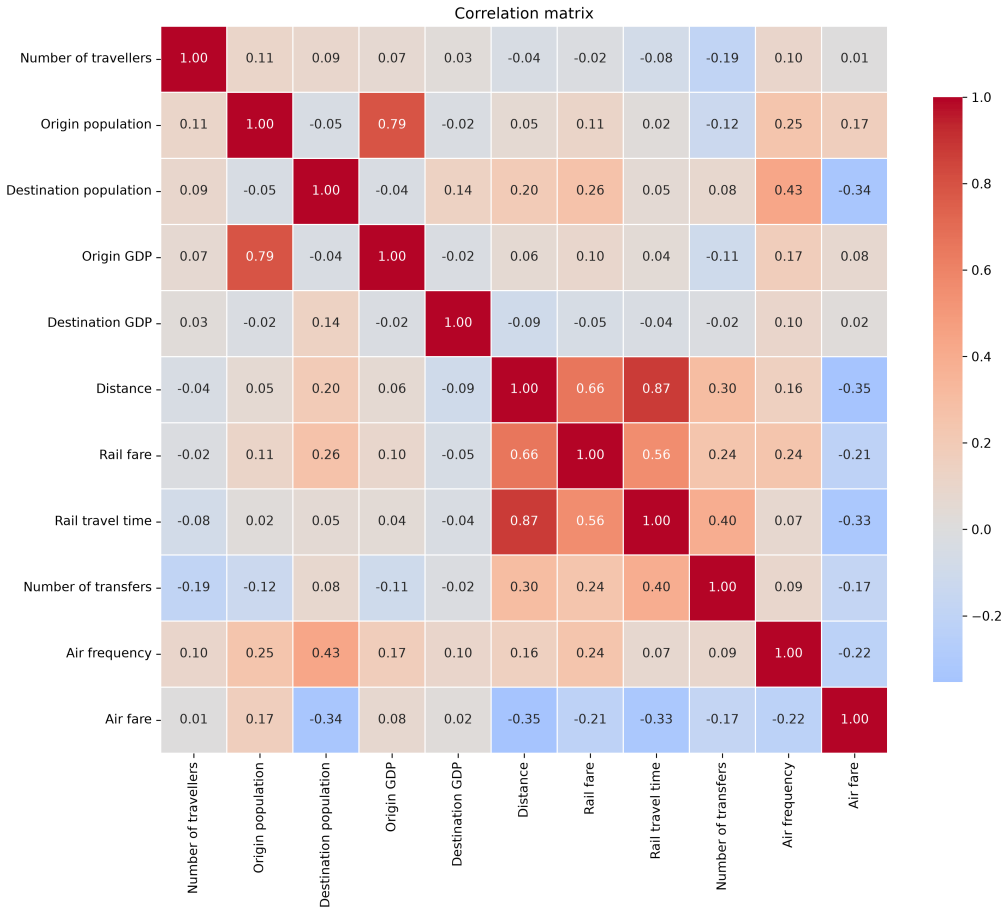


Figure 4.11: Correlation matrix of the ticket sales data

The correlation matrices show (Figure 4.10 and 4.11) the pairwise correlations for both datasets. In both cases the highest correlation is between rail travel time and distance, which is expected. In addition rail fare and distance are also highly correlated. An interesting difference is the correlation between origin GDP and origin population compared between both datasets. In the ticket sales dataset there is a high correlation (0.79) while for the passenger volume data the correlation is quite low (0.11). Air travel time was omitted from the models as it showed perfect correlation with distance.

Variance inflation factor

As described in section 3.6 the variance inflation factor (VIF) can be used to determine whether there is multicollinearity between the variables and if the coefficients become unstable. It describes how much of the variation of the variable is explained by the other regressors. Any value above 5.0 is reported.

Passenger volume data:

- Distance: 29.71
- Rail travel time: 30.37

Ticket sales data:

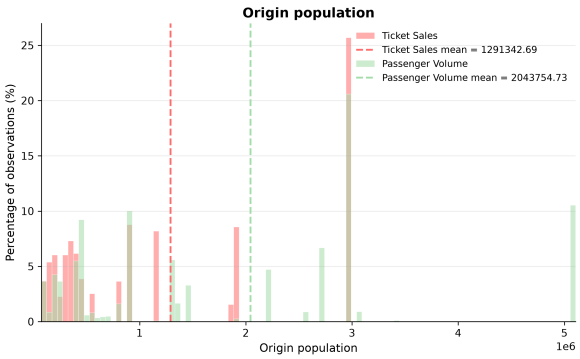
- Distance: 9.07
- Rail travel time: 7.22

Distance and rail travel time both show a high VIF, especially in the passenger volume data, where both values are around 30. In the ticket sales data the values are lower but still above the reporting threshold. As a result, their individual coefficients cannot be reliably separated in models where both are included.. Multicollinearity

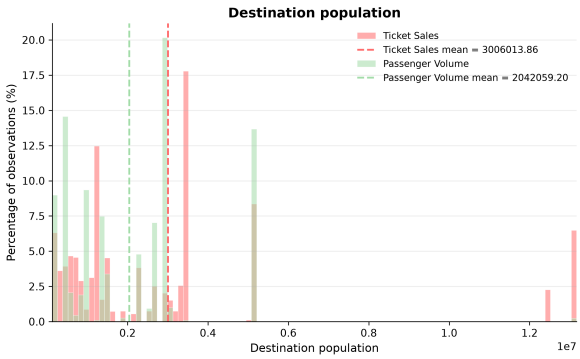
does not negatively influence the fit of the model, so it mainly affects the interpretation of the individual distance and travel time coefficients. No other variables have a VIF above 5.0.

4.5.4. Variable distributions

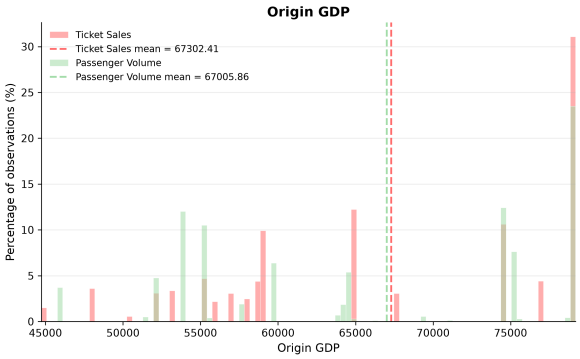
This section shows how the variables are distributed. For every variable, the distribution of a variable in both datasets is shown in one figure together with the mean. Overall the distributions show that the ticket sales data has a much wider range of values, and is not as biased by a few select OD pairs with a very large volume as in the passenger volume data. This is mainly due to the much larger number of OD pairs in the data.



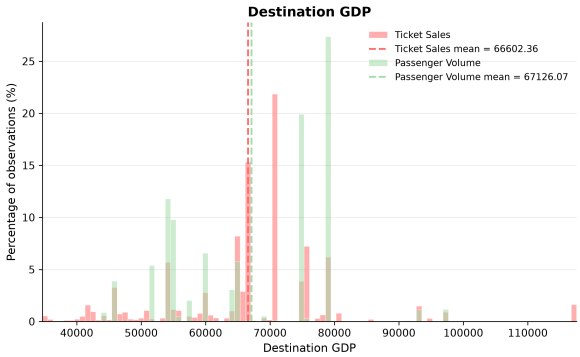
(a) Distribution of the origin population



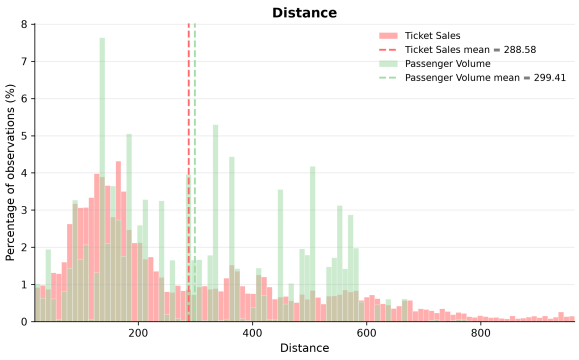
(b) Distribution of the number of the destination population



(c) Distribution of the origin GDP

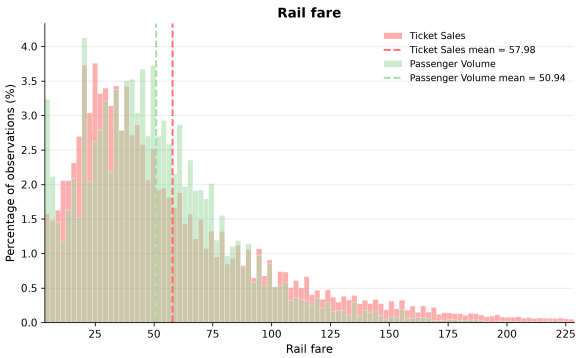


(d) Distribution of the number of the destination GDP

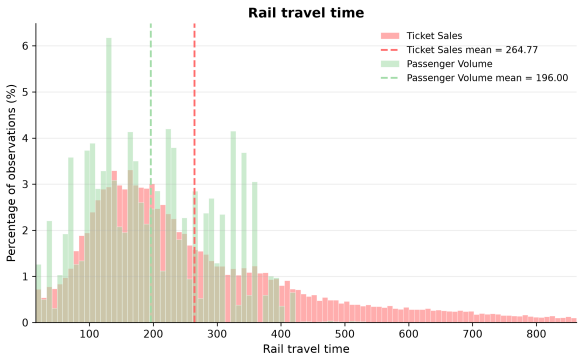


(e) Distribution of the distances

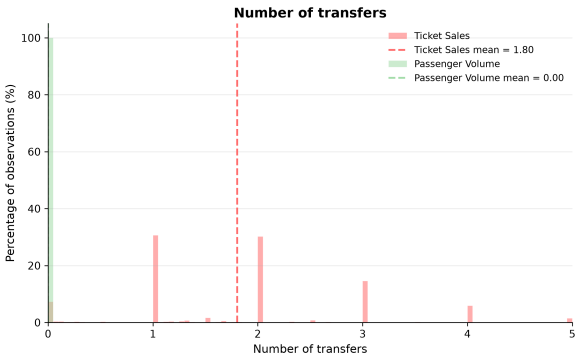
Figure 4.12: Distributions of the OD characteristics in both datasets



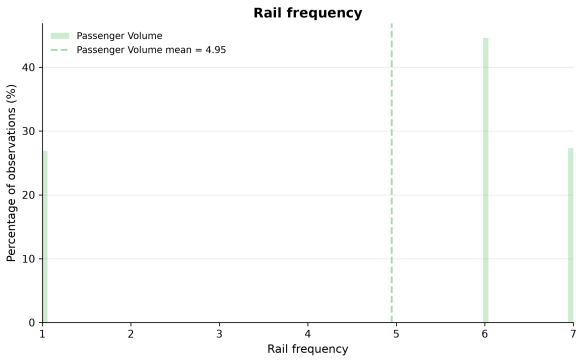
(a) Distribution of the rail fare



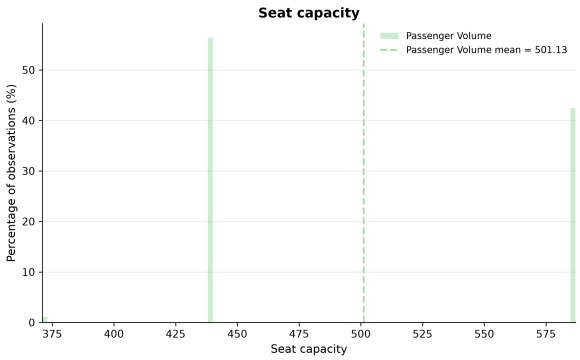
(b) Distribution of the rail travel time



(c) Distribution of the number of transfer

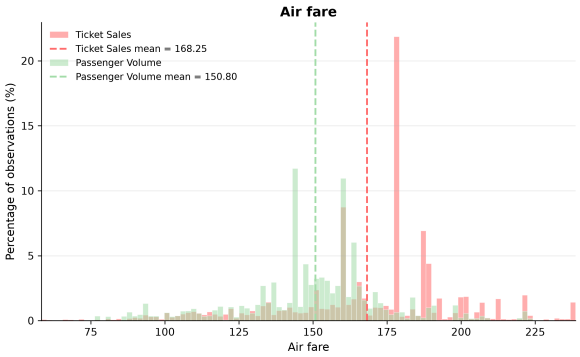


(d) Distribution of the rail frequency (passenger volume data only)

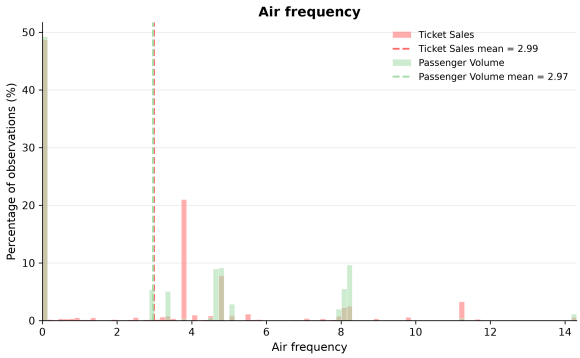


(e) Distribution of the rail seat capacity (passenger volume data only)

Figure 4.13: Distributions of the rail service attributes in both datasets



(a) Distribution of the number of the air fare



(b) Distribution of the number of the air frequency

Figure 4.14: Distributions of the air travel characteristics in both datasets

4.5.5. Data Limitations

As a result of the nature of the data as well as the data processing, there are a number of limitations which limit the quality of the analysis. These limitations are grouped into four categories: fare data, station and OD data, explanatory variables, and data coverage and resolution.

Fare data

- Using A-B as B-A fare for fare matching

The passenger volume dataset use the fare data from the ticket sales dataset because the passenger volume dataset does not include fare directly. To do this, it was assumed that the fare for station A to station B was the same for station B to station A on a specific date. This was done because the fare data is only available for origins in The Netherlands and destination outside of The Netherlands. For the Eurostar services fare data was already included in the passenger volume dataset, and therefore fare matching was not necessary and fare data is available for A-B and B-A separately.

- Using nearest date for fare matching

For the matching of the fare data to the volume data, for every entry in the volume dataset, the nearest date was found in the ticket sales data where for the same OD pair a ticket was sold, and this was used as the fare. This meant that for some observations, the nearest date with fare for the same OD pair, the nearest date was far away, and therefore this fare might not always be a representative fare.

- Average fare data

The available fare data is an average fare paid per traveller for a specific OD pair on a specific date. This means that there is no differentiating between fare classes as well as time of booking. Some services use dynamic pricing structures which are less expensive when booking far ahead but more expensive when booking close the departure date which could skew the average fare paid.

- Air fare on monthly level

The air fare data is only available at the monthly aggregated level which means that same value is used for all observations within one month.

Station and OD data

- Correction of boarding and alighting station with coordinates

For the ICE services, a correction of the boarding and alighting station needed to be made when the journey started or ended in The Netherlands. This was done based on the coordinates, where the it was assumed that the a traveller used the nearest station to board or alight the train. However, the geographically nearest station might not always make the most sense, which limits the accuracy of the boarding and alighting station.

- Name matching of train station

For the ticket sales data, to get coordinates for all the stations, the train station database was used. However, the naming of train station in the ticket sales data and train station database was not the same. Therefore fuzzy string matching was applied, where a 90% threshold was used to automatically map stations. This threshold was set to limit the amount of manual mapping needed, which was roughly 500 stations. However, this meant that there could wrongly mapped stations done by the automatic mapping, leading to ticket sold being mapped to the wrong station and therefore skewing the data.

Explanatory variables

- Construction works not included in timetable

On the railway there are often construction works. This can result in train services being rerouted, re-timed, shortened or cancelled. As most construction works is planned far in advance, it is usually known when booking. However, it was not known when construction took place on lines affecting the train services with passenger volume data, and therefore not integrated into the timetable data. This means that a change in passenger volume might be because of construction works, without the service change being reflected in the service supply attributes.

- Availability explanatory variables ticket sales and volume data

Due to the nature of the ticket sales and passenger volume data, not all variables are available for both datasets. For the ticket sales data it is not known what route travellers used, and therefore timetable data such as frequency is also not available. And for the volume data, fare data is not available except for Eurostar.

- FUA data not always complete and available

The data used for the functional urban areas (GDP and population) is not always available for the same time frame as the passenger volume or ticket sales data. This could limit the quality of the results but is not expected to have a large impact as these variables exhibit less variation over time then passenger volumes or ticket sales.

- Limited variation for estimation of fixed-effects

A number of variables in both datasets exhibit a limited or no amount of variation within OD-pairs which makes it hard or impossible to estimate the effects with the fixed-effects models. Population shows very minimal amount of variation over time which inflates the coefficients making them implausible. GDP shows variation over time but no recent data with variation for the same time frame as the observed demand and therefore cannot be estimated. Two variables are completely time-invariant, namely air frequency and distance. Seat capacity and rail frequency are nearly time invariant as described in subsection 4.4.1. The other variables (rail fare, rail travel time, transfers and air fare) do show enough temporal variation to be estimated in fixed-effects.

Data coverage and resolution

- Limited number of train services with passenger volume data

As shown in Table 4.1, there is only a limited number of train services for which passenger volume is available. In addition, the time range for which the data is available is not the same for all train services. This limits the quality of the analysis since there are not as many observations.

- No passenger volume data for services to Belgium

Since Eurostar and other services operated by NSI (Eurocity and Eurocity Direct) both operate and compete between the The Netherlands and Belgium, for only one of them passenger volume data could be provided. Eurostar was chosen since they offer longer distance services, but this does limit the amount of data available.

- Data resolution Eurostar

For Eurostar the passenger volume data is limited to monthly aggregation, limiting the granularity of the analysis and making splitting the passenger volume dataset necessary.

5 Results

This chapter presents the model results: first passenger volume models, then ticket sales models. Discussion of the results follows in chapter 7.

5.1. Passenger volume

This section presents the results from the models using the passenger volume data, first for the daily data in Table 5.1 and in Table 5.2 for the monthly aggregated data.

5.1.1. Daily results

Variable	Pooled panel model Parameter	p-value	OD fixed-effects Parameter	p-value	OD time fixed-effects Parameter	p-value
Model constant						
α	-65.639	0.0000	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	1.4983	0.0199	-	-	-	-
$\delta_{ICE,Basel}$	15.327	0.1784	-	-	-	-
$\delta_{ICE,Munich}$	15.109	0.1836	-	-	-	-
$\delta_{Eurostar,Paris}$	-	-	-	-	-	-
$\delta_{Eurostar,Seasonal}$	-	-	-	-	-	-
OD characteristics						
$\beta_{pop,i}$	0.8462	0.0000	2.0106	0.7663	-7.6397	0.6665
$\beta_{pop,j}$	0.8625	0.0000	-2.9703	0.7031	-11.307	0.5221
$\beta_{gdp,i}$	1.4642	0.0496	-	-	-	-
$\beta_{gdp,j}$	0.9409	0.1738	-	-	-	-
β_{D}	-0.0128	0.9923	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	0.0514	0.3734	-0.0086	0.5271	-0.0422	0.0035
$\beta_{TT_{rail}}$	-0.3104	0.8439	1.2398	0.0638	1.1592	0.1335
$\beta_{freq_{rail}}$	8.6249	0.1562	0.9053	0.0000	0.9091	0.0000
β_{cap}	-0.0049	0.9849	-0.3582	0.0181	-0.8543	0.0000
β_{TR}	-	-	-	-	-	-
Air travel characteristics						
$\beta_{freq_{air}}$	0.1794	0.5823	-	-	-	-
$\beta_{f_{air}}$	0.5419	0.0158	0.2655	0.0007	-0.0847	0.3980
Model statistics						
Unit of analysis	OD-day passengers		OD-day passengers		OD-day passengers	
Observations	38826		38826		38826	
R^2	0.4466		0.2803		0.3121	
Within R^2	0.2670		0.2803		0.2770	

Table 5.1: Results for models for daily passenger volume data, excluding Eurostar (green: $p \leq 0.01$, yellow: $0.01 \leq p \leq 0.05$)

Table 5.1 shows the results of the three models estimated on the daily passenger volume data, thereby excluding Eurostar data. For the dummy variables the ICE Berlin was used as a baseline, so the the dummies presented in the table are compared to the ICE Berlin. For every variable the coefficient and significance is shown using the p value. In the fixed-effects models only the variables that vary over time remain as all other time-invariant OD characteristics are absorbed in the models. For every significant variable the sign and magnitude is also compared with the hypotheses formulated in subsection 3.1.4.

Service type dummies
For the service type dummies, only the ICE Frankfurt dummy is statistically significant (1.498, p = 0.020) relative to the ICE Berlin baseline. The ICE Basel and ICE Munich dummies have large point estimates (15.327 and 15.109 respectively) but are far from significant (p = 0.178 and p = 0.184), indicating large standard errors.

OD characteristics
The population elasticities ($\beta_{pop,i}$, $\beta_{pop,j}$) are strongly significant and close to 1.0 in the pooled model, showing that population size of the origin and destination are a strong determinant of demand between OD pairs. This is agreement with hypothesis H1. In both fixed-effects specifications, the population coefficients are much larger, lose the expected sign and become insignificant. In the fixed-effects specifications only within-OD variation in population over time is used and this variation is not enough to get plausible estimates.

GDP per capita shows an asymmetric pattern in the pooled model, the origin ($\beta_{gdp,i} = 1.464$) is significant at the 5% level (p = 0.050), while the destination coefficient ($\beta_{gdp,j} = 0.941$) is not (p = 0.174). The origin coefficient is larger than hypothesis H2 but it is the expected sign. No recent data on GDP per capita was available so therefore there is no variation over time and thus it was not estimated in the fixed-effects models. Distance (β_D) is close to zero and far from significant in the pooled model (p = 0.992), thus not explaining demand in the model, distance is obviously time-invariant and therefore absorbed in the fixed-effects models.

Rail service attributes
Rail fare ($\beta_{f_{rail}}$) is not significant in the pooled model (0.051, p = 0.373) or the OD fixed-effects model (-0.009, p = 0.527), but is negative and significant in the OD+time fixed-effects model (-0.042, p = 0.004). This is the only specification in which the fare coefficient carries the expected negative sign hypothesized in H4, although not as strong as expected. This suggests that the demand reducing effect of higher fares is identified only once both OD-specific and time-specific unobserved heterogeneity are controlled for.

Rail travel time ($\beta_{TT_{rail}}$) is not significant in the pooled model (p = 0.844), where it carries a positive sign, which is unexpected. Rail capacity (β_{cap}) is not significant in the pooled model but is negative and significant in both fixed-effects specifications (-0.358, p = 0.018 and -0.854, p < 0.001), indicating that within a given OD pair, increases in capacity coincide with lower predicted volume once other time-varying factors are absorbed for. This is not expected as according to H7 the expected sign is positive.

In all three specification, the coefficient on rail service frequency ($\beta_{freq_{rail}}$) is positive and highly significant in the OD fixed-effects and OD+time fixed-effects models (0.905 and 0.909 respectively, p < 0.001), indicating that within a given OD pair, an increase in the number of daily services is associated with more passenger volume. The sign is in line with hypothesis H6, although the magnitude is larger. This suggests that the demand is mostly driven by changes in frequency within-OD over time rather than from difference in frequency between OD pairs.

Air travel characteristics
Air fare ($\beta_{f_{air}}$) is positive and significant in both the pooled model (0.542, p = 0.016) and the OD fixed-effects model (0.266, p = 0.001), suggesting that air and rail act as substitutes: a higher competing air fare is associated with higher rail demand. The positive air fare coefficient agrees with hypothesis H10 is This relationship is no longer significant in the OD+time fixed-effects model (-0.085, p = 0.398). Air frequency is not significant in the pooled model.

Model fit
In terms of overall fit, the pooled model has the highest R^2 (0.447), followed by the OD+time fixed-effects model (0.312) and the OD fixed-effects model (0.280). The within-R^2 values are more comparable across specifications (0.267, 0.280, and 0.277 respectively), indicating that the fixed-effects specifications explain a similar share of within-OD variation as the pooled model, despite absorbing the time-invariant OD characteristics that contribute to the pooled model's higher overall R^2.

5.1.2. Monthly results

Variable	Pooled panel model Parameter	p-value	OD fixed-effects Parameter	p-value	OD time fixed-effects Parameter	p-value
Model constant						
α	-44.862	0.0005	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	0.4868	0.4644	-	-	-	-
$\delta_{ICE,Basel}$	-0.2947	0.6665	-	-	-	-
$\delta_{ICE,Munich}$	1.2270	0.0437	-	-	-	-
$\delta_{Eurostar,Paris}$	-0.0959	0.8750	-	-	-	-
$\delta_{Eurostar,Seasonal}$	3.9962	0.0000	-	-	-	-
OD characteristics						
$\beta_{pop,i}$	0.8558	0.0000	49.480	0.0001	-14.193	0.6552
$\beta_{pop,j}$	0.6787	0.0002	42.762	0.0006	-19.691	0.5586
$\beta_{gdp,i}$	1.2683	0.0917	-	-	-	-
$\beta_{gdp,j}$	1.2093	0.1035	-	-	-	-
β_{D}	-2.2709	0.0816	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	0.1033	0.8069	-0.5557	0.0338	-0.5156	0.0745
$\beta_{TT_{rail}}$	1.4852	0.1002	3.2648	0.0004	3.2776	0.0289
$\beta_{freq_{rail}}$	1.2371	0.0000	1.0512	0.0000	1.1312	0.0000
β_{cap}	0.0896	0.8009	-0.2731	0.1514	-0.3767	0.0814
β_{TR}	-	-	-	-	-	-
Air travel characteristics						
$\beta_{freq_{air}}$	0.8963	0.1182	-	-	-	-
$\beta_{f_{air}}$	0.4625	0.1757	0.3567	0.0071	-0.1012	0.5369
Model statistics						
Unit of analysis	OD-day passengers		OD-day passengers		OD-day passengers	
Observations	2216		2216		2216	
R^2	0.5923		0.0598		0.0639	
Within R^2	0.1228		0.0598		0.0222	

Table 5.2: Results for models for monthly passenger volume data, including Eurostar (green: $p \leq 0.01$, yellow: $0.01 \leq p \leq 0.05$)

Table 5.2 shows the equivalent results estimated on the monthly passenger volume data, which additionally includes Eurostar services. The much smaller number of observations (2216, compared to 38826) reflects the larger temporal aggregation, and is an important context for interpreting the differences discussed below.

Service type dummies
For the service type dummies, the pattern of significance changes considerably: the ICE Frankfurt dummy, significant in the daily data, is no longer significant here (p = 0.464), while the ICE Munich dummy remains significant (1.227, p = 0.044). The newly included Eurostar seasonal dummy is strongly significant (3.996, p < 0.001), the largest and most significant dummy coefficient in either table, while the Eurostar Paris dummy is not significant (p = 0.875).

OD characteristics
The population elasticities again show a strong, expected-sign relationship in the pooled model ($\beta_{pop,i} = 0.856$, $\beta_{pop,j} = 0.679$, both p < 0.001), similar to the daily results and in agreement with the hypothesis H1. In the OD fixed-effects model, however, both coefficients are now statistically significant but extremely large (49.480 and 42.762, p < 0.001), in contrast to the daily data where the equivalent coefficients were insignificant. In the OD+time fixed-effects model the coefficients revert to being insignificant and negatively signed, consistent with the daily results but still very large. It is expected that these coefficients become very large as they try to capture a lot of variation in monthly demand from very small changes in population.

GDP per capita is no longer significant for either origin or destination in the pooled model (p = 0.092 and p =

0.104, compared to p = 0.050 and p = 0.174 in the daily data), and distance, while still not significant, is closer to significance and negatively signed (-2.271, p = 0.082).

Rail service attributes
Rail fare shows a different pattern of significance: it is significant and negative in the OD fixed-effects model (-0.556, p = 0.034), in which it was not significant in the daily data, while it is only slightly outside significance in the OD+time fixed-effects model (-0.516, p = 0.075), compared to being clearly significant there in the daily data.

Rail travel time is positive and significant in both fixed-effects specifications (3.265, p < 0.001 and 3.278, p = 0.029), a clear difference from the daily data, where this coefficient was not significant in any specification. The positive sign remains very implausible compared to the expected direction under hypothesis H7 in both datasets.

Rail capacity is not significant in any of the three specifications in the monthly data, in contrast to the daily data, where it was negative and significant in both fixed-effects models.

Rail service frequency remains positive and highly significant across all three specifications (1.237, 1.051, and 1.131, all p < 0.001), consistent with the daily data, indicating that this result is robust to the choice of temporal aggregation, and mostly in agreement with hypothesis H6, with the sign being in agreement but the magnitude larger.

Air travel characteristics
Air fare is significant and positive only in the OD fixed-effects model (0.357, p = 0.007), where it does agree with hypothesis H10. Whereas it was significant in both the pooled and OD fixed-effects models in the daily data. Air frequency remains insignificant, as in the daily data.

Model fit
The overall fit of the monthly models differs substantially from the daily models. While the pooled model's R^2 is comparable (0.592 versus 0.447 in the daily data), the fixed-effects specifications explain much less variation in the monthly data: R^2 of 0.060 and 0.064, compared to 0.280 and 0.312 in the daily data. The within-R^2 values are correspondingly lower (0.123, 0.060, and 0.022), suggesting that the fixed-effects specifications are considerably less able to explain within-OD variation at the monthly level than at the daily level, likely reflecting the smaller sample size available at this temporal resolution.

5.2. Ticket sales

Variable	Pooled panel model Parameter	p-value	OD fixed-effects Parameter	p-value	OD time fixed-effects Parameter	p-value
Model constant						
α	-3.9955	0.1698	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	-	-	-	-	-	-
$\delta_{ICE,Basel}$	-	-	-	-	-	-
$\delta_{ICE,Munich}$	-	-	-	-	-	-
$\delta_{Eurostar,Paris}$	-	-	-	-	-	-
$\delta_{Eurostar,Seasonal}$	-	-	-	-	-	-
OD characteristics						
β_{pop_i}	0.0166	0.8127	1.6146	0.0390	2.6731	0.1375
β_{pop_j}	0.0843	0.0076	0.0113	0.0000	0.0034	0.0001
β_{gdp_i}	0.1711	0.5551	-	-	-	-
β_{gdp_j}	0.3670	0.0003	-	-	-	-
β_D	0.2875	0.0132	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	-0.0253	0.3294	-0.1512	0.0000	-0.1479	0.0000
$\beta_{TT_{rail}}$	-0.3684	0.0152	-0.0299	0.3836	-0.0414	0.2414
$\beta_{freq_{rail,ij}}$	-	-	-	-	-	-
β_{cap}	-	-	-	-	-	-
β_{TR}	-0.3755	0.0000	-0.0898	0.0000	-0.1013	0.0000
Air travel characteristics						
$\beta_{freq_{air}}$	0.0927	0.1471	-	-	-	-
$\beta_{f_{air}}$	-0.3093	0.0003	0.2496	0.0000	0.0706	0.0002
Model statistics						
Unit of analysis	OD-day tickets		OD-day tickets		OD-day tickets	
Observations	511743		511743		511743	
R^2	0.1565		0.0223		0.0226	
Within R^2	-0.1041		0.0223		0.0208	

Table 5.3: Results for models for ticket sales data (green: $p < 0.01$, yellow: $0.01 \leq p \leq 0.05$)

Table 5.3 shows the results of the three models estimated on the daily ticket sales data. Unlike the passenger volume data, no service type dummy variables are estimated in this dataset, and rail frequency and capacity are not included as service attributes since they are not available in the data. The number of transfers (β_{TR}) is included as an additional rail service attribute.

OD characteristics

The population elasticities show mixed significance across specifications. The origin coefficient ($\beta_{pop,i}$) is not significant in the pooled model (p = 0.813), becomes significant in the OD fixed-effects model (1.615, p = 0.039), and loses significance again in the OD+time fixed-effects model (2.673, p = 0.138). The destination coefficient ($\beta_{pop,j}$) is significant across all three specifications, though its magnitude drops a lot from the pooled model (0.084, p = 0.008) to the OD fixed-effects model (0.011, p < 0.001) and OD+time fixed-effects model (0.003, p < 0.001). Therefore it somewhat agrees with hypothesis H1.

GDP per capita is significant only for the destination in the pooled model (0.367, p < 0.001). Distance is significant and positively signed in the pooled model (0.288, p = 0.013), and is absorbed into the fixed-effects in both fixed-effects specifications, agreeing with the magnitude of hypothesis H3.

Rail service attributes

Rail fare ($\beta_{f_{rail}}$) is not significant in the pooled model (p = 0.329), but is negative and highly significant in both fixed-effects specifications (-0.151 and -0.148, both p < 0.001). The sign agrees with hypothesis H4 while the magnitude is smaller. Rail travel time is negative and significant only in the pooled model (-0.368, p = 0.015),

and is not significant in either fixed-effects specification. In the pooled model it agrees with hypothesis H5. The number of transfers (β_{TR}) is negative and highly significant across all three specifications (-0.376, -0.090, and -0.101, all $p < 0.001$), indicating that itineraries requiring more transfers are consistently associated with lower ticket sales.

Air travel characteristics
Air fare shows a sign reversal across specifications: it is negative and significant in the pooled model (-0.309, $p < 0.001$), but positive and significant in both fixed-effects specifications (0.250, $p < 0.001$ and 0.071, $p < 0.001$). Therefore it only agrees with hypothesis H10 in the fixed-effects models but not in the pooled OLS model. Air frequency is not significant in the pooled model, the only specification in which it is estimated.

Model fit
In terms of the model fit, the pooled model has the highest R^2 (0.157), followed by the OD+time fixed-effects model (0.023) and the OD fixed-effects model (0.022). The within-R^2 of the pooled model is negative (-0.104), while the fixed-effects specifications within R^2 values are positive but very low (0.022 and 0.021), indicating that all three specifications explain very little of the within-OD variation in ticket sales.

5.3. Comparison of results

This section shows a comparison of the model results by visualizing the coefficient, the 95% confidence interval and whether it is significant (at $p < 0.05$). In the previous sections the results where compared for different model specifications for the same dataset. In this comparison the models are compared differently by comparing the results of the different datasets for the same model.

Figure 5.1 shows the results for the different datasets for the pooled OLS models. What is clear from the figure is that the confidence interval for the ticket sales coefficients is always smaller than for the volume datasets, and more of the variables are significant. The extra observations in the daily vs monthly volume datasets also clearly help significance, as origin GDP and air fare are significant in the daily model but not in the monthly model, while the magnitude is similar. An obvious outlier is rail frequency, as with a coefficient of 8.6 it lies outside of the figure. Another interesting discrepancy is rail travel time, as it is significant with a confidence interval fully in the negative part of the figure, while for both volume datasets the confidence interval is very large spanning both the negative and positive parts of the figure.

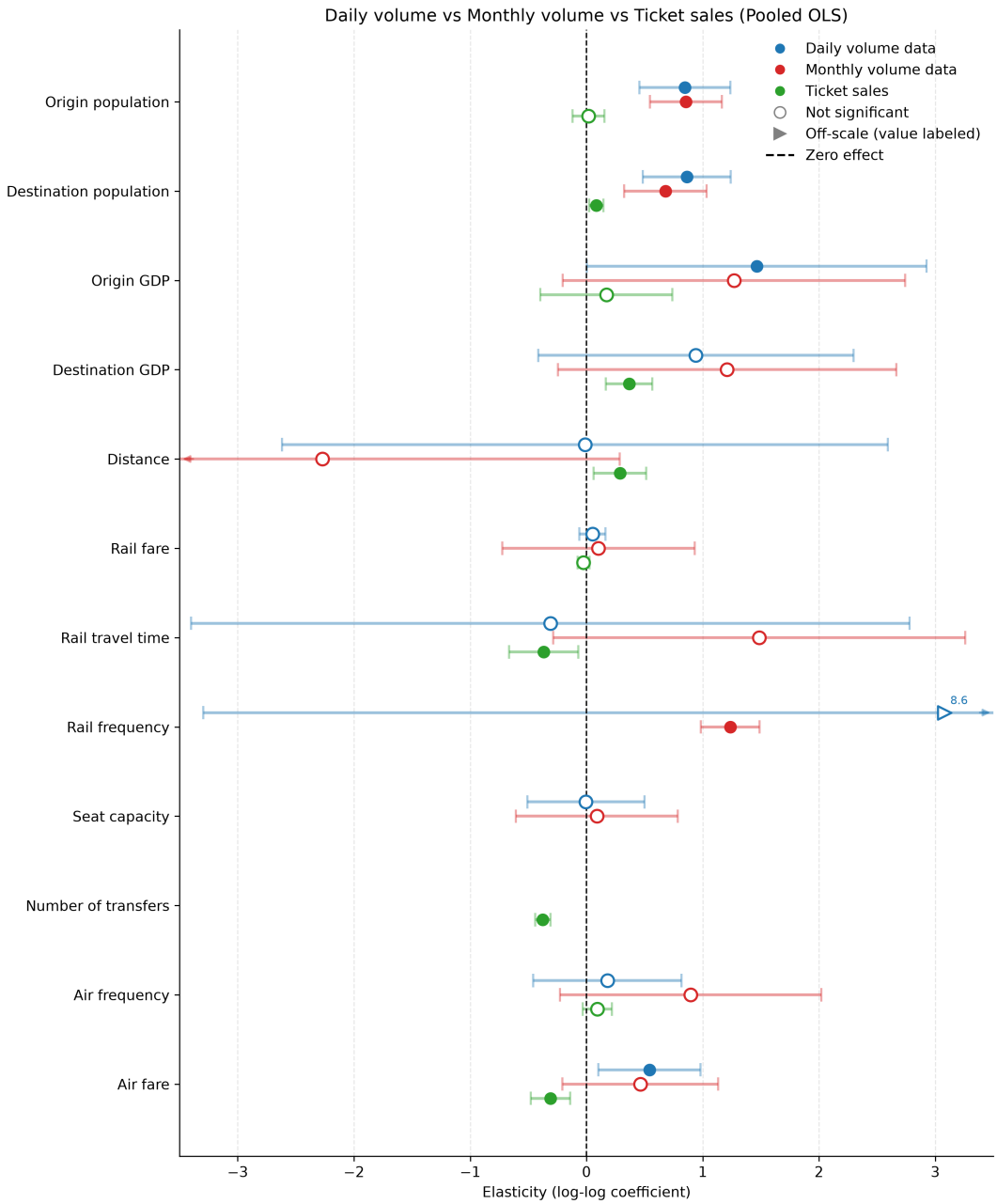


Figure 5.1: Pooled OLS model coefficients, with 95% confidence interval

Figure 5.2 shows the coefficients of the models for the OD fixed-effects specification. Some interesting observations can be seen again. For the origin and destination population coefficients, the magnitude of the coefficients is much larger for the monthly data is much larger than for the daily data, and even have the opposite sign for the destination population, although they are also insignificant. They also have a large confidence interval which is not fully captured in this figure. Rail fare went from insignificant for all datasets in the pooled model to significant and negative in the monthly volume and ticket sales data. The difference in significance for monthly vs daily data for the rail fare is likely resulting from the addition of Eurostar data, as the monthly data has less observations but includes more OD pairs. Air fare also became significant with fixed-effects and it is the only variable for which all datasets agree.

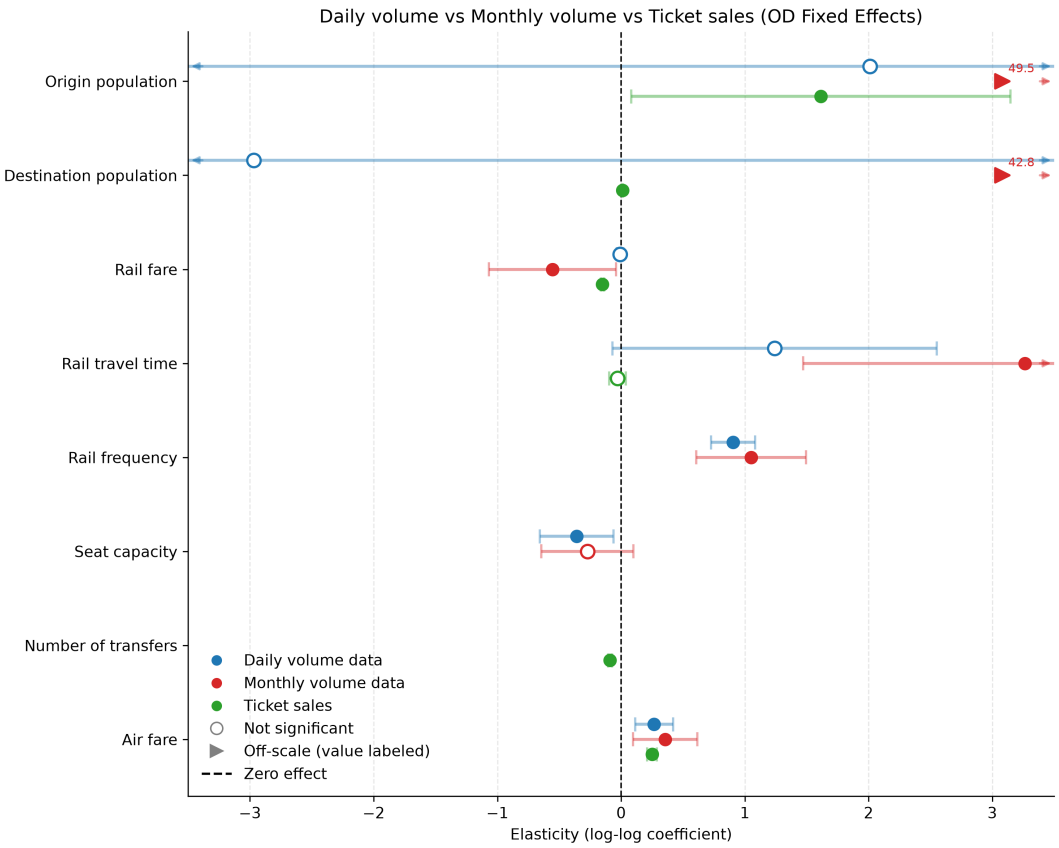


Figure 5.2: One way fixed-effects model coefficients, with 95% confidence interval

Figure 5.3 shows the coefficients of the model for the two way fixed-effects models. This shows that when introducing time fixed-effects the confidence intervals for origin and destination population for both volume datasets increases even more. Other than that the changes compared to one way fixed-effects are relatively minimal. Rail frequency is clearly the most stable variable across all specifications and datasets, especially for the monthly data.

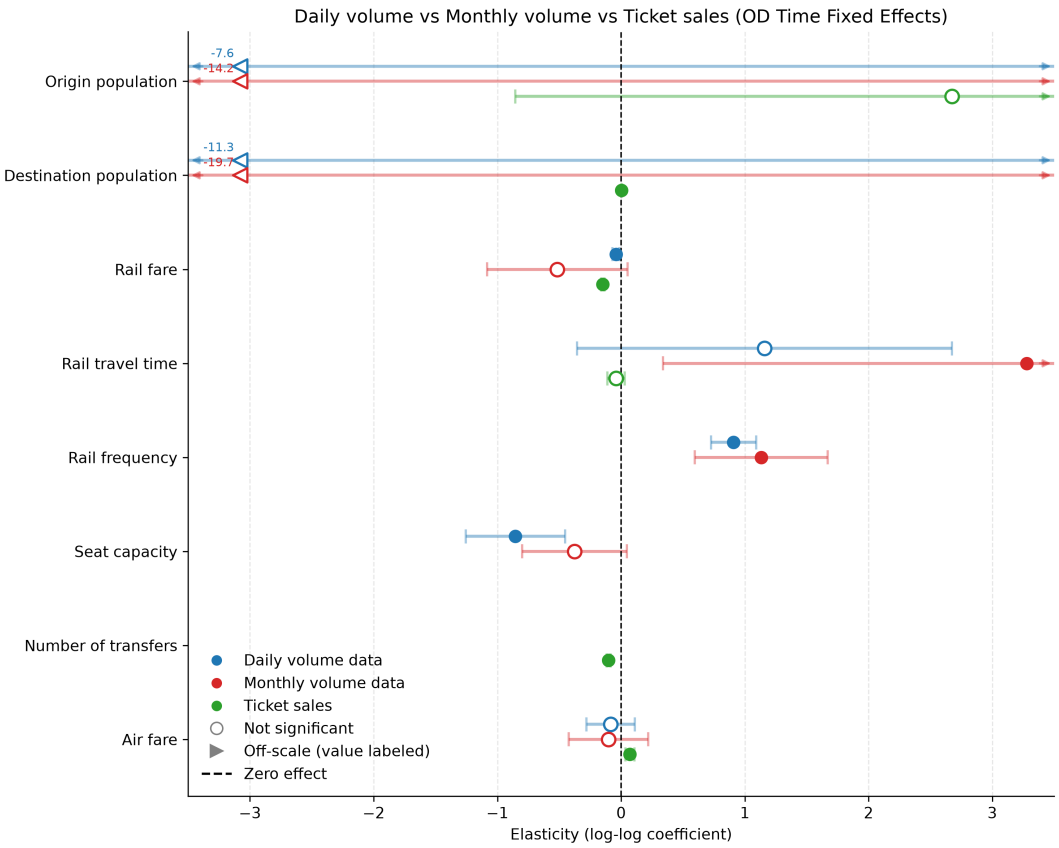


Figure 5.3: Two way fixed-effects model coefficients, with 95% confidence interval

6

Case study: corridor forecasting

This chapter shows the application of the model by forecasting rail demand for a number of corridors under varying scenarios. It follows the methodology introduced in section 3.8.

6.1. Forecasting model

For the forecast either the passenger volume pooled OLS model or the ticket sales pooled OLS model can be used as the fixed-effects models can not be applied for unobserved corridors. The ticket sales model provides the most plausible results, as the estimated coefficients are more plausible than for the passenger volume models. chapter 7 provides more detail on why these results are considered to be better.

The model used for the forecasting is shown in Equation 6.1, where V_{ijt} is the estimated demand on OD pair ij on average day t, α is the intercept of the model, $X_{k,ijt}$ represent the characteristics of the OD pair, $\hat{\beta}_k$ is the estimated coefficient of the explanatory variable, $\hat{\phi}$ is the duan smearing factor and δ is the ticket sales correction factor.

$$\hat{V}_{ijt} = \frac{\exp\left(\hat{\alpha} + \sum_k \hat{\beta}_k \ln(X_{k,ijt})\right) \cdot \hat{\phi}}{\delta} \tag{6.1}$$

The duan smearing indicator was estimated to be 2.25, and the ticket sales correction factor estimated in subsection 4.5.2 is 0.25.

6.2. Candidate corridors

From a set of corridors three main corridors were selected. They can be split up into one high speed rail corridor with two branches and two conventional intercity rail corridors. The corridors for this case study are shown below with the served stations and the travel times for the base scenario. All corridors are assumed to have a frequency of 1 train per day and run during the daytime. The estimated demand will not vary with frequency as it is not included as a variable in the ticket sales model. The three corridors and the corresponding baseline fare and travel time are:

- Corridor 1A: HSR Amsterdam - Barcelona
- Corridor 1B: HSR Amsterdam - Milan
- Corridor 2: Amsterdam - Copenhagen
- Corridor 3: Amsterdam - Prague

These corridors were selected from a number of candidate corridors proposed by NSI, and adjusted to fit the requirements listed in section 3.8. This meant that some of the train station were excluded because there were multiple stations in the same FUA or because they were located outside of any FUA. The feature construction also meant some adjustments had to be made because of a lack of air travel alternatives or GDP data, this is described in subsection 6.2.1. The travel time, fare and excluded stations are all shown in tables 6.1 to 6.4.

Train station	Travel time	Total travel time	Fare from first origin
Amsterdam Centraal	-	00:00	-
Schiphol Airport			*Excluded because of same FUA*
Rotterdam Centraal	0:50	0:50	€49
Antwerpen Centraal	0:35	1:25	€49
Bruxelles Midi	0:36	2:01	€49
Lille Europe	0:47	2:48	€49
Aéroport Charles de Gaulle TGV			*Excluded because of same FUA*
Marne-la-Vallée Chessy	1:16	4:04	€59
Lyon Part-Dieu	1:23	5:27	€69
Valence TGV	0:36	6:03	€79
Nîmes	0:48	6:51	€79
Montpellier-Saint-Roch	0:29	7:20	€79
Narbonne			*Excluded because located outside of any FUA*
Perpignan	1:30	8:50	€79
Figueres-Vilafant			*Excluded because located outside of any FUA*
Girona	0:43	9:33	€89
Barcelona Sants	0:42	10:15	€99

Table 6.1: Corridor 1A: Amsterdam - Barcelona

Train station	Travel time	Total travel time	Fare from first origin
Same as Corridor 1A up to Lyon Part-Dieu	-	-	-
Lyon Part-Dieu	-	5:27	€69
Chambéry-Challes-les-Eaux	1:36	7:03	€79
St Jean de Maurienne-Arvan, Modane, Oulx			*Excluded because located outside of any FUA*
Torino Porta Susa	2:47	9:50	€89
Milano Centrale	0:55	10:45	€99

Table 6.2: Corridor 1B: Amsterdam - Milan

Train station	Travel time	Total travel time	Fare from first origin
Amsterdam Centraal	-	00:00	-
Amersfoort Centraal	0:33	0:33	€29
Deventer	0:46	1:19	€29
Osnabrück Hbf	1:30	2:49	€29
Bremen Hbf	1:02	3:51	€39
Hamburg Hbf	1:02	4:53	€39
Kolding St			*Excluded because located outside of any FUA*
Odense	3:34	8:27	€69
København	1:11	9:38	€79

Table 6.3: Corridor 2: Amsterdam - Copenhagen

Train station	Travel time	Total travel time	Fare from first origin
Amsterdam Centraal	-	00:00	-
Amersfoort Centraal	0:36	0:36	€29
Deventer	0:42	1:18	€29
Osnabrück Hbf	1:36	2:54	€29
Hannover Hbf	1:10	4:04	€39
Berlin Hbf	1:38	5:42	€45
Berlin Südkreuz			*Excluded because of same FUA*
Dresden Hbf	1:42	7:24	€69
Děčín hl.n.			*Excluded because located outside of any FUA*
Praha hl.n.	2:15	9:39	€79

Table 6.4: Corridor 3: Amsterdam - Prague

6.2.1. Feature construction

For every origin and destination in the corridors the explanatory variables needed to be constructed to estimate the demand. This is using the same methodology as for the model estimation, with two exceptions explained below.

In the original estimation, any OD pair without air travel alternative was removed for the model estimation, which allowed the coefficients for the air travel characteristics to be estimated. For these corridors however, numerous OD combinations did not have any available air travel alternative. To mitigate this, the pooled ticket model was estimated again without the air travel explanatory variables. This was deemed acceptable as the impact of the air travel characteristics on estimated demand turned out to be very small (see Table 5.3). The difference in estimated demand with the adjusted model was also very minimal (usually less than 0.5 passengers per OD pair) and therefore it was chosen to use the model without air travel characteristics for OD pairs without an air travel alternative and the normal model for all other OD pairs. The results from the adjusted model are shown in Appendix E.

The dataset used for GDP data in the model estimation ("OECD Data Explorer • Economy - FUAs", n.d.) only included GDP per capita for every a functional urban area if that FUA had at least 250000 inhabitants. Stations in smaller FUAs were excluded from the estimation which was not considered a large issue since it was a relatively small proportion of the data (see subsection 4.5.2). GDP per capita turned out to be an important explanatory variable for demand and therefore the data quality of this variable is important for the forecasting. Four of the stations included in the corridors are located in FUAs smaller than 250000 inhabitants and therefore no GDP data was available. To mitigate this a different dataset from the same source was used which includes the GDP per different type of city or urban area for every country (OECD, n.d.-a). As the number of stations was very limited this was feasible to do manually.

The tables 6.1 to 6.4 only mention the travel times between each consecutive station. For all other OD combinations they are interpolated. Only the fare from the corridor's first station to each other station was available in the source tables. To fill in the missing fares, a fare vs travel time curve is built from the origin (time, fare) observations. For any OD pair, the segment travel time is obtained by subtracting cumulative travel times, and this segment time is then mapped to a fare via linear interpolation on that curve. A minimum fare is applied to every interpolated value to avoid implausibly low fares near the origin. The minimum fare is the lowest fare from the source tables.

6.2.2. Scenarios

For every corridor, alternative scenarios(s) is/are analysed to evaluate how the estimated demand responds to a change in fare, travel time or transfers. Capacity and frequency are not varied in any scenario, as these variables were not included in the ticket sales model used for the forecasts.

For corridors 1A and 3, the fare is doubled relative to the base scenario to evaluate the resulting change in estimated demand. In addition, for corridor 1A a scenario is estimated where there is a transfer in Paris. The

travel time is kept the same to isolate the effect of introducing a transfer. Only OD pairs whose route passes through Paris as an intermediate stop (so the origin and destination lie on opposite sides of Paris), get a value of 1 for transfer while all other OD pairs remain without a transfer. For corridors 1B and 2, the scenarios reflect the opening of new infrastructure and the resulting reduction in travel time. For corridor 1B, this is the opening of the Lyon-Turin base tunnel and its associated high-speed line, which reduces travel time by 2 hours for all stops beyond the tunnel ("Turin-Lyon Base Tunnel", n.d.). For corridor 2, this is the opening of the Fehmarnbelttunnel, which reduces travel time between Hamburg and Copenhagen by 2 hours and removes Kolding and Odense as stops on the corridor ("Why we're building the Fehmarnbelt fixed link", n.d.).

Corridor	Change	Difference compared to base scenario
Corridor 1A (HSR Barcelona)	Fare doubling	Example fare €198 vs €99 for Amsterdam Centraal to Barcelona Sants
Corridor 1A (HSR Barcelona)	Transfer vs direct service	Change in Paris while travel times remain the same
Corridor 1B (HSR Milan)	Opening Lyon-Turin base tunnel	Reduction in travel time of 2 hours from Torino and Chambéry station omitted
Corridor 2 (Copenhagen)	Opening Fehmarnbelttunnel	Odense and Kolding omitted and reduction of travel time of 2 hours to Copenhagen
Corridor 3 (Prague)	Fare doubling	Example fare €158 vs €79 Amsterdam Centraal to Praha hl.n.

Table 6.5: Alternative scenarios for corridors

6.3. Forecast results

The section present the results of the forecast, first an overview of the corridors and corresponding loads is shown. Afterwards per corridor the OD matrix under baseline and alternative scenario(s) is shown as well as the load profile in both directions. This section only presents the results, and in section 6.4 the results are interpreted. Discussion about the forecasting methodology is done in chapter 7.

Table 6.6 summarizes the demand for the corridors. It shows the total average demand for one day (total of both directions), the number of stops in the corridor, the average demand per OD pair and the demand between the endpoints. Figure 6.1 shows the corridors and the corresponding loads on each section for the baseline scenarios in the forward direction. The forward direction is considered away from Amsterdam. The thickness indicates the load on each section of the corridors. From the figure it is clear that the middle section of the corridor always is the busiest section.

Corridor	Total demand (total of both directions)	Stops	Average demand per OD pair	Demand Amsterdam to endpoint
Corridor 1A (HSR Barcelona)	*Removed for confidentiality*	14		
Corridor 1B (HSR Milan)	*Removed for confidentiality*	10		
Corridor 2 (Copenhagen)	*Removed for confidentiality*	8		
Corridor 3 (Prague)	*Removed for confidentiality*	8		

Table 6.6: Estimated demand for corridors, all demand forecast measurements are in estimated passengers/day

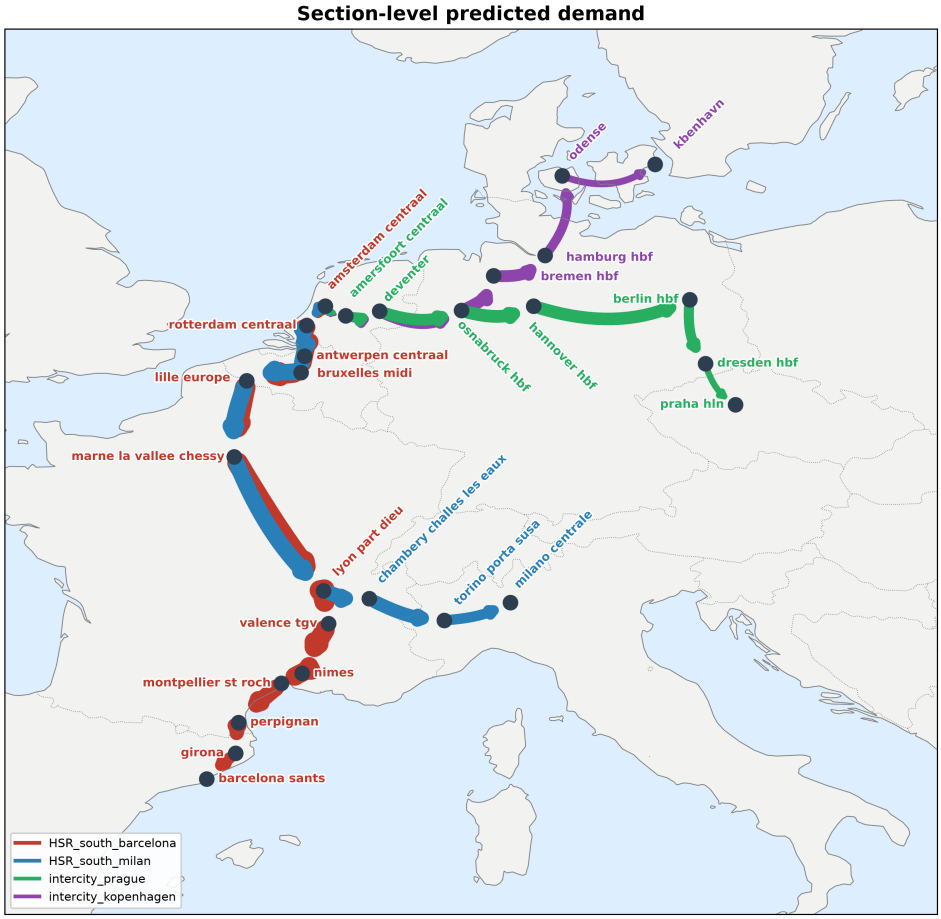


Figure 6.1: Corridors and estimated demand loads. Line thickness shows load on section

6.3.1. Corridor 1A: Barcelona HSR

Removed for confidentiality

Figure 6.2: OD matrix of estimated demand of corridor 1A, values in passengers/day

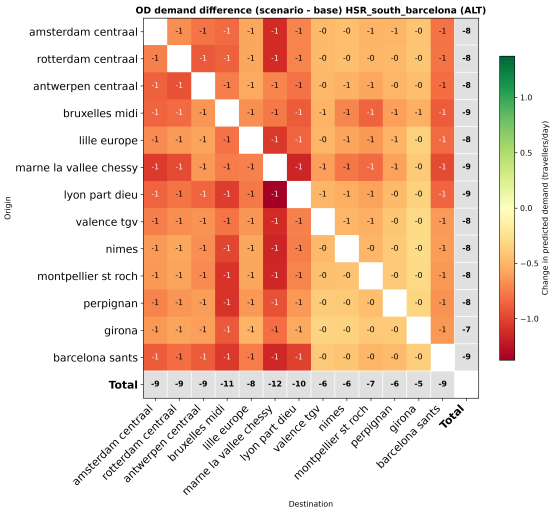


Figure 6.3: Difference in OD matrix under alternative scenario with fare doubling for corridor 1A

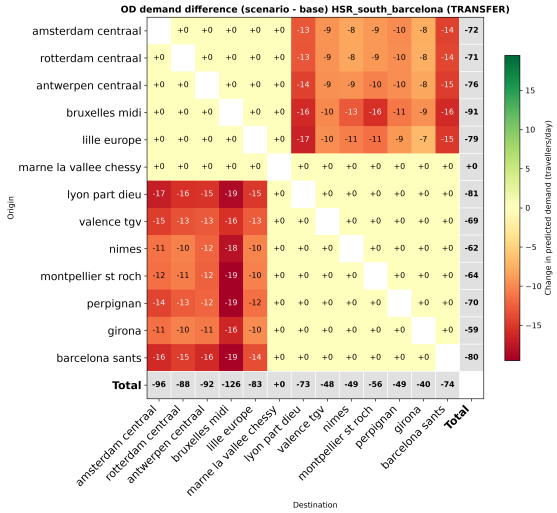


Figure 6.4: Difference in OD matrix under alternative scenario with transfer in Paris for corridor 1A

Removed for confidentiality

Figure 6.5: Corridor load per section for corridor 1A under baseline scenario, values in passengers/day

Figure 6.2 shows the OD matrix for corridor 1A under the baseline scenario, while Figure 6.3 shows the difference in demand under fare doubling and Figure 6.4 shows the difference in demand with a transfer in Paris. Table 6.7 shows the demand measurements for all three scenarios. Under the baseline scenario the largest demand occurs between Marne-la-Vallée Chessy (located within the Paris FUA) and Lyon Part-Dieu. All stations up to Lyon generate a considerable amount of demand, while the stops between Lyon and Barcelona produce relatively less demand. Figure 6.5 shows the load on every section for both directions. This shows that the backward direction (from Barcelona towards Amsterdam) generates more demand than the forward direction, with the largest asymmetry occurring at Paris, where the corridor's diverging pattern of stations switches to a converging one.

Under the alternative scenario with fare doubling, the difference in demand is just 0 or 1 passenger per OD pair per day. However, under the alternative scenario with a transfer the impact is much larger. All OD pairs where a transfer is involved show a significance reduction in the number of passengers per day, for example, the estimated number of passengers from Amsterdam Centraal to Barcelona Sants reduces from 46 to 32 which is reduction of roughly 30%. The largest decrease is on the OD pair Montpellier-Saint-Roch to Bruxelles-Midi, which is the OD pair with the largest demand that requires a transfer.

Scenario	Total demand (total of both directions)	Stops	Average demand per OD pair	Demand Amsterdam to endpoint
Baseline	*Removed for confidentiality*			
Fare doubling	*Removed for confidentiality*			
Transfer in Paris	*Removed for confidentiality*			

Table 6.7: Estimated demand for corridor 1A under alternative conditions compared to baseline, all measurements are in passengers/day

6.3.2. Corridor 1B: Milan HSR

Removed for confidentiality

Figure 6.6: OD matrix of estimated demand of corridor 1B, values in passengers/day

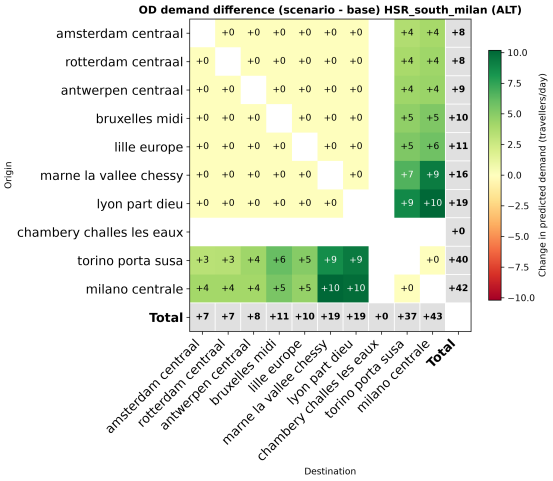


Figure 6.7: Difference in OD matrix under alternative scenario for corridor 1B

Removed for confidentiality

Figure 6.8: Corridor load per section for corridor 1B under baseline scenario, values in passengers/day

Figure 6.6 shows the OD matrix for corridor 1B under the baseline scenario, and Figure 6.7 shows the difference in demand under the alternative scenario where the Lyon-Turin base tunnel is finished together with the corresponding high speed line. This reduces the travel time for all stations after Lyon with 2 hours and omits the stop at Chambéry station as the line would be in a tunnel at that point. The estimated demand between al OD pairs before Lyon is the same as for corridor 1A as they serve the same stops with all the same characteristics. Turin and Milan also both generate a considerable amount of demand from all other origins. Under the alternative scenario the model predicts substantial amount of extra demand for Turin and Milan. The increase in demand is the largest for OD pairs where the reduction in travel time is relatively the largest.

6.3.3. Corridor 2: Kopenhagen

Removed for confidentiality

Figure 6.9: OD matrix of estimated demand of corridor 2, values in passengers/day

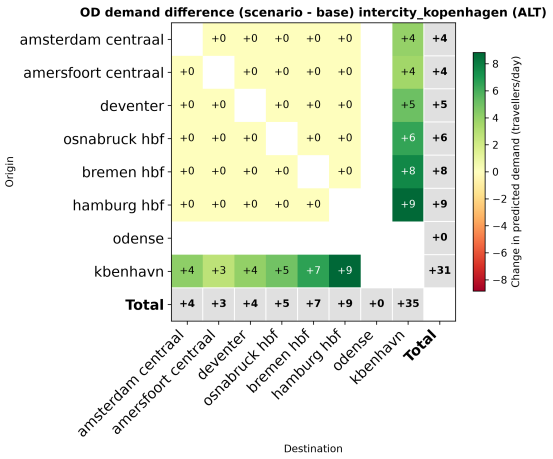


Figure 6.10: Difference in OD matrix under alternative scenario for corridor 2

Removed for confidentiality

Figure 6.11: Corridor load per section for corridor 2 under baseline scenario, values in passengers/day

Figure 6.9 shows the OD matrix for corridor 2 under the baseline scenario, and Figure 6.10 shows the difference in demand under the alternative scenario where the Fehmarnbelttunnel is finished. Again it is clear that the largest cities generate the most demand with Amsterdam, Hamburg and Copenhagen dominating the generated demand. The total generated demand is a less compared to corridors 1A and 1B. Under the alternative scenario only demand for Copenhagen is influenced as this is the only station that benefits from the reduced travel time. Again Figure 6.11 shows that there is some asymmetry in the demand with the forward direction showing more demand for most of the corridor.

6.3.4. Corridor 3: Prague

Removed for confidentiality

Figure 6.12: OD matrix of estimated demand of corridor 3, values in passengers/day

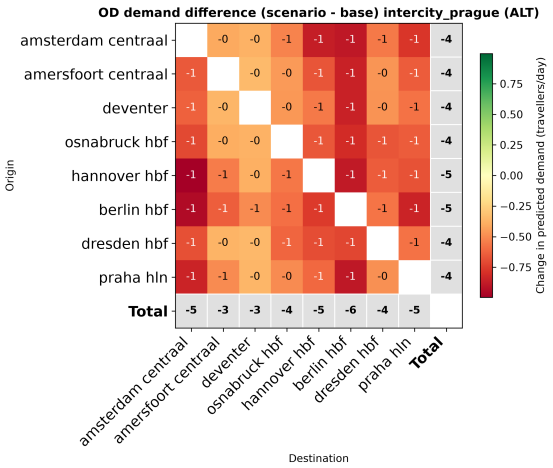


Figure 6.13: Difference in OD matrix under alternative scenario for corridor 3

Removed for confidentiality

Figure 6.14: Corridor load per section for corridor 3 under baseline scenario, values in passengers/day

Figure 6.12 shows the OD matrix for corridor 3 under the baseline scenario, and Figure 6.13 shows the difference in demand under the alternative scenario where all ticket fares are doubled. Again it is clear that the largest cities generate the most demand with Amsterdam and Berlin being the busiest OD pair. Under the alternative scenario demand is only slightly influenced with OD pairs showing a maximum change of 1 travellers per day. Again Figure 6.14 shows that there is some asymmetry in the demand with the forward direction showing more demand for most of the corridor.

6.4. Forecasting uncertainty and interpretation

The results from the forecast might give the impression to be accurate and useful when looking at the numbers but is is important to read them as indicative of how this model predicts demand for different corridors with different characteristics instead of precise estimates. This section aims to explain how the forecast results are formed and what the drivers of uncertainty are in the forecast.

The model used for the prediction the pooled OLS model based on the ticket sales data, for which the results are summarized in Table 5.3. This shows that the model is mostly driven by GDP, distance, travel time, transfers and air fare, with population, rail fare and air frequency having a much smaller influence. As a larger distance will result in a larger travel time so both variables increase for a further destination from an origin, they partially cancel their effects although the relation is not constant as different train services can have different speeds. In addition, the coefficient for destination population and destination GDP is larger than for the origin coefficients which means that the model will predict asymmetric demand on the same OD pair when the origin and destination are not the same size.

When looking at the estimated demand for any of the corridors this is result that can be seen. OD pairs with two different sizes produce asymmetric demand and demand is mostly driven by the OD characteristics. Demand is

not strongly reduced even by very long travel times (10+ hours), since in this model distance actually increases the estimated demand while travel time reduces demand at the same time. With only marginal effects from rail fare and air travel characteristics, the model always estimates demand mostly based on the population and GDP per capita of the FUA that the train station is located in. This means that the magnitude of the estimated demand is similar for similar sized OD pairs partially regardless of the rail and air travel characteristics. One exception is the influence of transfers, as including a transfer in corridor 1A in Paris reduced the demand for all OD pairs with a transfer by around 30%. This clearly indicates that there is value in running a direct train service, even if travel times are the same.

The corridors are used to predict demand for new train services, but demand is predicted separately for each OD pair. This means that adding more stops will always increase demand according to the model, to mitigate this slightly, only one stop per FUA was implemented and more stops does result in longer travel times, but this does not reduce demand significantly. Therefore a corridor with more stops will result in more total demand, and this can be seen when comparing corridor 1A to corridor 2. The total demand is much higher for corridor 1A compared to corridor 2 even when the travel times and city sizes are similar.

Since the pooled OLS model is estimated based on ticket sales data, where every observation is a number of tickets sold between OD pairs with the origin in The Netherlands and the destination outside of The Netherlands, the model actually predicts how many tickets would be sold by NSI for this OD pair instead of total demand. To get magnitudes for the expected volume instead of ticket sold by NSI the ticket sales correction factor was applied based on the share of tickets for OD pairs with a distance >300km for the lines for which passenger volume was also available. This introduces extra uncertainty into the prediction, as this share is not representative for overall share, and is expected to be even lower for other OD pairs not served by train services directly from The Netherlands. Other than the ticket sales correction factor the reported demand levels also incorporate the Duan smearing factor described section 3.8, which is an estimate derived from the residuals of the data on which the model is estimated. This correction was not evaluated based on the candidate corridors, since there is no observed demand for these corridors. Therefore this factor introduces another level of uncertainty into the prediction.

The use pooled OLS specification introduces further uncertainty. The fitted equation retains variables that are not statistically significant, which was done to be able to predict demand level. This mean that their actual effect could be smaller, zero, or reversed in sign. The specification also has a relatively low goodness of fit (Table 5.3), meaning a substantial share of demand variation remains unexplained regardless of the characteristics of the corridors. Finally, because pooled OLS omits OD-pair fixed-effects, it does not control for time-invariant unobserved heterogeneity. This is likely the case, as the coefficients for some of the variables flip sign in the fixed-effects specifications. Pooled OLS is used because it is the only specification capable of predicting absolute demand for new corridors while using time-invariant characteristics (such as distance), at the cost of a weaker explanatory power than the fixed-effects models.

Another source of uncertainty lies in the data limitations. First, several intermediate stations were omitted because either they lie in the same FUA as another station or they are located outside of any FUA. This means that the model can not properly predict demand for these stations. Secondly, all corridors have close to the maximum travel times observed in the data (see Figure 4.13b), where there are very few observations for these long travel times. This means that prediction for these OD pairs requires some extrapolation and thus are less accurate.

6.5. Corridor ranking

Based on the baseline predicted demand for the corridors, the four candidate corridors are compared. This is based on the demand overview from Table 6.6 and the estimated demand per OD pair. Corridor 1A has the highest total predicted demand, but this is partly due to the larger number of stations, and as described in the previous section, more stations automatically results in more total demand. The average demand and endpoint to endpoint demand are less sensitive to these effects since they are normalised for the corridor length and stops. Based on these metrics corridor 1B scores the best considering average demand per OD pair, while corridor 1A shows the highest total demand.

Using these normalized metrics, the ranking becomes: corridor 1B, 1A, 3 and 2. However, this ranking only reflects the demand estimated by this model and is subject to many limitations and should therefore be interpreted carefully. This ranking does not incorporate the demand effects of the infrastructure improvements,

fare changes or transfer from the scenario analysis.

7

Discussion

This chapter interprets and discusses the results from this research and places it into a more general context. It uses the following structure, first the main findings of this research are interpreted, discussed and put into the broader context in section 7.1. Next the case study is discussed, by considering the developed methodology and results in section 7.2. Section 7.3 discusses the main limitations of this research and how they can be addressed in future research. Finally section 7.4 translates the findings into some practical recommendations.

7.1. Results discussion

This section discusses the main results of the research (without the case study). First, the mechanisms behind the results, nature of the data and plausibility of the results are interpreted. Next the choices made in the model specification are discussed and how these influence the results. Finally the results are compared to the literature.

7.1.1. Interpretation of the individual coefficients

The results are interpreted per variable, and discussed across model specifications and datasets. Any important differences and anomalies are reported. Next, differences at the model level are discussed, and finally which model results are considered the best is discussed.

Service type dummies

The service type dummies are only part of the passenger volume model specifications as the service types in the ticket sales data are not known (results from Table 5.1 and 5.2). The ICE Berlin is taken as a baseline and all the other service dummies are compared to this service. In the daily results only ICE Frankfurt is significant, meaning that the ICE Frankfurt has somewhat higher demand than ICE Berlin on average. An explanation for this could be that Frankfurt has more available onward connections thus making it more attractive, which is unobserved in the model. The other service dummies are insignificant, likely because they have much fewer observations and/or more heterogeneity. The ICE Basel and Munich have very large point estimates, but as the sample size is relatively small these should not be over interpreted. The monthly results show that the seasonal Eurostar (Snow and Sun) generally has more demand, as despite having only a few observations, demand for these seasonal services is mostly at capacity, giving a precise estimated effect despite the small sample.

OD characteristics

Population appears to be an important driver of demand in all pooled models, being significant for both origin and destination in the volume data models, and for the destination only in the ticket sales model. The coefficients from the volume models become quite large in both fixed-effects models, especially estimated on the monthly data. This is expected as they try to explain variation in demand over time within the same OD-pair, with very limited variation in population over time which results in large coefficients. The larger population coefficient in the monthly fixed-effects model likely reflects the larger OD pair composition of the monthly panel, which includes Eurostar corridors not present in the daily panel. From the ticket results, in all three model specifications, destination population is significant while origin population is not (except for in the OD fixed-effects model). A possible explanation is the way the ticket dataset is constructed and ticket are sold. NSI encourages to book from your 'home' station, which results in a large set of origins (nearly all train stations within The Netherlands). With a large number of origins, the number of observations per origin is relatively small. A large share of the tickets sold are on corridors which are served by a direct train from

the Netherlands. This results in many observations in the ticket sales dataset being formatted like this: home station - large station in the Netherlands - large station outside of The Netherlands served by a direct service. Therefore there exist many observations for the same destination compared to the origin and therefore this can explain why the destination population is significant while the origin is not. The same pattern occurs for GDP per capita, where the destination coefficient is significant but the origin is not. GDP per capita is only weakly significant in the daily volume origin GDP but not for any other variable from the ticket sales data. This is likely because the volume data only considers a few corridors in western Europe which all have a somewhat similar GDP per capita and therefore there is not enough variation to reach significance.

Distance presents a bit of an anomaly. In both the daily and monthly volume models, the distance coefficient is negative and insignificant, which is the theoretically expected sign as in hypothesis H3, where demand decreases with distance. In the ticket sales pooled model however, the coefficient is positive and significant, suggesting that within this sample, OD pairs with greater distance are associated with higher demand. There are a few possible explanations for this positive distance coefficient. First, distance is correlated with rail travel time, although imperfectly as different services vary in speed (HSR vs slower regional connections). This means that the two variables cannot be interpreted fully independently, part of the distance effect may reflect travel time variation not fully captured by the travel time variable itself. Another more speculative explanation is that longer-distance destinations may carry additional unobserved attractiveness such as tourism appeal, weather, or other factors that are not captured by the model and correlated with distance.

Rail service attributes

Rail fare elasticity is among the most researched variables in the literature, making hypothesis H4 one of the strongest hypotheses in this research. The effect of moving from pooled OLS to a fixed-effects specification is particularly clear for this variable. In the daily volume model, the fare coefficient flips sign and becomes significant when moving from the pooled to the two-way fixed-effects specification. The monthly volume model shows a similar sign flip, though the coefficient remains insignificant. In the ticket sales model, the coefficient becomes more negative and remains significant across specifications. This pattern is consistent with the endogeneity problems explained in the methodology. Fares may be correlated with unobserved OD characteristics. For example, higher fares on high-speed services or higher quality services that independently attract more demand, or dynamic pricing systems that respond to demand itself. In pooled OLS, this correlation can bias the fare coefficient toward the observed positive/insignificant pattern. Once OD (and time) fixed-effects absorb this unobserved heterogeneity, the remaining variation is closer to within OD fare changes over time, resulting in the theoretically expected negative relationship between fare and demand.

Rail travel time shows a plausible negative coefficient in the daily volume pooled model. In the monthly volume pooled model and across the fixed-effects specifications for both volume datasets, however, travel time yields unexpected (positive) or insignificant coefficients. As with population, this can be attributed to the fixed-effects models relying only on within-OD variation, and travel time showing very limited within-OD variation over the sample period (see subsection 4.4.1), this both inflates the magnitude of the estimated coefficient and reduces its precision. In the ticket sales results, the coefficient remains negative (as expected) across specifications, though it does not reach significance in the fixed-effects models, again consistent with limited within-OD variation in travel time.

Rail frequency is only part of the volume data models and clearly shows that a higher service frequency is strongly associated with demand, both in the pooled model specification and both fixed-effects models.

Seat capacity is something of an anomaly. This variable was included because it was available in the passenger volume data and exhibits some within-OD variation over time. Hypothesis H7 formulated a positive sign, as greater capacity was expected to be associated with higher demand. However, the realised sign is negative across specifications. Two considerations help explain this result, and can explain why the model suggests that seat capacity reduces demand. First, in the fixed-effects models, capacity per train is nearly time-invariant for most OD pairs, with only a single rolling stock change observed across the sample period (see Table 4.3). As with population and travel time above, this leaves very limited within-OD variation to identify the coefficient, so the fixed-effects estimates for capacity should not be over-interpreted. Secondly, relevant for the pooled model where between-OD variation exists, capacity is moderately positively correlated with service frequency (see Figure 4.10). Frequency is strongly significant and positive across specifications, consistent with theoretical expectations that more frequent service increases demand through greater schedule convenience. With this correlation, it is possible that frequency captures most of the effects from more service, more demand. This

could leave capacity's coefficient to reflect a residual effects shaped by the shared variation rather than an independent effect. While considering this, this explanation could mean the negative coefficient is more of an effect of multicollinearity than a finding that more seat capacity would reduce demand.

The number of transfers, included only in the ticket sales model, produces interpretable and expected results. Unlike the other rail service variables discussed above, number of transfers is entered in levels rather than logs, so its coefficient is interpreted as a semi-elasticity: the percentage change in demand per extra transfer, rather than an elasticity with respect to a percentage change in transfers. More transfers are associated with lower demand, consistent with hypothesis H8. This effect holds across the model specifications. The fixed-effects coefficients are smaller in magnitude than the pooled estimate, but stay negative and significant. This pattern reflects the two sources of variation captured by number of transfers: between OD pairs, where less transfers attract more tickets sold, and within OD pairs, where travellers show some preference for the same route with fewer transfers when alternatives exist. That the within-OD (fixed-effects) effect is weaker than the between-OD (pooled) effect is expected, as most of the variation in transfers reflects fundamental differences in route directness between OD pairs as there are not always multiple viable routes available.

Air travel characteristics
Both air travel variables (frequency and fare) can be interpreted as cross-elasticities: how much rail demand changes is response to the change in the air travel variable.

Air frequency is time-invariant, as the underlying data provides only a single cross-sectional datapoint with no time dimension. It is therefore absorbed under the fixed-effects specifications (as with distance) and testable only in pooled OLS model, where it is not significant. Unfortunately it cannot be tested if there is a response of rail demand to changes in air supply.

Air fare illustrates the effect of the fixed-effects specifications. In the daily volume model, the coefficient is positive and significant under pooled OLS, remains positive and significant but smaller under OD fixed-effects, and turns negative and insignificant under OD+time fixed-effects. The initial positive, significant pooled coefficient is unexpected, implying complementarity rather than substitution, and likely reflects unobserved between-OD heterogeneity, as corridors with higher air fares may also have generally higher rail demand, unrelated to any substitution effect. Once fixed-effects absorb this heterogeneity, air fare loses its explanatory power for rail demand entirely in this dataset.

The monthly volume model shows that both the pooled and OD+time fixed-effects estimates are insignificant, with air fare reaching significance only under OD fixed-effects alone.

The ticket sales model produces a more interpretable result. Air fare is significant across all three specifications, with the coefficient moving from negative under pooled OLS to positive under OD fixed-effects, and remaining positive and smaller, under OD+time fixed-effects. As above, the negative pooled coefficient likely reflects between-OD correlation rather than an independent effects. The positive coefficients under both fixed-effects specifications, however, are consistent with rail-air substitution: within a given OD pair, a higher air fare is associated with modestly higher rail demand. Although the effect is small, its consistency and significance across both fixed-effects specifications makes shows the most clear evidence of cross-modal substitution found in this research.

7.1.2. Interpretation of overall models

Comparing the full set of results across models and datasets, the ticket sales results seem most plausible. They align most closely with the formulated hypotheses, and most variables are statistically significant. However, most coefficient magnitudes are smaller than commonly found in the literature and than anticipated in the hypotheses (see subsection 3.1.4). A likely explanation lies in the nature of the ticket sales data itself. As it contains only tickets sold through NSI, likely mostly representing Dutch travellers rather than the full market of travellers on a given OD pair. Multiple booking channels exist for international rail travel, and the share of travellers booking through NSI is unlikely to be consistant across OD pairs, likely declining the further a destination lies from the Netherlands, as competing national operators and booking platforms become more favourable alternatives. If this share itself responds to variable changes, the estimated elasticities could reflect changes in NSI booking share as much as changes in underlying rail demand, understating the true demand response.

The results from both volume datasets should also be interpreted carefully. Both datasets cover a limited number

of services, with substantial differences in demand levels between them. For example the Eurostar service to Paris carries much higher volumes than the ICE service to Munich. This combination of few entities and large scale heterogeneity across them makes the pooled estimates sensitive to a small number of high-volume services. Furthermore, the limited within-OD variation discussed throughout this section means the fixed-effects models have little variation to identify effects from, producing the large, imprecise, or unexpected signs for coefficients noted for several variables above. Together, these factors suggest the volume datasets results should not be considered very plausible, with the ticket sales results providing a better basis for interpretation.

7.1.3. Model specification

This research estimated three different specifications, namely pooled OLS, OD fixed-effects and OD+time fixed-effects (two-way fixed-effects). These multiple model types were used because some of variables are time-invariant, which meant they cannot be estimated in fixed-effects models. fixed-effects is expected to yield more reliable estimates because it absorbs any unobserved OD-level heterogeneity. These unobserved OD-specific characteristics are likely correlated with regressors such as fare and travel time. This is illustrated by the rail fare elasticity estimated on the daily volume data, as insignificant in the pooled OLS and OD fixed-effects models, but in the OD+time fixed-effects the sign is flipped and it is significant.

Pooled OLS remains necessary, as the time-invariant variables are absorbed under FE and they are important drivers of demand. In addition, the population, capacity, and travel time variables do vary over time but exhibit so little variation that the produced estimates are not reliable. Furthermore, in order to produce a forecast for unseen OD pairs in the case study FE cannot be used.

7.1.4. Comparison to literature

This subsection summarizes the findings across all models and datasets and compares them to the existing literature using the formulated hypotheses. Table 7.1 shows a summary of the most plausible elasticities found in this research. Here both significant and insignificant variables are used, as the most plausible estimate is reported.

The OD characteristics are broadly consistent with the literature and the hypotheses H1 and H2. Only the pooled OLS coefficients are considered in this paragraph, as described in subsection 7.1.1 the FE coefficients are not reliable. Multicollinearity between GDP per capita and population was observed for both datasets, with a high VIF, especially in the passenger volume dataset (subsection 4.5.3), which is an issue that was also identified by Wardman (2006).

For both volume datasets population and GDP per capita turned out to be important drivers of demand, which was also found in the literature (Wardman & Batley, 2022; Wardman et al., 2018; Wijeweera et al., 2014). Distance is sometimes used for heterogeneity analysis but not estimated directly as in this research. The positive coefficient found here does not align with hypothesis H3. The multicollinearity between distance and rail travel time and fare could be part of the cause of this as described in the previous section.

The rail service attributes from the ticket sales model are mostly consistent with the literature, while those from the volume data are not. Fare and travel time elasticities have been studied extensively, and both are hypothesized to be negative (H4, H5). The ticket sales fare elasticity matches this expectation but is somewhat smaller in magnitude than typical literature estimates, plausibly reflecting that ticket sales capture only NSI own booking share rather than total demand. Travel time is negative as expected, in the daily volume and ticket sales pooled model, the remaining specifications are affected by limited within-OD variation (subsection 7.1.1) and are not considered reliable for comparison. Frequency and capacity, estimated only in the volume models, are correlated and therefore not fully separable in interpretation; capacity in particular is not considered a reliable estimate. Frequency's magnitude is much larger than reported in the literature (Kristoffersson & Liu, 2024; Paulley et al., 2006) and than formulated in hypothesis H6.

The air travel characteristics results show that air frequency is never significant, while air fare does show some mode substitution effect, as was also formulated in hypothesis H10, although the magnitude is smaller than typically found in the literature (Hortelano et al., 2016; Román et al., 2014; Wardman et al., 2018).

The formulated hypotheses and the results from this study are summarized in Table 7.1. For every variable, the hypothesis, expected sign and magnitude and finding in this study are reported. The reported value is the finding that is considered the most plausible result from all the models and datasets rather than from a single model or dataset. This means that in some cases an insignificant result is reported rather than a significant

result (such as for distance). Since the results are reported for different models they should not be combined in a single forecast, as the coefficients also depend on the other regressors in their respective model and are affected by multicollinearity. The forecast should use the coefficients from one complete model.

Variable	Hypothesis	Expected sign and magnitude	Value found in this study	References
Origin-destination characteristics				
Origin Population	H1	+0.1 to 4.0	0.85** (V-OLS)	(Wardman, 2006; Wardman et al., 2007; Wijeweera et al., 2014)
Destination population	H1	+0.1 to 4.0	0.68** (V-OLS)	(Wardman, 2006; Wardman et al., 2007; Wijeweera et al., 2014)
Origin GDP per capita	H2	+0.3 to 1.0	0.17 (T-OLS)	(Börjesson, 2014; Paulley et al., 2006; Wardman, 2006; Wardman & Batley, 2022; Wardman et al., 2018)
Destination GDP per capita	H2	+0.3 to 1.0	0.37** (T-OLS)	(Börjesson, 2014; Paulley et al., 2006; Wardman, 2006; Wardman & Batley, 2022; Wardman et al., 2018)
Distance	H3	-0.5 to +0.5	-0.01 (V-OLS)	(Hortelano et al., 2016; Paulley et al., 2006)
Rail service supply attributes				
Rail fare	H4	-0.4 to -0.8	-0.148** (T-TWFE)	(Börjesson, 2014; Hortelano et al., 2016; Kristoffersson & Liu, 2024; Román et al., 2014; Wardman, 2006, 2022b; Wardman & Batley, 2022; Wijeweera et al., 2014)
Rail travel time	H5	-0.3 to -0.8	-0.37* (T-OLS)	(Börjesson, 2014; Paulley et al., 2006; Román et al., 2014; Wardman, 2006, 2022b)
Rail frequency	H6	+0.1 to 0.5	0.9** (V-TWFE)	(Kristoffersson & Liu, 2024; Paulley et al., 2006; Román et al., 2014; Wijeweera et al., 2014)
Seat capacity	H7	+ (no magnitude forecast)	-0.85** (V-TWFE), not recommended to use	*Variable not reported in reviewed literature*
Transfers	H8	- (no magnitude forecast)	-0.38** (T-OLS)	(Börjesson, 2014; Paulley et al., 2006; Wardman, 2006)
Air travel characteristics				
Air frequency	H9	-0.1 to -0.5	0.09 (T-OLS)	(Román et al., 2014)
Air fare	H10	+0.1 to 0.7	0.07** (T-TWFE)	(Hortelano et al., 2016; Román et al., 2014; Wardman et al., 2018)

Table 7.1: Overview of hypotheses and findings from this research, reported finding is most plausible finding from all results and reported as elasticities. V = passenger volume model, T = ticket sales model, OLS = Pooled OLS, FE = fixed-effects (OD only), TWFE = two-way fixed-effects (OD + time), *: $0.01 \leq p \leq 0.05$, **: $p \leq 0.01$

7.2. Case study discussion

This section discusses the methodology applied to the case study, the interpretation of the case study results was already covered in section 6.4.

The case study relies on the pooled OLS model, as the fixed-effects models cannot produce demand estimates

for OD pairs that are unobserved. Therefore the forecast is not based on the 'best' results from this research. This weakness is amplified by the requirement to include all coefficients, including the insignificant coefficients or coefficients with implausible signs (such as distance and air fare). This means that forecast is partially driven by relationships that are not deemed reliable.

A further part of the forecasting methodology is that it predicts demand per OD pair rather than for a full corridor at once. Since the model has no sense of multiple OD pairs belonging to a single train service. Although the travel time to further stops does increase when adding another stop, the effects from this is relatively small. A further effect of this is that when adding another stop, the demand always increases, this is partially mitigated by having a maximum of one stop per FUA but will always result in higher total estimated demand when adding more stops.

Another limitation of the forecasting methodology is that it has no sense of the availability of existing train services. Demand is predicted on an OD level without considering current services, which means demand is both under- and over-predicted. For example, the Amsterdam Centraal-Berlin Hbf pair (part of corridor 3, Amsterdam-Prague) is predicted to have around 51 passengers per day (see Figure 6.12), whereas the passenger volume dataset shows around 500 passengers per day (see **??**). Note that the prediction is based on ticket sales estimates, with both the Duan and ticket sales correction factors while the observed volume is a total amount of passengers for which no correction is necessary.

However, the gap is too large to be explained by this alone. A likely explanation is that the model is estimated on all OD pairs in the ticket sales data, which includes both high- and low-volume OD pairs. The estimated relations are therefore averages across all these pairs and do not capture factors specific to established connections, such as service frequency, passenger familiarity, brand recognition or tourism attractiveness. As a result, demand on OD pairs that already have well-established high quality train services established connections tend to be under-predicted.

The model can also over predict demand, for example for OD pairs connecting relatively small cities over very long travel times. Because the distance coefficient is positive and the travel time coefficient is negative, these two effects partially cancel each other out, reducing the demand penalty that would otherwise be expected for long travel times. Combined with the fact that for example tourism attractiveness is not included as a variable, this raises the possibility that the model over-predicts demand for such pairs, but this cannot be confirmed directly, as no observed demand data is available. An example is the Lille - Perpignan OD pair within corridor 1A (Amsterdam - Barcelona HSR), for which the model predicts 29 passengers per day, which is only half the predicted demand for Brussels - Barcelona, despite Lille and Perpignan being considerably smaller cities with a similarly long travel time and likely much less tourism attractiveness.

As mentioned before the absolute demand forecasts should not be seen as the actual expected level of demand, since it is based on the partial share of the total demand and relies on two correction factors (Duan and ticket sales) which amplify the uncertainty. The relative ranking is still considered useful as a first indication of which corridors are the most promising. Both correction factors scale the forecasts of all corridors in the same way and therefore affect the level of demand rather than the order of the corridors. In addition, all corridors considered in the case study originate in the Netherlands, which is also the origin of the ticket sales data, and are assumed to be subject to similar biases.

Finally, there are many more aspects of rail travel which are not covered in the model and are also difficult to capture. First, fares in this study are treated as inputs, whereas operators often use dynamic pricing that responds to anticipated demand, which is part of reverse causality. Secondly, the model has no way of representing how easily tickets can be purchased for a route. For rail travel this often a barrier, especially in long-distance travel. Finally when a new service is introduced it takes time for travellers to respond and possible demand to be induced. These factors are not part of the model.

7.3. Limitations and future research

This research is subject to a number of limitations which affect the interpretation of of the results. This section describes the limitations of the data, method and generalizability. The limitations of the case study are already covered in section 6.4 and 7.2.

Data

This research is heavily reliant on the quality of the data which means that any data problems directly impact the results. Subsection 4.5.5 already covers the technical limitations which influence the quality of the panel datasets, while this sections focusses on the data limitations which influence the interpretation of the results.

First, the passenger volume data in this research is limited to a small number of services. This biases the results toward how these specific services are used, but more importantly, the number of OD pairs is too limited to give a general picture of travel behaviour and thus representative elasticities. In addition, the volume per service is very unevenly distributed, which biases the estimates (see Table 4.5). Furthermore, without transfer data, it is unclear whether a passenger's actual journey started or ended at the station where they boarded or alighted. Since the data also covers only a limited number of specific services, even for known OD pairs it does not capture the full number of passengers travelling on that pair, as other services are often also available. Together, these limitations result in a fragmented picture of passenger volumes. This meant that the especially the within-OD variation was too limited to get interpretable results.

Secondly, the ticket sales data is subject to multiple sources of bias. First, since it only represents tickets sold by NSI, it is biased toward OD pairs reachable via direct services from the Netherlands, as most tickets are sold for these connections (see Figure 4.6). Second, it is biased by which travellers book through NSI, these are likely mostly Dutch travellers, who are generally familiar with NS and therefore more likely to know of NSI. This affects which destinations are booked, how much travellers are willing to spend, and which tickets are even offered through this booking channel. As a result, it is assumed that the further a destination lies from the Netherlands, the less likely a ticket for that journey was booked through NSI (see Figure 4.7). Even for OD pairs that are represented in the dataset, NSI's share of total ticket sales varies and is unknown, except for the OD pairs that overlap with the passenger volume data. Together, these limitations mean the ticket sales data provides a biased picture of demand for international rail travel, and the resulting elasticity estimates reflect NSI's booking behaviour instead of total rail demand.

Third, the additional data used in this research also shows limitations. GDP per capita data is only available up to an earlier year than the passenger volume and ticket sales data covered, so the latest available year was used throughout the sample period. This means the variable has no variation over time in the processed dataset, making it impossible to estimate in the fixed-effects models. Air travel frequency was also only available as an average over multiple years, and although population data did vary over time, the changes were too small to provide any useful variation. As a result, none of these variables offered meaningful temporal variation, resulting in to useful estimation in fixed-effects.

Future research would benefit a lot from a more complete demand dataset. This should cover a large, balanced set of OD pairs across multiple countries and multiple years, capturing full passenger volumes rather than tickets sold by a specific operator or a subset of train services. Having such a dataset would allow addressing several of the limitations discussed above: a wider set of OD pairs with more even volume distribution would reduce the sensitivity to a small number of dominant services and improve the reliability of the elasticities. Furthermore with a larger dataset more within-OD variation is expected, especially if OD characteristics and air travel data is also improved, which allows fixed-effects models to achieve better results.

In practice, no single operator holds this kind of large, complete, international dataset, as rail operators generally do not share commercially sensitive data with each other. Overcoming this would require either data-sharing agreements between operators or a centralised data source, such as one coordinated by a European rail body, that operators could contribute volume data to without sharing it directly with competitors, and for example making it available for this kind of research.

Methodology

The OD definition used in this research also suffers from some limitations. First of all the catchment area for the airports is set at a constant distance, while research has shown it can vary depending on the type of airport, other airports that are available and the flights offered at an airport (Lieshout, 2012). In addition, air travel is only considered to be available when there is a direct flight between the origins and destinations while more OD pairs actually have air alternative available as transferring is not included.

In addition, the access and egress time to or from the train station or airport is not considered. This full journey time can therefore differ as some airports are relatively close to a city while others require longer access and egress times. In the model also only OD pairs with air travel available and located within a FUA are considered.

This means that smaller cities are not considered which can bias the estimates for the elasticities. Another issue is that the model assumes that the passenger volume or ticket sales represents the demand. However, if a train is at capacity this value does not represent actual demand. Currently the model has no way to deal with this, which can dampen the demand estimate in the case that trains are at capacity.

Endogeneity is also a problem in the applied methodology, as discussed in subsection 3.6.2. There are a few sources of endogeneity in this research. First, there is the problem of omitted variable bias, especially in the pooled OLS model, where unobserved OD characteristics are not absorbed. Variables such as tourism attractiveness or weather were not included in the model, as sourcing suitable data for these variables is complex, time-consuming and was considered outside the scope of this research. Second, there is the problem of reverse causality, for which the most obvious example is fare. In the model, rail fare is assumed to affect demand, but operators also use estimated demand to set fares. In this way, demand causes the fare instead of fare causing the demand.

Testing and mitigating endogeneity is possible with the use of an instrumental variable (IV) (Wooldridge, 2016). An IV is correlated with the endogenous regressor but affects demand only through that regressor. This isolates the effect of the regressor on demand from the reverse effect of demand on the regressor. For rail fare, a possible IV is a variable that influences the fare but is not set in anticipation of demand, such as track access charges set by infrastructure managers. These raise the costs of operators and therefore the fares, but are not based on demand. For rail frequency, planned infrastructure maintenance could serve as an IV, as it can reduce service frequency for reasons unrelated to passenger demand, but only if the timing is unrelated to expected demand. Implementing IVs is difficult because it needs sufficient variation across OD pairs and over time. In addition, using instruments for multiple regressors is difficult because you need an IV for every endogenous regressor which makes finding suitable variables even more difficult.

The methodology also did not examine any heterogeneity in the results. For some aspects, no suitable data was available. For example, the literature has shown that elasticity values generally differ between business and leisure travellers, with business travellers typically less sensitive to fare and more sensitive to travel time (Brons et al., 2002; Wardman & Batley, 2022). Another aspect worth investigating is heterogeneity across regions or countries, as differences in rail systems and how they are used could produce different elasticities. This would, require a substantially larger and more evenly spread dataset to estimate reliable elasticities per country or region. Heterogeneity across distance bands could also be a promising direction, as different distance bands likely represent different travel markets. It is possible for example, that longer-distance rail travel shows less sensitivity to travel time among those who choose rail, while competition from air travel becomes more important. Heterogeneity analysis can also help with the significance and model fit within, as Hortelano et al. (2016) showed that just one elasticity value often does not generalize very well.

Future research could use a more flexible definition of origins and destinations, for example by constructing more dynamic catchment areas for population and airport access. Also by using a finer grained geographical unit than the FUA to allow smaller cities to be included in the model. Addressing the capacity constraint issue discussed above, possibly through excluding affected observations or applying censored regression methods, where the dependent variable is allowed to be known up to a certain threshold (in this capacity). This would also improve the reliability of the demand estimates. Finally, a heterogeneity analysis, for example by trip purpose, region, or distance band, as described above would provide valuable insight into how elasticities vary, provided sufficient data is available.

Generalizability

The method developed in this research is not tied to NSI data and can in principle be applied to any observed rail demand dataset. In practice however, this study's own results illustrate how much the method depends on data quality: a large, evenly distributed, and temporally extensive dataset is needed to produce reliable and interpretable coefficients, conditions that this study's data only partially met. The forecasting methodology itself is similarly generalizable to other corridors, with accuracy depending mostly on the reliability of the coefficients and the quality of the input data used to construct corridor features. In this research, a correction factor was needed to convert ticket sales into an estimate of absolute demand, this step would be unnecessary and should ideally be avoided if the model were estimated on full passenger volume dataset rather than one operator ticket sales.

The specific coefficients estimated in this study require more caution before being applied elsewhere. Only some estimates are plausible and consistent with the literature, while others are not (especially in the volume

data results), and the low overall model fit means it is difficult to determine whether the estimated relationships actually reflect how rail demand is formed, or just reflect the bias introduced by this study's specific dataset. This uncertainty is amplified for the forecasting model specifically, since the ticket sales data it relies on only contains observations from journeys originating in the Netherlands with destinations outside of The Netherlands. As a result, the estimated coefficients, and any forecasts derived from them, should not be assumed to generalize to other operators, origin countries, or rail markets more broadly.

7.4. Practical implications

In this model demand is found to be driven mostly by population and GDP per capita, therefore NSI should treat market size as a first indication when evaluating candidate corridors, using rail and air service characteristics as secondary considerations. NSI should also be aware of induced demand when interpreting model based forecasts. Forecasts for new corridors should therefore be treated as a starting point, with the expectation that realised demand may grow over time as the market develops. Finally, if NSI wants to apply this type of methodology more broadly and with greater reliability, obtaining more complete demand data, for example through data sharing agreements with partner operators, would help the most and directly address several of the limitations identified in this research.

8

Conclusion

This chapter concludes the research by answering the research questions. First the sub questions are answered, after which they are used to answer the main research question. The interpretation, discussion of the results and future research are covered in chapter 7.

Sub-question 1: *What are the key determinants of passenger demand and how can these variables be quantified?*

The literature review (chapter 2) has shown that there are many factors which can influence rail passenger demand. The most important variables influencing rail demand are travel time, ticket fare and transfers, while GDP (per capita) and population generally drive transport demand. Depending on the distance context, different competing modes also play a role, as for urban, inter-urban and medium distance demand car (travel time and associated costs) plays an important role while for long-distance air travel is more dominant. The effects on these explanatory variables can be expressed as elasticities, which is how much (rail) demand changes in response to a change in an explanatory variable.

Sub-question 2: *How much variation is there in international rail supply and demand to and from The Netherlands over the study period and is it sufficient to identify demand responses?*

The variation in the data is somewhat sufficient to identify demand responses for the ticket sales dataset, but not for the passenger volume datasets. The pooled OLS model exploits cross-OD variation, while the fixed-effects model exploits within-OD variation. In the passenger volume data, the within-OD variation in supply is very limited, with only five specific changes to the supply over the study period. This resulted in implausible and inflated coefficients for the capacity, travel time and population variables, while many other variables remained insignificant. In addition, the limited number of OD pairs provided too little cross-OD variation in the daily and monthly passenger volume datasets to identify plausible demand responses. The ticket sales data provided a more even distribution of the explanatory variables and demand, and therefore led to more plausible results. Enough within-OD variation was available in this dataset, as illustrated by the air and rail fare variables, for which the fixed-effects model gave plausible estimates. Overall, it was possible to identify plausible demand responses for rail fare, travel time, transfers, air fare and GDP using the fixed-effects models, and for population using pooled OLS, but only for the ticket sales dataset. The model fit remained very low, which limits how much weight can be placed on these estimates.

Sub-question 3: *What are the estimated relations between the explanatory variables and passenger demand?*

As the model fit is relatively low, not all variables are significant and not all variables are plausible or agree with the literature, the interpretation of the relations between the explanatory variables and passenger demand requires careful interpretation. Only the significant and plausible results are mentioned to answer this sub-question. The most important drivers of demand are the population and GDP per capita of the destination, of which GDP per capita has the largest influence. These result in elasticities of 0.0843 ($p = 0.0076$) and 0.3670 ($p = 0.0003$) respectively. Rail fare has a negative elasticity of -0.1479 ($p = 0.0000$) thus higher rail fare reduces demand. Similar for rail travel time a elasticity of -0.3684 ($p = 0.0152$) was found so increased travel time reduces demand. Transfers also negatively influence demand with a elasticity of -0.3755 ($p = 0.0000$) in the pooled OLS model and -0.1013 ($p = 0.0000$) in the two-way fixed-effects model. Air fare shows a small substitution effect between air and rail with an elasticity of 0.0706 ($p = 0.0002$) in the two-way fixed-effects model.

Sub-question 4: *How can the estimated relations be applied to forecast demand for potential new international rail services and to identify the most promising corridors to and from The Netherlands.*

The estimated coefficients from the pooled OLS model were applied in the developed forecasting methodology. Pooled OLS was the only suitable model as it is the only model that can predict absolute demand levels for OD pairs that do not have an observed service. For every candidate corridor, the explanatory variables were constructed for every OD pair, using assumed service characteristics. Demand was predicted at an OD level and aggregated to the corridor level.

The method was applied to four candidate corridors, namely: 1A: Amsterdam - Barcelona (HSR), 1B: Amsterdam - Milan (HSR), 2: Amsterdam - Copenhagen, 3: Amsterdam - Prague. Similar demand levels were predicted when considering the OD level, with corridor 1B having the highest average demand per OD pair and corridor 1A having the highest total demand. The alternative scenarios showed that improving travel time does substantially increase demand while increasing fare has very little influence. Transfers show a very clear negative impact on demand, with demand reducing with around 30% in the scenario with a transfer for corridor 1A. However, these results are highly dependent on the coefficients used in the model, and with a more suitable dataset result will likely be different.

The accuracy of these forecasts is limited. It depends on the reliability of the estimated coefficients, which come from a pooled OLS model with a low fit, and on the quality of the input data. Furthermore the predicted tickets sold had to be extrapolated to reach absolute demand levels which introduces further uncertainty. There these forecasting results should be interpreted as a relative ranking of corridors rather than as absolute demand forecasts.

Main question: *To what extent do origin-destination characteristics, rail service attributes, and competing air services influence international rail passenger demand, and how can these estimated relationships be used to forecast demand for new international rail services?*

Origin-destination characteristics, rail service attributes and competing air services all influence international rail passenger demand, but the confidence of the strength and direction of these relationships could only partially be estimated using the ticket sales dataset. It showed that higher rail fares, longer rail travel times and more transfers reduce demand, while a higher GDP per capita and population at the destination increased demand. Higher air fare showed a small positive effect on rail demand, indicating some substitution between these two modes.

The estimated relationships can be used to forecast demand for new international rail services using the developed methodology. By constructing the explanatory variables for each OD pair along the candidate corridor, average daily demand can be predicted at the OD level, which can be aggregated to the corridor level. Applied to four candidate corridors originating in Amsterdam, the forecast showed similar demand levels per OD pair, with Amsterdam - Milan having the highest average demand per OD pair and Amsterdam - Barcelona having the highest total demand. With these coefficients, the scenario analysis showed that reducing travel time substantially increases demand while increasing fare only has a marginal effects. Introducing a transfer has the large impact, with the estimated demand reducing with around 30%. This highlights that there is value in running a direct service compared to itineraries with transfers even when the travel times are equal. However, because the underlying model has a low fit, insignificant variables were included, extrapolation was required and the service characteristics are assumed, these results are best interpreted as a relative ranking rather than absolute demand forecasts.

References

Allard, R. F., & Moura, F. (2018). Effect of transport transfer quality on intercity passenger mode choice. *Transportation Research Part A: Policy and Practice*, *109*, 89–107. https://doi.org/10.1016/j.tra.2018.01.018

Avogadro, N., & Redondi, R. (2023). Diverted and induced demand: Evidence from the London-Paris passenger market. *Research in Transportation Economics*, *100*, 101304. https://doi.org/10.1016/j.retrec.2023.101304

Baltagi, B. H. (2008). Forecasting with panel data [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.1047]. *Journal of Forecasting*, *27*(2), 153–173. https://doi.org/10.1002/for.1047

Baltagi, B. H. (2021). *Econometric Analysis of Panel Data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-53953-5

Banerjee, N., Morton, A., & Akartunalı, K. (2020). Passenger demand forecasting in scheduled transportation. *European Journal of Operational Research*, *286*(3), 797–810. https://doi.org/10.1016/j.ejor.2019.10.032

Binnen 6 uur van Amsterdam naar Berlijn | NS International. (n.d.). Retrieved May 26, 2026, from https://www.nsinternational.com/nl/vragen-en-antwoorden/www.nsinternational.com/nl

Birolini, S., Cattaneo, M., Malighetti, P., & Morlotti, C. (2020). Integrated origin-based demand modeling for air transportation. *Transportation Research Part E: Logistics and Transportation Review*, *142*, 102050. https://doi.org/10.1016/j.tre.2020.102050

Börjesson, M. (2014). Forecasting demand for high speed rail. *Transportation Research Part A: Policy and Practice*, *70*, 81–92. https://doi.org/10.1016/j.tra.2014.10.010

Brons, M., Pels, E., Nijkamp, P., & Rietveld, P. (2002). Price elasticities of demand for passenger air travel: A meta-analysis. *Journal of Air Transport Management*, *8*(3), 165–175. https://doi.org/10.1016/S0969-6997(01)00050-3

Castillo-Manzano, J. I., Pozo-Barajas, R., & Trapero, J. R. (2015). Measuring the substitution effects between High Speed Rail and air transport in Spain. *Journal of Transport Geography*, *43*, 59–65. https://doi.org/10.1016/j.jtrangeo.2015.01.008

Clewlow, R. R., Sussman, J. M., & Balakrishnan, H. (2014). The impact of high-speed rail and low-cost carriers on European air passenger traffic. *Transport Policy*, *33*, 136–143. https://doi.org/10.1016/j.tranpol.2014.01.015

Diaz-Gutierrez, J., & Ranjbari, A. (2025). How effective are fixed-effects models in fixing the transit supply–demand bidirectional interaction? [_eprint: https://doi.org/10.1080/19427867.2024.2422713]. *Transportation Letters*, *17*(7), 1199–1212. https://doi.org/10.1080/19427867.2024.2422713

Dobruszkes, F. (2011). High-speed rail and air transport competition in Western Europe: A supply-oriented perspective. *Transport Policy*, *18*(6), 870–879. https://doi.org/10.1016/j.tranpol.2011.06.002

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association*, *78*(383), 605–610. https://doi.org/10.2307/2288126

Eurostar Sun | Thalys Zon treinen naar Zuid-Frankrijk. (n.d.). Retrieved May 18, 2026, from https://www.eurostar.com/nl-nl/speciale-aanbiedingen/eurostar-zon-je-reis-naar-de-provence

Fale, M., Kramberger, T., Rupnik, B., Lixin, S., & Vizinger, T. (2025). Evaluating Middle-Distance Passenger Transportation: A Comparison of the Time Efficiency, Sustainability, and Break-Even Distances Between High-Speed-Rail and Air Travel in Sweden. *Sustainability*, *17*(2), 386. https://doi.org/10.3390/su17020386

Faraway, J., & Chatfield, C. (1998). Time Series Forecasting with Neural Networks: A Comparative Study Using the Airline Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *47*(2), 231–250. Retrieved March 2, 2026, from https://www.jstor.org/stable/2988352

Fildes, R., Wei, Y., & Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting*, *27*(3), 902–922. https://doi.org/10.1016/j.ijforecast.2009.06.002

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2005). How (In)accurate Are Demand Forecasts in Public Works Projects?: The Case of Transportation [_eprint: https://doi.org/10.1080/01944360508976688]. *Journal of the American Planning Association*, *71*(2), 131–146. https://doi.org/10.1080/01944360508976688

Fourth railway package of 2016 - Mobility and Transport. (2016). Retrieved September 29, 2026, from https://transport.ec.europa.eu/transport-modes/rail/railway-packages/fourth-railway-package-2016_en

Givoni, M. (2007). Environmental Benefits from Mode Substitution: Comparison of the Environmental Impact from Aircraft and High-Speed Train Operations. *International Journal of Sustainable Transportation*, *1*(4), 209–230. https://doi.org/10.1080/15568310601060044

Gundelfinger-Casar, J., & Coto-Millán, P. (2017). Intermodal competition between high-speed rail and air transport in Spain. *Utilities Policy*, *47*, 12–17. https://doi.org/10.1016/j.jup.2017.06.001

Holmgren, J. (2007). Meta-analysis of public transport demand. *Transportation Research Part A: Policy and Practice*, *41*(10), 1021–1035. https://doi.org/10.1016/j.tra.2007.06.003

Hortelano, A. O., Guzman, A. F., Preston, J., & Vassallo, J. M. (2016). Price Elasticity of Demand on the High-Speed Rail Lines of Spain: Impact of the New Pricing Scheme. *Transportation Research Record*, *2597*(1), 90–98. https://doi.org/10.3141/2597-12

Ibarra-Rojas, O. J., Delgado, F., Giesen, R., & Muñoz, J. C. (2015). Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological*, *77*, 38–75. https://doi.org/10.1016/j.trb.2015.03.002

Improvements to German railways tracks | NS International. (n.d.). Retrieved May 18, 2026, from https://www.nsinternational.com/en/questions-and-answers/www.nsinternational.com/en

Kampp, M., Sedelmeier, J., Schüth, J., Thust, M., Kaiser, D., Scherr, W., Schlaich, J., & Senk, P. (2025). Design of a European high-speed rail network and use of passenger demand forecasting to test European policy targets. *European Transport Research Review*, *17*(1), 23. https://doi.org/10.1186/s12544-025-00715-y

Kitamura, R. (1990). Panel analysis in transportation planning: An overview. *Transportation Research Part A: General*, *24*(6), 401–415. https://doi.org/10.1016/0191-2607(90)90032-2

Kristoffersson, I., & Liu, C. (2024). Estimation of demand models for long-distance international travel - Key determinants for Swedes' travel abroad [Num Pages: 27 Web of Science ID: WOS:001388740200004]. *European Journal of Transport and Infrastructure Research*, *24*(4), 62–88. https://doi.org/10.59490/ejtir.2024.24.4.6696

Lieshout, R. (2012). Measuring the size of an airport's catchment area. *Journal of Transport Geography*, *25*, 27–34. https://doi.org/10.1016/j.jtrangeo.2012.07.004

Litman, T. (2025). How Prices and Other Factors Affect Travel Behavior.

Nicolaisen, M. S., & Driscoll, P. A. (2014). Ex-Post Evaluations of Demand Forecast Accuracy: A Literature Review [_eprint: https://doi.org/10.1080/01441647.2014.926428]. *Transport Reviews*, *34*(4), 540–557. https://doi.org/10.1080/01441647.2014.926428

OECD. (n.d.-a). Economic indicators by territorial typology on access to functional urban areas. https://data-explorer.oecd.org/vis?lc=en&df%5Bds%5D=dsDisseminateFinalDMZ&df%5Bid%5D=DSD_REG_ECO%40DF_TYPE_METRO&df%5Bag%5D=OECD.CFE.EDS&df%5Bvs%5D=1.0&dq=A..AUT%2BBEL%2BCHE%2BCHL%2BCOL%2BCZE%2BDEU%2BDNK%2BESP%2BEST%2BFIN%2BFRA%2BGBR%2BGRC%2BHUN%2BIRL%2BITA%2BJPN%2BKOR%2BLTU%2BLUX%2BLVA%2BNLD%2BNOR%2BNZL%2BPOL%2BPRT%2BSVK%2BSVN%2BSWE%2BTUR%2BUSA..GDP..Q.USD_PPP_PS&lom=LASTNPERIODS&lo=5&to%5BTIME_PERIOD%5D=false&vw=ov

OECD. (n.d.-b). OECD Data Explorer • Population and built-up density - Cities and FUAs. Retrieved May 4, 2026, from http://data-explorer.oecd.org/s/1e6

OECD Data Explorer • Economy - FUAs. (n.d.). Retrieved June 8, 2026, from https://data-explorer.oecd.org/vis?lc=en&df%5Bds%5D=dsDisseminateFinalDMZ&df%5Bid%5D=DSD_FUA_ECO%40DF_ECONOMY&df%5Bag%5D=OECD.CFE.EDS&dq=AT001F%2BAT002F%2BAT003F%2BAT004F%2BAT005F%2BAT006F.A.LAB_PROD..&lom=LASTNPERIODS&lo=5&to%5BTIME_PERIOD%5D=false&vw=tb

OECD Definition of Cities and Functional Urban Areas. (n.d.). Retrieved May 4, 2026, from https://www.oecd.org/en/data/datasets/oecd-definition-of-cities-and-functional-urban-areas.html

Ortúzar, J. d. D., & Willumsen, L. G. (2024, February). *Modelling Transport* [Google-Books-ID: kMn1EAAAQBAJ]. John Wiley & Sons.

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J., & White, P. (2006). The demand for public transport: The effects of fares, quality of service, income and car ownership. *Transport Policy*, *13*(4), 295–306. https://doi.org/10.1016/j.tranpol.2005.12.004

Polzin, S. E., Batur, I., & Pendyala, R. M. (2025). Sustaining the relevance of transportation planning and forecasting in a fast-changing world. *Transport Policy*, *168*, 279–287. https://doi.org/10.1016/j.tranpol.2025.04.019

Rich, J., & Mabit, S. L. (2012). A Long-Distance Travel Demand Model for Europe. *European Journal of Transport and Infrastructure Research*, *12*(1). https://doi.org/10.18757/ejtir.2012.12.1.2946

Rocco di Torrepadula, F., Napolitano, E. V., Di Martino, S., & Mazzocca, N. (2024). Machine Learning for public transportation demand prediction: A Systematic Literature Review. *Engineering Applications of Artificial Intelligence*, *137*, 109166. https://doi.org/10.1016/j.engappai.2024.109166

Román, C., Martín, J. C., Espino, R., Cherchi, E., Ortúzar, J. d. D., Rizzi, L. I., González, R. M., & Amador, F. J. (2014). Valuation of travel time savings for intercity travel: The Madrid-Barcelona corridor. *Transport Policy*, *36*, 105–117. https://doi.org/10.1016/j.tranpol.2014.07.007

Train Stations in Europe. (n.d.). Retrieved June 8, 2026, from https://www.kaggle.com/datasets/headsortails/train-stations-in-europe

Treinarchief. (n.d.). Retrieved June 22, 2026, from https://www.rijdendetreinen.nl/treinarchief

Tsekeris, T., & Tsekeris, C. (2011). Demand Forecasting in Transport: Overview and Modeling Advances [_eprint: https://doi.org/10.1080/1331677X.2011.11517446]. *Economic Research-Ekonomska Istraživanja*, *24*(1), 82–94. https://doi.org/10.1080/1331677X.2011.11517446

Turin-Lyon Base Tunnel. (n.d.). Retrieved September 10, 2026, from https://www.fsitaliane.it/content/fsitaliane/en/strategic-projects/turin-lyon-base-tunnel.html

Wardman, M. (1997). Inter-urban rail demand, elasticities and competition in Great Britain: Evidence from direct demand models. *Transportation Research Part E: Logistics and Transportation Review*, *33*(1), 15–28. https://doi.org/10.1016/S1366-5545(97)83195-X

Wardman, M. (2006). Demand for rail travel and the effects of external factors. *Transportation Research Part E: Logistics and Transportation Review*, *42*(3), 129–148. https://doi.org/10.1016/j.tre.2004.07.003

Wardman, M. (2022a). Meta-analysis of British time-related demand elasticity evidence: An update. *Transportation Research Part A: Policy and Practice*, *157*, 198–214. https://doi.org/10.1016/j.tra.2022.02.001

Wardman, M. (2022b). Meta-analysis of price elasticities of travel demand in great britain: Update and extension. *Transportation Research Part A: Policy and Practice*, *158*, 1–18. https://doi.org/10.1016/j.tra.2022.01.020

Wardman, M. (2025). Investigating demand models with more flexible elasticity functions: Empirical insights from rail demand analysis. *Transportation*, *52*(4), 1457–1483. https://doi.org/10.1007/s11116-024-10462-z

Wardman, M., & Batley, R. (2022). The demand impacts of train punctuality in great britain: Systematic review, meta-analysis and some new econometric insights. *Transportation*, *49*(2), 555–589. https://doi.org/10.1007/s11116-021-10186-4

Wardman, M., Lythgoe, W., & Whelan, G. (2007). Rail Passenger Demand Forecasting: Cross-Sectional Models Revisited. *Research in Transportation Economics*, *20*, 119–152. https://doi.org/10.1016/S0739-8859(07)20005-8

Wardman, M., Toner, J., Fearnley, N., Flügel, S., & Killi, M. (2018). Review and meta-analysis of inter-modal cross-elasticity evidence. *Transportation Research Part A: Policy and Practice*, *118*, 662–681. https://doi.org/10.1016/j.tra.2018.10.002

Why we're building the Fehmarnbelt fixed link. (n.d.). Retrieved September 10, 2026, from https://femern.com/why-we-re-building-the-fehmarnbelt-fixed-link/

Wijeweera, A., To, H., Charles, M. B., & Sloan, K. (2014). A time series analysis of passenger rail demand in major Australian cities. *Economic Analysis and Policy*, *44*(3), 301–309. https://doi.org/10.1016/j.eap.2014.08.003

Wong, H., & Yap, M. (2023). A data driven approach to update public transport service elasticities. *Journal of Public Transportation*, *25*, 100066. https://doi.org/10.1016/j.jpubtr.2023.100066

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, *68*(1), 115–132. https://doi.org/10.1016/0304-4076(94)01645-G

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (Sixth edition, student edition). Cengage Learning.

Yao, E., & Morikawa, T. (2005). A study of an integrated intercity travel demand model. *Transportation Research Part A: Policy and Practice*, *39*(4), 367–381. https://doi.org/10.1016/j.tra.2004.12.003

A

Data

Appendix removed for confidentiality

B

Scientific paper

Scientific paper starts on next page

Keeping rail demand on track: Using panel data to estimate international long-distance rail demand elasticities

W.S.J. (Wouter) van der Hoorn
Delft University of Technology, The Netherlands
Public version

Rail travel emits considerably less than air travel, which makes improving rail travel an important element of the goals of the EU. However, to achieve this, many more services and much more infrastructure are needed. Furthermore, introducing new services is costly and a risk for operators. Therefore, operators want to get more insight into what the expected demand is before introducing a new service. Demand for new services cannot be observed and must be inferred from existing services. This research estimates demand elasticities for international long-distance rail and develops a method to apply them to forecast demand on candidate corridors. Origin-destination (OD) demand is modelled with pooled OLS, OD fixed-effects and OD+time fixed-effects panel models, using two demand measures: passenger volumes on a limited number of services and ticket sales from NS International. The passenger volume results showed some implausible estimates (such as a positive coefficient for travel time) and are therefore not considered reliable. The ticket sales models give more plausible estimates: higher rail fares, longer travel times and more transfers are associated with lower demand (elasticities of -0.15 for fare under fixed-effects, and -0.36 for travel time and -0.38 for transfers in pooled OLS), while a higher air fare is associated with slightly higher rail demand within OD pairs (0.07 under fixed-effects), suggesting a small substitution effect between air and rail. GDP and population drive demand, with the destination attracting more demand than the origin generates (0.017 and 0.084 elasticity for origin and destination population respectively, and 0.171 and 0.367 GDP elasticity). Model fit is low (R^2 of 0.157 for pooled OLS and within-R^2 of 0.022 under fixed-effects). The coefficients from the pooled OLS model are applied to four candidate corridors from Amsterdam, namely Milan (HSR), Barcelona (HSR), Prague and Copenhagen. Scenario analysis showed that reducing travel time increases demand by 8 to 20%, changing fare has little to no effect and introducing a transfer decreases demand by around 30%. The Milan (HSR) corridor showed the highest average demand per OD pair and is therefore considered most promising.

Key words: International rail demand, long-distance, demand elasticities, panel data, fixed-effects, air-rail competition, demand forecasting

1. Introduction

Rail travel emits considerably less emissions than air travel (Givoni 2007). Shifting passengers from air to rail is therefore an important element of reducing the negative impacts of passenger transport and an important part of the goals of the EU (Kampp et al. 2025). Achieving this shift requires more and better rail services including new high-speed and long-distance rail connections as well as infrastructure (Kampp et al. 2025). NS International (NSI) can help this shift by identifying new long-distance international rail routes where demand exists but competitive rail options are not available. To achieve this an understanding of how demand for long-distance rail is formed is needed (Banerjee et al. 2020).

Forecasting demand for new services is difficult, since no historical observations exist for these specific connections. Instead demand must be inferred from relationships observed on existing services. Previous research has shown that explanatory variables such as travel time, fares and service frequency play an important role in explaining passenger demand (Avogadro and Redondi 2023, Wardman and Batley 2022,

Kristoffersson and Liu 2024). Demand is also influenced by external factors such as GDP, population and competing transport modes (Wardman 2006). Air transport tends to dominate on the medium-distance market (300 to 1000 km) (Fale et al. 2025), and reductions in rail travel time have been shown to shift passengers from air to rail (Gundelfinger-Casar and Coto-Millán 2017). Rail demand on these routes therefore depends not only on the rail service itself, but also on what air transport offers on the same connection.

The rail market is getting more deregulated and competition is stimulated, which means that operators such as NSI get more freedom in how and to which destinations they run their international train services (Fou 2016). Each new service is costly and a risk, therefore operators want more insight into how rail demand is formed and what corridors would be the most promising for introducing new services. This requires estimates of how international rail demand responds to origin-destination characteristics, rail service attributes and competition from air, as well as a method to apply the estimates to currently unserved corridors.

The existing literature only partly provides this. First, empirical rail demand elasticities mostly originate from single-country or market studies, and international estimates for long-distance rail are rare, which makes generalization difficult (Rich and Mabit 2012). Second, studies that estimate the effects of multiple demand drivers at the same time, such as fare and travel time, origin-destination characteristics and competition from other modes are limited. Finally, long-distance and international rail demand remains understudied compared to single country, inter-urban and commuting markets.

This study aims to address these gaps in two ways. First, by exploiting variation between origin-destination (OD) pairs and over time, econometric regression models are used to estimate long-distance rail demand elasticities for origin-destination characteristics, rail service attributes and air travel competition. Second, it develops a method to apply the estimated elasticities to forecast demand for currently unserved corridors, which can be used to rank the corridors with the most promising demand. The research is guided by the following research question: *To what extent do origin-destination characteristics, rail service attributes, and competing air services influence international rail passenger demand, and how can these estimated relationships be used to forecast demand for new international rail services?*

2. Literature review

2.1. Demand forecasting methods

The focus of this research lies in estimating long-distance international rail demand elasticities and forecasting demand for new services, therefore the focus lies on forecasting methods on the strategic level. Banerjee et al. (2020) reviewed more than 120 papers on passenger demand forecasting for scheduled transportation and identified time series methods, causal methods and artificial intelligence (AI) methods as the main quantitative approaches. Time series methods derive their prediction from past observations, typically through extrapolation techniques. They are useful for small changes to existing services, but not suitable for new services, since no historical observations exist from which to extrapolate. Causal methods establish the factors that explain past demand and use these to model a forecast, which makes them more suitable for new services.

Causal models generally predict travel volume, modal choice or route choice. Logit models are used extensively to estimate choices between travel modes and are typically applied to disaggregate data, whereas regression and econometric models are usually estimated on aggregate origin-destination (OD) level data (Ortúzar and Willumsen 2024). Disaggregate models capture individual-level behavioural responses, while aggregate models reflect the combined effect of these responses across a market or geographic area. Econometric models describe demand as a function of explanatory variables such as travel time, service frequency, ticket price and measures of market size, through linear or log-linear regression (Tsekeris and Tsekeris 2011).

They can also incorporate a gravity model, and are usually direct demand models, meaning that the model directly deals with variations in aggregate rail demand rather than separate components of trip generation, distribution, and mode choice (Wardman et al. 2007). Machine learning methods such as neural networks can be accurate due to their ability to capture non-linear relationships, but risk over-fitting and are subject to poor interpretability. These limitations make them less suitable for this research.

2.2. Determinants of passenger demand

The effect of explanatory variables on demand is commonly expressed as an elasticity, which is how much demand changes in relation to the change in the explanatory variable. The most researched variables are travel time and ticket fare. These can be combined in a generalized cost function using value of time, in a generalised journey time and fare approach, or specified separately (Wardman et al. 2007, Wardman 2025). However, elasticities depend on the context. They vary by travel mode, journey purpose, distance and whether they are short-run or long-run, with long-distance trips typically showing different elasticities than inter urban trips and business travel generally having a lower fare elasticity than leisure travel (Wardman and Batley 2022, Brons et al. 2002). Travel time elasticities are also context dependent (Wardman 2022a), demand responds more strongly to service degradations than to equivalent improvements (Wong and Yap 2023), and transfers consistently negatively influence the travel experience and mode choice (Allard and Moura 2018).

Other than fare and travel time, Wardman (2006) showed that (the growth of) GDP is important, while car time, fuel cost, and population also influence demand. This is reinforced by Holmgren (2007) and Paulley et al. (2006), who identify car ownership, income, and level of service as key determinants of public transport demand. The relative importance of these determinants varies considerably by geography and context (Wijeweera et al. 2014).

When using service supply variables as determinants, the endogeneity between demand and supply should be considered. Supply influences demand, but demand also influences supply. For example, a service with high frequency can be the result of high demand, but high frequency can also induce new demand (Wardman et al. 2007, Tsekeris and Tsekeris 2011). Without accounting for this, the model can produce biased and/or inconsistent results (Diaz-Gutierrez and Ranjbari 2025). Fixed-effects approaches partially mitigates these effects as they absorb time-invariant unobserved characteristics, and instrumental variables (Diaz-Gutierrez and Ranjbari 2025, Birolini et al. 2020) can be used to isolate the effect without reverse causality. Service improvements also induce new demand rather than only shifting mode choice (Yao and Morikawa 2005), as shown by the introduction of Eurostar services between Paris and London, where total flows increased (Avogadro and Redondi 2023).

2.3. Mode competition on the medium to long-distance market

Medium distance in the passenger transport market is commonly defined as trips between 300 and 1000 km (Fale et al. 2025). Air-rail competition research has shown that the modes are most competitive within this distance category, as high speed rail can effectively compete on total travel time. Fale et al. (2025) found that for Sweden the break-even point between air and rail is 400 km, and Gundelfinger-Casar and Coto-Millán (2017) found that the reduction in travel time due to high speed rail influenced the air travel market. Travel time and fare are the most important factors in influencing if rail can replace air travel.

Several studies provide cross-elasticities that quantify the magnitude of air and rail influencing the other mode. Wardman et al. (2018) report air-rail and rail-air cross-elasticities of 0.28 and 0.15 respectively, which tend to decrease with distance. Improved rail quality and pricing were also found to decrease air travel demand in specific corridors (Clewlow et al. 2014, Román et al. 2014, Hortelano et al. 2016). Not all demand shifted to rail is diverted demand, as Castillo-Manzano et al. (2015) found that only 13.9% of high speed rail demand in Spain can be attributed to substitution from air travel.

2.4. Research gap

The literature shows that a substantial body of empirical evidence exists on the determinants of rail demand and on the competition between rail and air. However, several gaps remain. First, most empirical elasticity estimates originate from single-country studies, such as the UK (Wardman 2006, Wardman et al. 2007, Wardman 2022b) or Spain (Román et al. 2014, Hortelano et al. 2016, Castillo-Manzano et al. 2015). Some meta-analyses exist (Rich and Mabit 2012, Wardman et al. 2018, Wardman 2022a), but international elasticity estimates for long-distance rail are rare. Second, studies that estimate rail demand as a function of own price and travel time, origin-destination characteristics and cross-modal effects simultaneously are uncommon, as most focus on one or two of these. The influence of air travel on rail demand is well researched for individual corridors and countries, but rarely incorporated as an explanatory variable in a broader, international rail demand model. Finally, although long-distance trips account for a small proportion of total trips, they account for a large proportion of the distance and thus emissions. (Rich and Mabit 2012). Long-distance and international rail demand remains comparatively understudied (Kristoffersson and Liu 2024).

3. Methodology

This section describes the methodology used to answer the main research question and estimate the effects of origin-destination characteristics, rail service attributes and air travel competition on international rail demand, and to apply the estimates to currently unserved corridors. The main steps are summarized in Figure 1. Every box links to the subsection where the step is described, except for data, results, literature review and case study, which are covered in their own sections.

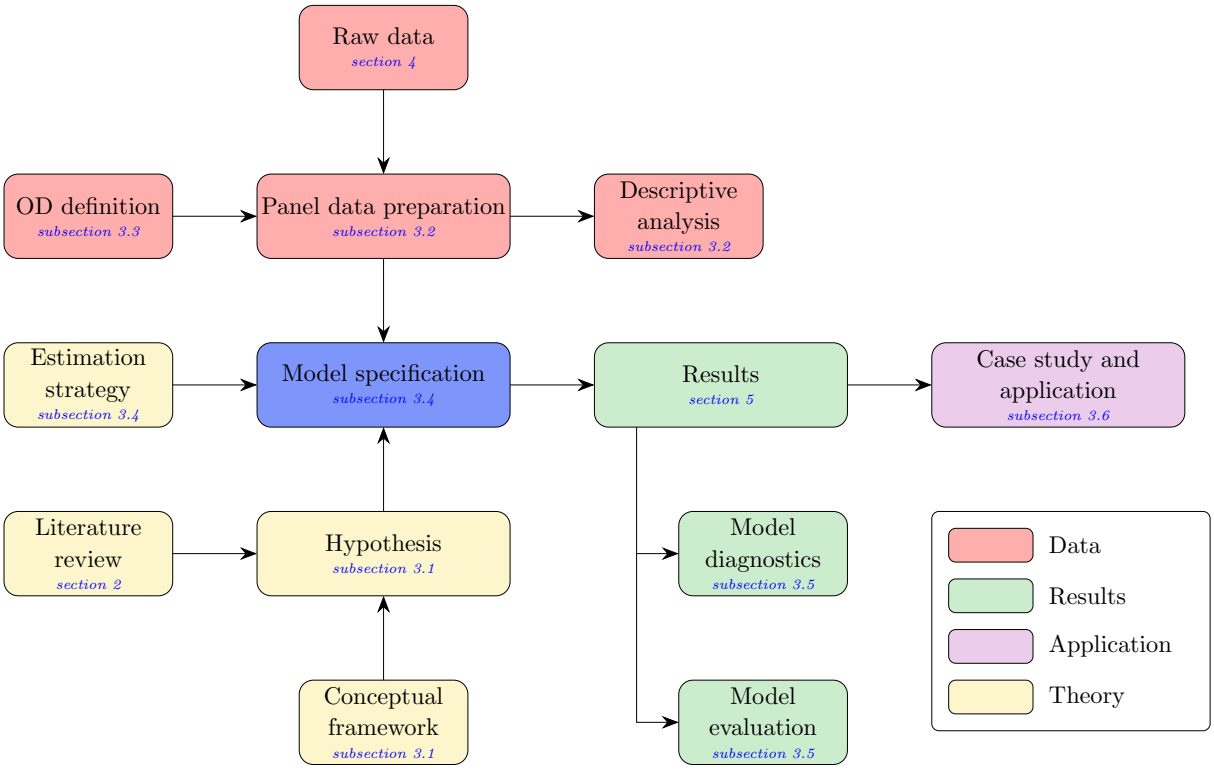


Figure 1 **Methodology overview**

3.1. Conceptual framework and hypotheses

This research models rail demand between an origin and destination as a function of rail service characteristics, the attractiveness of the origin and destination and air travel characteristics as a competing alternative (Figure 2). Fare, travel time and number of transfers are assumed to reduce rail demand when they increase, while frequency and seat capacity are assumed to increase rail demand by improving the level of service. An increase in air fare is assumed to increase rail demand, and an increase in air frequency to reduce it, since air is assumed to serve as a substitute for rail in long-distance travel. Rail demand is further assumed to depend on the population, GDP and distance between the origins and destinations. The framework implies one-way causality, which in practice is not fully correct since operators set fares and frequencies in anticipation of demand. This is discussed in subsection 3.5.

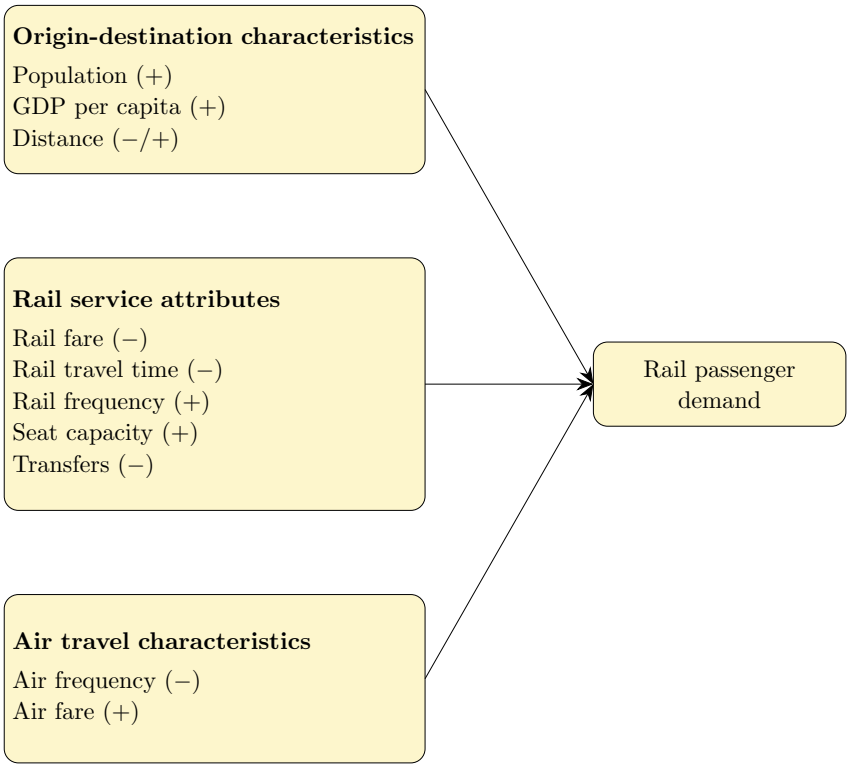


Figure 2 **Conceptual model**

For every explanatory variable a hypothesis is formulated based on the reviewed literature and theoretical judgment (Table 1). Some of the reviewed studies are disaggregate choice studies, which capture mode diversion, whereas the aggregate models used here also captures trip generation and destination substitution. The hypotheses are used to evaluate the model results.

Variable	Notation	Expected sign and magnitude	Hypothesis	References
Origin-destination characteristics				
Population	pop_{it}	+0.1 to 4.0	H1	(Wijeweera et al. 2014, Wardman 2006, Wardman et al. 2007)
GDP per capita	gdp_{it}	+0.3 to 1.0	H2	(Wardman et al. 2018, Wardman and Batley 2022, Wardman 2006, Paulley et al. 2006, Börjesson 2014)
Distance	D_{ij}	-0.5 to +0.5	H3	(Paulley et al. 2006, Hortelano et al. 2016)
Rail service supply attributes				
Rail fare	$f_{rail,ijt}$	-0.4 to -0.8	H4	(Wardman 2022b, Wardman and Batley 2022, Wardman 2006, Román et al. 2014, Kristoffersson and Liu 2024, Börjesson 2014, Hortelano et al. 2016, Wijeweera et al. 2014)
Rail travel time	$TT_{rail,ijt}$	-0.3 to -0.8	H5	(Wardman 2022b, 2006, Román et al. 2014, Paulley et al. 2006, Börjesson 2014)
Rail frequency	$freq_{rail,ijt}$	+0.1 to 0.5	H6	(Wijeweera et al. 2014, Román et al. 2014, Kristoffersson and Liu 2024, Paulley et al. 2006)
Seat capacity	cap_{ijt}	+ (no magnitude forecast)	H7	*Variable not reported in reviewed literature*
Transfers	TR_{ijt}	- (no magnitude forecast)	H8	(Wardman 2006, Paulley et al. 2006, Börjesson 2014)
Air travel characteristics				
Air frequency	$freq_{air,ij}$	-0.1 to -0.5	H9	(Román et al. 2014)
Air fare	$fare_{air,ijt}$	+0.1 to 0.7	H10	(Hortelano et al. 2016, Román et al. 2014, Wardman et al. 2018)

Table 1 Overview of variables and their expected relation to demand, reported values are elasticities

3.2. Panel data

The raw data sources are harmonized into two panel datasets in which each observation is an OD pair on a specific date with all associated attributes. One uses ticket sales as the demand measure and the other passenger volume. They are kept separate because ticket sales cover only tickets sold through NSI, while passenger volume counts all travellers on the observed train services. Part of the passenger volume data is only available per month, for which the time index is monthly.

3.3. Origin-destination definition

Train stations are the leading origins and destinations, and all other variables are linked to them. For the origin and destination characteristics the functional urban area (FUA) is used, defined as a city of at least 50,000 inhabitants plus the surrounding area from which at least 15% of the population commutes to the city (OEC b). This is used because the focus lies on long-distance travel, where it is assumed that international train services have a relatively large catchment area. Stations within the same FUA receive the same characteristics, and stations outside any FUA are excluded.

Airports within 75 km of a station are considered alternatives, with a maximum of five airports per station. This is somewhat shorter than the catchment areas found in the literature (Lieshout 2012), since the OD pairs are all at short-haul flight distances. Air frequency is the sum of all flights between the airports of the origin and destination. Air fare is the weighted mean fare between all airports in the catchment areas, where first the weighted mean fare per carrier is determined for every airport pair.

3.4. Estimation strategy and model specification

Panel data allows the analysis to exploit variation both between OD pairs and within OD pairs over time (Baltagi 2021). Three specifications are estimated for each dataset. A pooled ordinary least squares (OLS) model is the baseline and uses both between and within-OD variation, but assumes that unobserved characteristics are not correlated with the explanatory variables, which is likely to be violated. For example, OD pairs with high demand will likely receive high frequency services. Fixed-effects models partially mitigates this by allowing each OD pair to have its own constant α_{ij}, so that relationships are estimated from changes within the same OD pair over time. Finally, time fixed-effects λ_t are added to account for factors affecting all OD pairs at the same time, such as seasonal demand patterns.

All variables except the number of transfers are log-transformed, so the coefficients β_k are directly interpreted as elasticities. This also reduces the influence of a few high-volume OD pairs on the right skewed demand data. The number of transfers TR_{ijt} is kept in levels as it takes small integer values including zero. With V_{ijt} the demand measure (tickets sold or passenger volume) of OD pair ij on day t, the three specifications are:

$$\ln(V_{ijt}) = \alpha + \sum_k \beta_k \ln(X_{k,ijt}) + \sum_s \delta_s \, servicedummy_{s,ij} + \epsilon_{ijt} \tag{1}$$

$$\ln(V_{ijt}) = \alpha_{ij} + \sum_k \beta_k \ln(X_{k,ijt}) + \epsilon_{ijt} \tag{2}$$

$$\ln(V_{ijt}) = \alpha_{ij} + \lambda_t + \sum_k \beta_k \ln(X_{k,ijt}) + \epsilon_{ijt} \tag{3}$$

The explanatory variables X_k include both time-invariant and time varying variables. Variables that do not vary over time are absorbed by the OD fixed-effects and can only be estimated in the pooled model. All specifications use entity-clustered standard errors, which are robust to heteroscedasticity across OD pairs of different sizes and to correlation of the error terms within each OD pair over time (Wooldridge 2016).

3.5. Model diagnostics and evaluation

Multicollinearity is checked with the variance inflation factor, $VIF_k = 1/(1 - R_k^2)$, where R_k^2 is obtained by regressing X_k on all other explanatory variables. Values above 5 are reported. Endogeneity is a problem because fare, frequency, capacity are set by the operator. OD fixed-effects protect against time-invariant omitted variables, but not against reverse causality over time, and no instrumental variable was available

to formally test for reverse causality. The coefficients of these variables should therefore be interpreted with this limitation in mind.

The models are evaluated on three criteria. The first is statistical: R^2, within R^2, and the significance of the coefficients (p-value). The second are the hypotheses from Table 1, testing whether the sign and magnitude of the estimates are as expected. The third is theoretical plausibility given the characteristics of the data, since a model can fit well without giving plausible relationships.

3.6. Forecasting approach

The results are applied to the following methodology to forecast demand for currently unserved corridors. Forecasting demand for corridors without historical data follows two stages (Figure 3). In the first stage a baseline demand is estimated for each candidate corridor, and in the second stage the rail service attributes are varied to see how the estimated demand responds.

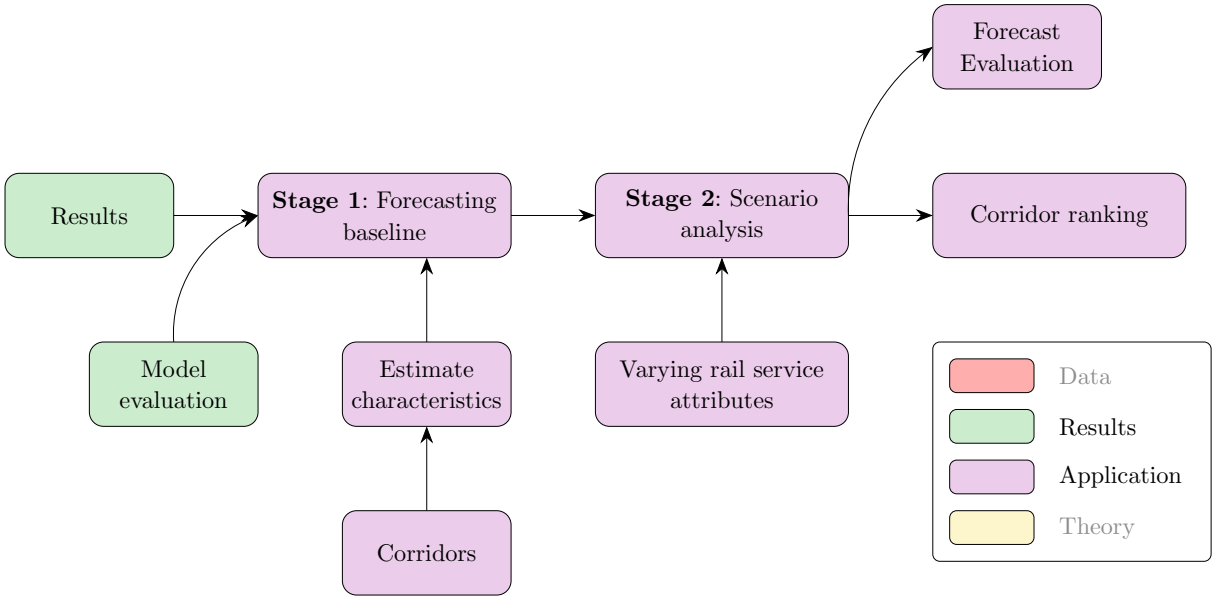


Figure 3 **Application methodology overview**

The pooled OLS model is the only specification able to produce an absolute demand level for OD pairs outside the estimation sample, since the fixed-effects models absorb all cross-sectional variation and the constants α_{ij} are OD specific. The full pooled model, including insignificant coefficients, is used because dropping them biases the estimated demand. Whether the ticket sales or passenger volume model is used depends which results are the most plausible.

Corridors are defined as ordered sequences of stations served by a single train service. They are selected to be longer than 300 km with a maximum travel time of 12 hours, not currently served by direct services from The Netherlands. They should be relevant for NSI to operate, and to have origins and destinations within a FUA, with a maximum of one stop per FUA. This prevents the same population from being counted twice. For every corridor the predictor variables are constructed from the same sources as the panel data. The base rail service attributes are based on expert knowledge within NSI.

Exponentiating the predicted log demand gives the median instead of the mean of demand. This is corrected with Duan's smearing factor $\hat{\phi} = \frac{1}{n}\sum \exp(\hat{\epsilon}_{ijt})$, calculated from the residuals of the estimation sample (Duan

1983). When the ticket sales model is used, the forecast also represents only NSI ticket sales. It is therefore divided by δ, the average share of tickets sold by NSI observed on corridors with passenger volume data longer than 300 km. Corridors are ranked by three metrics: total demand, average demand per OD pair and demand between the endpoints. The model to forecast demand becomes:

$$\hat{V}_{ijt} = \frac{\exp\left(\hat{\alpha} + \sum_k \hat{\beta}_k \ln(X_{k,ijt})\right) \cdot \hat{\phi}}{\delta} \tag{4}$$

4. Data

This section describes the data sources, the processing into two panel datasets, and the characteristics of the data that are used to help interpreting the results.

4.1. Data sources

Five raw data sources data are used. Historical passenger volume data was provided by NSI for the ICE services from Amsterdam to Berlin, Frankfurt, Munich and Basel, and for the Eurostar services to Paris and the seasonal Eurostar Snow and Sun. The ICE data is available per day, OD pair and train for periods between 2020 and April 2026, depending on the service. The Eurostar data is available per month between January 2024 and December 2025, and includes the average fare. Timetable data (travel time, frequency and seat capacity) was provided by NSI or derived from the Train Archive (Tre). Ticket sales data from NSI covers all tickets sold from The Netherlands to international destinatiosn(395 origins and 5660 destinations) between July 2023 and April 2026 at daily OD level, with the average fare, travel time and number of transfers. Because it only includes tickets sold by NSI, it does not represent total demand and is likely biased towards Dutch travellers and shorter distances. Air travel data (OAG, processed by TU Delft) covers 284 airports and contains the flight frequency and duration per carrier and the monthly fare, between March 2022 and December 2025. Finally, GDP and population per FUA (2019 to 2025) (OECD b, OEC a), the FUA boundaries (OEC b) and a database of about 36,000 European train stations (Tra) are used.

4.2. Data processing

The processing harmonizes these sources into two panel datasets, as shown in Figure 4. The train station database is the leading dataset. Each station is matched to its functional urban area (FUA) (with GDP per capita and population) and to up to five airports within 75 km (see subsection 3.3). Station names in the passenger volume and ticket sales data were matched to the database with fuzzy matching, automatically at 90% similarity or higher (about 95% of stations) and manually for the rest. Entries with incorrect stations or negative values were removed, and entries were aggregated to the number of passengers or tickets per OD pair and date. For air travel, flight times were corrected for time zones and aggregated per OD pair, using frequency weighted averages.

Because the passenger volume data contains no fares except for Eurostar, the fare of the nearest date with a ticket sold on the same OD pair was used, assuming that the fare from A to B equals that from B to A. For Eurostar, the NSI fare was verified to be representative of the overall fare. Eurostar is only available per month and is therefore excluded from the daily passenger volume dataset and included in a separate monthly dataset.

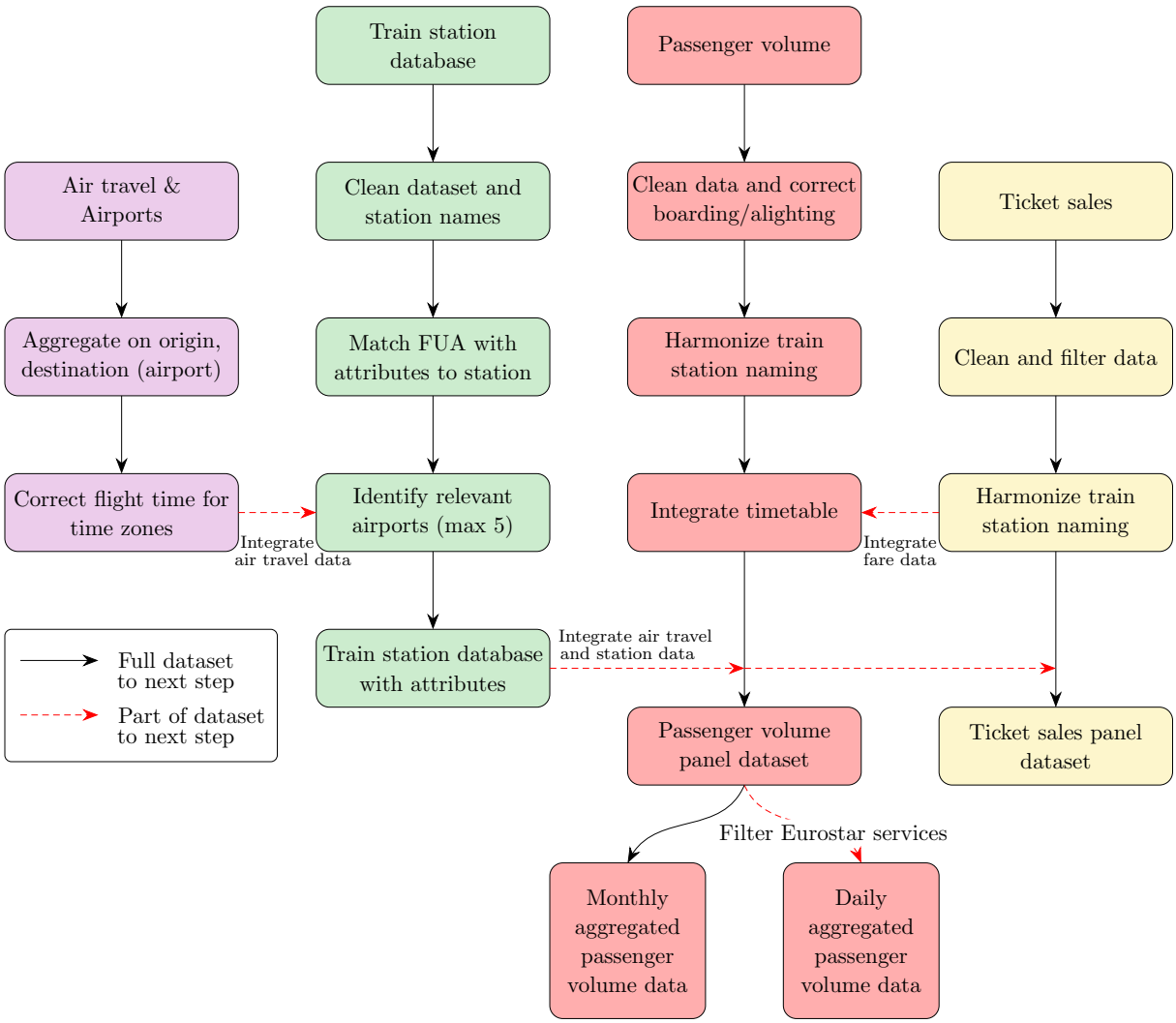


Figure 4 **Data processing steps**

4.3. Processed datasets

The size of both processed datasets is shown in Table 2 and 3. The variables included in the datasets are shown in Table 5. Not all are available in both datasets: the ticket sales data has no frequency or seat capacity, since the route taken is unknown, while the passenger volume data has no transfers, since all services are direct.

The data has several characteristics that are relevant for interpreting the results. Passenger volume is dominated by a few OD pairs, with Eurostar Paris accounting for more than half of the total volume. Stations outside a FUA (excluded) account for 2.10% of the passenger volume and 7.37% of the ticket sales, and 7.07% and 28.77% respectively have no air alternative. NSI sells 25.4% of all tickets on OD pairs longer than 300 km where passenger volume is also known. This share is used in the forecasting approach and is probably an overestimate, since the passenger volume data only contains direct services from The Netherlands, which NSI sells the most.

The variation in the data is limited. Rail supply variables only vary at few specific moments in time in the passenger volume data, which is shown in Table 4. Population shows very little variation over time,

and GDP is not available for the most recent years. Only rail fare, rail travel time, transfers and air fare show more varition. Distance and rail travel time are also strongly collinear (VIF of 29.71 and 30.37 in the passenger volume data, 9.07 and 7.22 in the ticket sales data), so their individual coefficients cannot be reliably separated. Air travel time was excluded since it is perfectly correlated with distance.

Other limitations affect the quality of individual variables. Fares are averages per OD pair and date, so fare class and booking time are not reflected, and fares matched to the passenger volume data may come from a different date. Construction works and reroutings are not included in the timetable, and Eurostar is only available per month. There is no passenger volume data on services to Belgium, where NSI also operates Eurocity services, and the number of train services with passenger volume data is limited.

Observation period	July 2023 - December 2025
Number of observations	106393
Total passenger volume	*Removed for confidentiality*
Number of train services	6
Number of origins	52
Number of destinations	52

Table 2 Passenger volume data characteristics

Observation period	July 2023 - April 2026
Number of observations	1,309,729
Total number of tickets	*Removed for confidentiality*
Number of origins	395
Number of destinations	5660

Table 3 Ticket sales data characteristics

Train service	**Variable change**	**Time period**	**Change**
IC(E) Berlin	Travel time reduction & stops removed	From December 2023	33 minutes reduction
IC(E) Berlin	Capacity decrease & rolling stock change	From September 2025	Seat capacity reduction of 148 seats
ICE Frankfurt	Travel time increase & stops changed	April, July, August and October 2025	24 minutes increase and routing via Venlo
ICE Basel/Munich	Destination change	Until July 2024 and from December 2024	Destination change
Eurostar Sun	Termination of service	Until 2024	Service terminated

Table 4 Temporal variation of rail service supply

Variable	Description	Resolution	Notation	Unit
Demand measurements				
Passenger volume	Number of passengers between OD pair during a certain time period	OD daily/monthly	V_{ijt}	passengers
Ticket sales volume	Number of tickets sold between specific OD pair on a specific date	OD daily	T_{ijt}	tickets
Origin-destination characteristics				
Population	Number of inhabitants per FUA in a specific year	Yearly per FUA	pop_{it}	inhabitants
GDP per capita	Gross domestic product per capita per FUA in a specific year	Yearly per FUA	gdp_{it}	€
Distance	Geographical distance between every station pair	OD pair	D_{ij}	km
Rail service attributes				
Rail fare	Average price paid for train ticket between an OD pair on a specific date	OD daily	$f_{rail,ijt}$	€
Rail travel time	Travel time between OD pair including transfers (if applicable) on a specific date	OD pair	$TT_{rail,ijt}$	minutes
Rail frequency	Number of direct trains between an OD pair on a specific day	OD pair	$freq_{rail,ijt}$*	trains/day
Seat capacity	Number of seats available (all fare classes combined) on a specific train between OD pair	OD daily	cap_{ijt}*	seats/train
Transfers	Number of transfers in itinerary on ticket sold on specific OD pair and day	OD daily	TR_{ijt}**	Number of transfers
Air travel characteristics				
Air frequency	Average number of flights between OD airports	OD pair	$freq_{air,ij}$	flights/day
Air fare	Average monthly fare between OD airport pairs, for all included carriers	OD monthly	$f_{air,ijt}$	€

Table 5 **Overview of available variables. *: only in passenger volume models, **: only in ticket sales models**

5. Results

Table 6, Table 7 and Table 8 show the estimated elasticities for the daily passenger volume data (excluding Eurostar), the monthly passenger volume data (including Eurostar) and the ticket sales data. For the service dummies the ICE Berlin is the baseline. Variables that do not vary over time are absorbed by the fixed-effects and are therefore not estimated in those specifications.

	Pooled panel model		OD fixed-effects		OD time fixed-effects	
Variable	**Parameter**	**p-value**	**Parameter**	**p-value**	**Parameter**	**p-value**
Model constant						
α	-65.639**	0.0000	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	1.4983*	0.0199	-	-	-	-
$\delta_{ICE,Basel}$	15.327	0.1784	-	-	-	-
$\delta_{ICE,Munich}$	15.109	0.1836	-	-	-	-
$\delta_{Eurostar,Paris}$	-	-	-	-	-	-
$\delta_{Eurostar,Seasonal}$	-	-	-	-	-	-
OD characteristics						
$\beta_{pop,i}$	0.8462**	0.0000	2.0106	0.7663	-7.6397	0.6665
$\beta_{pop,j}$	0.8625**	0.0000	-2.9703	0.7031	-11.307	0.5221
$\beta_{gdp,i}$	1.4642*	0.0496	-	-	-	-
$\beta_{gdp,j}$	0.9409	0.1738	-	-	-	-
β_D	-0.0128	0.9923	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	0.0514	0.3734	-0.0086	0.5271	-0.0422**	0.0035
$\beta_{TT_{rail}}$	-0.3104	0.8439	1.2398	0.0638	1.1592	0.1335
$\beta_{freq_{rail}}$	8.6249	0.1562	0.9053**	0.0000	0.9091**	0.0000
β_{cap}	-0.0049	0.9849	-0.3582*	0.0181	-0.8543**	0.0000
β_{TR}	-	-	-	-	-	-
Air travel characteristics						
$\beta_{freq_{air}}$	0.1794	0.5823	-	-	-	-
$\beta_{f_{air}}$	0.5419*	0.0158	0.2655**	0.0007	-0.0847	0.3980
Model statistics						
Unit of analysis	OD-day passengers		OD-day passengers		OD-day passengers	
Observations	38826		38826		38826	
R^2	0.4466		0.2803		0.3121	
Within R^2	0.2670		0.2803		0.2770	

Table 6 **Results for daily passenger volume data, excluding Eurostar (**: $p \leq 0.01$, *: $0.01 < p \leq 0.05$)**

Variable	Pooled panel model Parameter	p-value	OD fixed-effects Parameter	p-value	OD time fixed-effects Parameter	p-value
Model constant						
α	-44.862**	0.0005	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	0.4868	0.4644	-	-	-	-
$\delta_{ICE,Basel}$	-0.2947	0.6665	-	-	-	-
$\delta_{ICE,Munich}$	1.2270*	0.0437	-	-	-	-
$\delta_{Eurostar,Paris}$	-0.0959	0.8750	-	-	-	-
$\delta_{Eurostar,Seasonal}$	3.9962**	0.0000	-	-	-	-
OD characteristics						
$\beta_{pop,i}$	0.8558**	0.0000	49.480**	0.0001	-14.193	0.6552
$\beta_{pop,j}$	0.6787**	0.0002	42.762**	0.0006	-19.691	0.5586
$\beta_{gdp,i}$	1.2683	0.0917	-	-	-	-
$\beta_{gdp,j}$	1.2093	0.1035	-	-	-	-
β_D	-2.2709	0.0816	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	0.1033	0.8069	-0.5557*	0.0338	-0.5156	0.0745
$\beta_{TT_{rail}}$	1.4852	0.1002	3.2648**	0.0004	3.2776*	0.0289
$\beta_{freq_{rail}}$	1.2371**	0.0000	1.0512**	0.0000	1.1312**	0.0000
β_{cap}	0.0896	0.8009	-0.2731	0.1514	-0.3767	0.0814
β_{TR}	-	-	-	-	-	-
Air travel characteristics						
$\beta_{freq_{air}}$	0.8963	0.1182	-	-	-	-
$\beta_{f_{air}}$	0.4625	0.1757	0.3567**	0.0071	-0.1012	0.5369
Model statistics						
Unit of analysis	OD-month passengers		OD-month passengers		OD-month passengers	
Observations	2216		2216		2216	
R^2	0.5923		0.0598		0.0639	
Within R^2	0.1228		0.0598		0.0222	

Table 7 Results for monthly passenger volume data, including Eurostar (: $p \leq 0.01$, *: $0.01 < p \leq 0.05$)**

Variable	Pooled panel model		OD fixed-effects		OD time fixed-effects	
	Parameter	p-value	Parameter	p-value	Parameter	p-value
Model constant						
α	-3.9955	0.1698	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	-	-	-	-	-	-
$\delta_{ICE,Basel}$	-	-	-	-	-	-
$\delta_{ICE,Munich}$	-	-	-	-	-	-
$\delta_{Eurostar,Paris}$	-	-	-	-	-	-
$\delta_{Eurostar,Seasonal}$	-	-	-	-	-	-
OD characteristics						
β_{pop_i}	0.0166	0.8127	1.6146*	0.0390	2.6731	0.1375
β_{pop_j}	0.0843**	0.0076	0.0113**	0.0000	0.0034**	0.0001
β_{gdp_i}	0.1711	0.5551	-	-	-	-
β_{gdp_j}	0.3670**	0.0003	-	-	-	-
β_D	0.2875*	0.0132	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	-0.0253	0.3294	-0.1512**	0.0000	-0.1479**	0.0000
$\beta_{TT_{rail}}$	-0.3684*	0.0152	-0.0299	0.3836	-0.0414	0.2414
$\beta_{freq_{rail,ij}}$	-	-	-	-	-	-
β_{cap}	-	-	-	-	-	-
β_{TR}	-0.3755**	0.0000	-0.0898**	0.0000	-0.1013**	0.0000
Air travel characteristics						
$\beta_{freq_{air}}$	0.0927	0.1471	-	-	-	-
$\beta_{f_{air}}$	-0.3093**	0.0003	0.2496**	0.0000	0.0706**	0.0002
Model statistics						
Unit of analysis	OD-day tickets		OD-day tickets		OD-day tickets	
Observations	511743		511743		511743	
R^2	0.1565		0.0223		0.0226	
Within R^2	-0.1041		0.0223		0.0208	

Table 8 Results for ticket sales data (**: $p \leq 0.01$, *: $0.01 < p \leq 0.05$)

5.1. Passenger volume

In the pooled models the population elasticities are significant and close to one (0.85 and 0.86 in the daily data, 0.86 and 0.68 in the monthly data), in agreement with H1. In the fixed-effects models they are insignificant or implausibly large (up to 49.5), because population varies very little within OD pairs. Origin GDP per capita is significant in the daily pooled model (1.46, above the range of H2), and distance and air frequency are not significant in any model. Rail frequency is positive and significant in all fixed-effects models of both datasets (0.91 to 1.13), with the expected sign but above the range of H6. Rail fare is negative and significant in the daily two-way model (-0.04) and the monthly OD fixed-effects model (-0.56), where only the latter falls within the range of H4. Rail travel time is not negative in any passenger volume model and significantly positive in the monthly fixed-effects models (3.26 and 3.28), in contrast to H5. Seat capacity is significantly negative in the daily fixed-effects models (-0.36 and -0.85), in contrast to H7. Air fare is positive and significant in the OD fixed-effects models (0.27 daily, 0.36 monthly), in agreeement with H10, but not once time effects are added. The overall fit of the fixed-effects models is low for the monthly data (R^2 of 0.06), and somewhat higher for the daily data (within R^2 of 0.28).

5.2. Ticket sales

Rail fare is negative and significant in both fixed-effects models (-0.15), in agreement with H4 but smaller in magnitude than expected. Rail travel time is only significant in the pooled model (-0.37, within the range

of H5), and the number of transfers is negative and significant in all models (-0.38, -0.09 and -0.10), in agreement with H8. Air fare changes sign: negative in the pooled model (-0.31, contrary to H10), and positive and significant in the fixed-effects models (0.25 and 0.07). In the pooled model destination GDP per capita (0.37) and distance (0.29) are significant and within the ranges of H2 and H3, while the population elasticity is only significant for the destination, and small (0.08). The model fit is low: R^2 is 0.02 for the fixed-effects models, and the within R^2 of the pooled model is negative.

5.3. Comparison of datasets

The ticket sales data gives the more reliable and plausible estimates, with more significant coefficients than the passenger volume data. Rail frequency is the most stable coefficient across the passenger volume models, and air fare in the OD fixed-effects models is the only coefficient that is significant with the same sign in all three datasets. The fixed-effects estimates for the passenger volume data (population, travel time and seat capacity) are implausible, while the ticket sales data gives plausible estimates for rail fare, transfers and air fare.

6. Case study

This section applies the forecasting approach from subsection 3.6 to four candidate corridors originating in Amsterdam. The pooled OLS model based on the ticket sales data is used, since its estimated coefficients are considered more plausible than those of the passenger volume models (see section 7). The demand is forecast using Equation 4, where the Duan smearing factor $\hat{\phi}$ is 2.25 and the ticket sales correction factor δ is 0.25.

6.1. Candidate corridors

Four corridors were selected from candidates proposed by NSI and adjusted to fit the requirements in subsection 3.6. One high-speed-rail (HSR) corridor with two branches, to Barcelona (1A) and to Milan (1B), and two conventional intercity corridors, to Copenhagen (2) and to Prague (3). The baseline travel times and fares are shown in Table 9 and Table 10. Stations in the same FUA as another station (such as Schiphol in the same FUA as Amsterdam Centraal) or outside any FUA (such as Narbonne) were excluded. All corridors are assumed to receive one train per day, however this does does not influence the estimated demand since frequency is not in the ticket sales model.

Table 9 **Corridors 1A and 1B: travel time and fare from Amsterdam**

Corridor 1A: Barcelona			**Corridor 1B: Milan**		
Station	**Time**	**Fare**	**Station**	**Time**	**Fare**
Amsterdam Centraal	0:00				
Rotterdam Centraal	0:50	€49			
Antwerpen Centraal	1:25	€49			
Bruxelles Midi	2:01	€49			
Lille Europe	2:48	€49			
Marne-la-Vallée Chessy	4:04	€59			
Lyon Part-Dieu	5:27	€69	As 1A up to Lyon Part-Dieu		
Valence TGV	6:03	€79	Chambéry	7:03	€79
Nîmes	6:51	€79	Torino Porta Susa	9:50	€89
Montpellier-Saint-Roch	7:20	€79	Milano Centrale	10:45	€99
Perpignan	8:50	€79			
Girona	9:33	€89			
Barcelona Sants	10:15	€99			

Table 10 Baseline corridors 2 and 3: travel time and fare from Amsterdam

Corridor 2: Copenhagen			**Corridor 3: Prague**		
Station	**Time**	**Fare**	**Station**	**Time**	**Fare**
Amsterdam Centraal	0:00	-	Amsterdam Centraal	0:00	-
Amersfoort Centraal	0:33	€29	Amersfoort Centraal	0:36	€29
Deventer	1:19	€29	Deventer	1:18	€29
Osnabrück Hbf	2:49	€29	Osnabrück Hbf	2:54	€29
Bremen Hbf	3:51	€39	Hannover Hbf	4:04	€39
Hamburg Hbf	4:53	€39	Berlin Hbf	5:42	€45
Odense	8:27	€69	Dresden Hbf	7:24	€69
København	9:38	€79	Praha hl.n.	9:39	€79

6.2. Features and scenarios

The explanatory variables are constructed in the same way as for the model estimation, with two exceptions. First, a few stations on the corridors have no air alternative, so for these the pooled model was estimated again without the air travel variables. This changed the estimated demand by less than 0.5 passengers per OD pair and thus is considered acceptable. Second, GDP per capita was missing for four stations in FUAs below 250,000 inhabitants, and was taken manually from another dataset of the same source (OECD a). Travel times between non-consecutive stations are cumulated, and fares, which were only available from each corridor first station, were interpolated.

For each corridor alternative scenario(s) is/are analysed with the same model. The fare is doubled for corridors 1A and 3, and for corridor 1A a scenario with a transfer in Paris and the same travel times is also estimated. For corridor 1B the Lyon-Turin base tunnel reduces travel time by 2 hours beyond the tunnel and removes Chambéry as a stop (Tur). For corridor 2 the Fehmarnbelttunnel reduces travel time between Hamburg and Copenhagen by 2 hours and removes Odense (Why).

6.3. Forecast results

Table 11 and Figure 5 summarize the baseline forecast. In all corridors the largest cities generate the most demand, and the middle sections are the busiest. Corridor 1A has the highest total demand, but also the most stops. Corridor 1B has the highest average demand per OD pair and demand between the endpoints.

Corridor	**Total demand (total of both directions)**	**Stops**	**Average demand per OD pair**	**Demand Amsterdam to endpoint**
Corridor 1A (HSR Barcelona)	*Removed for confidentiality*	14		
Corridor 1B (HSR Milan)	*Removed for confidentiality*	10		
Corridor 2 (Copenhagen)	*Removed for confidentiality*	8		
Corridor 3 (Prague)	*Removed for confidentiality*	8		

Table 11 Estimated demand for corridors, all demand forecast measurements are in estimated passengers/day

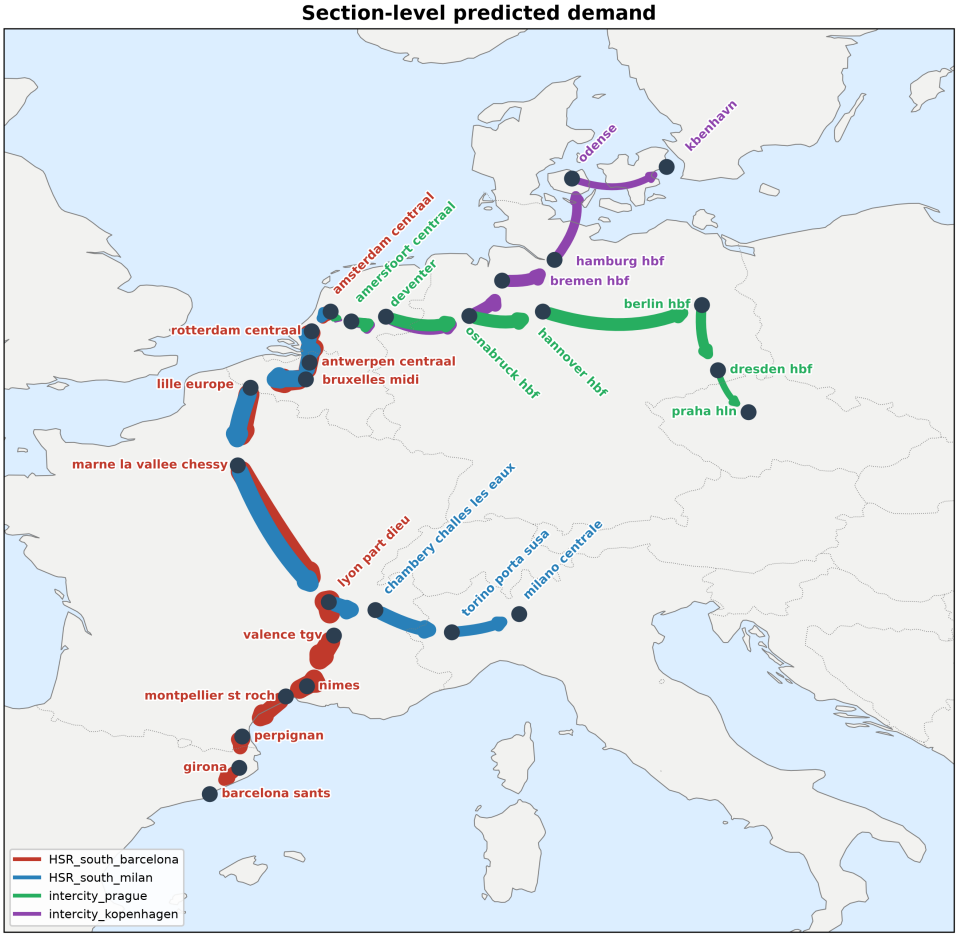
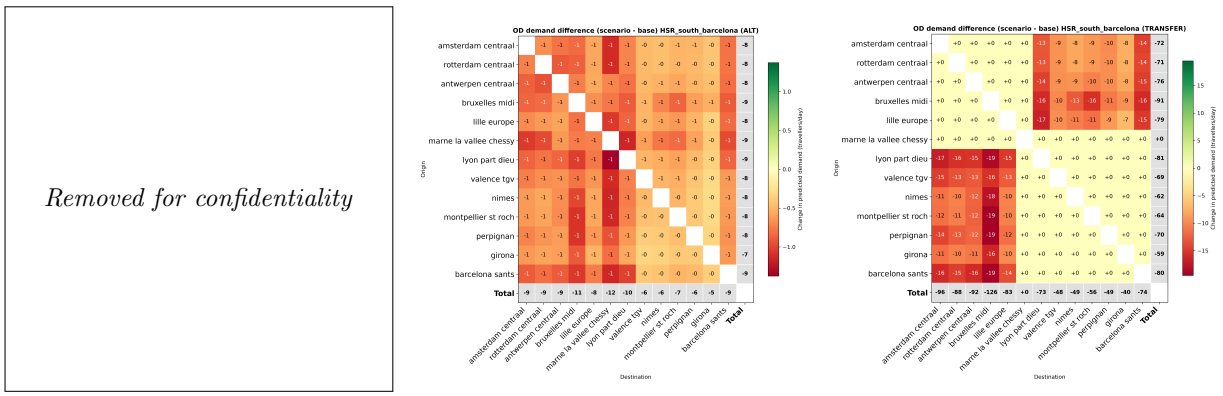


Figure 5 **Corridors and estimated demand loads in the forward direction (away from Amsterdam). Line thickness shows the load on each section**

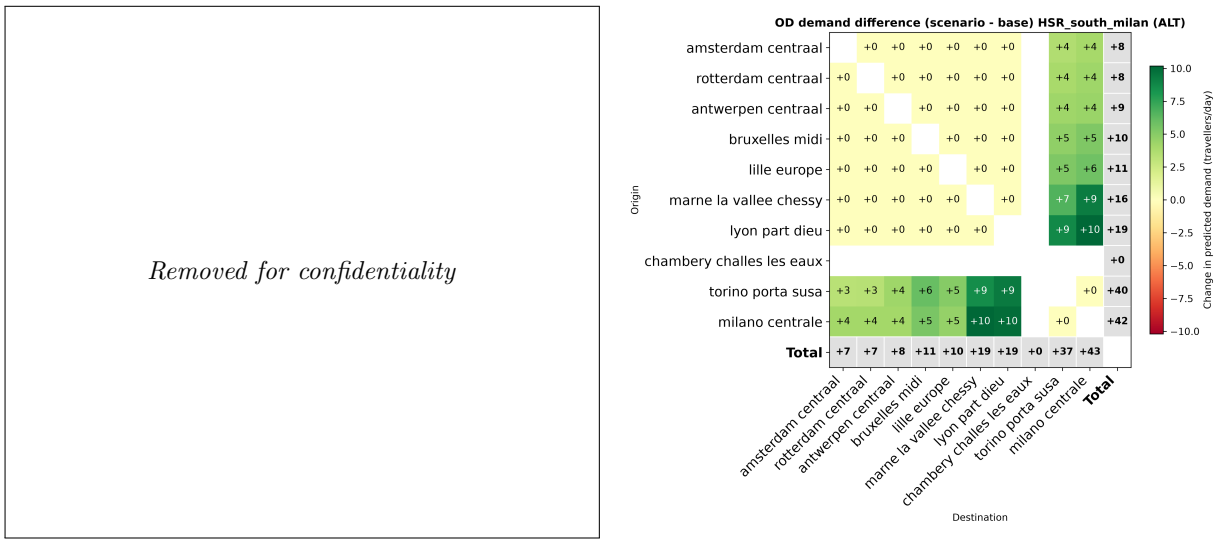
Figure 6 shows the OD matrices of corridor 1A. The largest demand is between Marne-la-Vallée Chessy (in the Paris FUA) and Lyon Part-Dieu, and the stops up to Lyon generate considerably more demand than those between Lyon and Barcelona. Doubling the fare reduces the total demand very slightly by 105 passengers per day (1.7%), with a change of 0 or 1 passengers per OD pair. The transfer in Paris has a larger effect, total demand falls by 873 passengers per day (14%), and for Amsterdam Centraal to Barcelona Sants from 46 to 32 passengers per day, which is a reduction of roughly 30%.



(a) Baseline (b) Difference, fare doubling (c) Difference, transfer in Paris

Figure 6 OD matrices of estimated demand (passengers/day) for corridor 1A

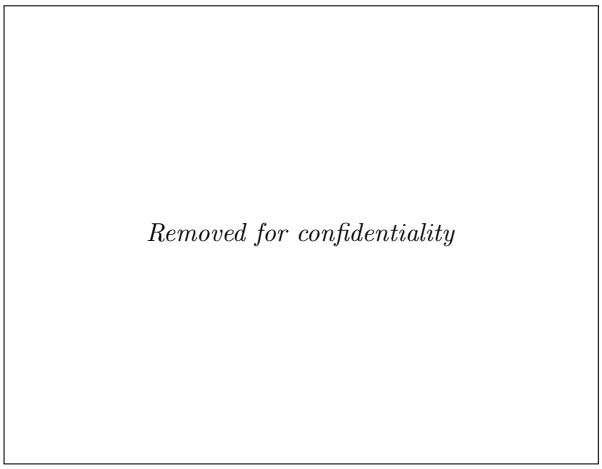
Figure 7 shows the estimated OD matrices of corridor 1B. The OD pairs up to Lyon are the same as for corridor 1A, and Turin and Milan generate considerable demand from all other origins. After a reduction of travel time by 2 hours hours through the opening of the Lyon-Turin base tunnel the model predicts a substantial increase in demand for Turin and Milan, largest on the OD pairs where the relative reduction in travel time is the largest.



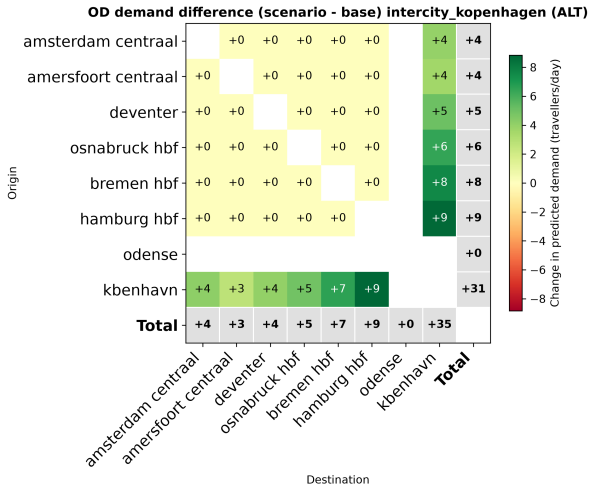
(a) Baseline (b) Difference 2 hour travel time reduction

Figure 7 OD matrices of estimated demand (passengers/day) for corridor 1B

Figure 8 shows the OD matrices of corridor 2, where Amsterdam, Hamburg and Copenhagen dominate the demand. The total demand is lower than for corridors 1A and 1B. Under the alternative scenario with a travel time reduction of 2 hours due to the opening of the Fehmarnbelttunnel. Only the demand to Copenhagen changes, as it is the only station that benefits from the reduced travel time.



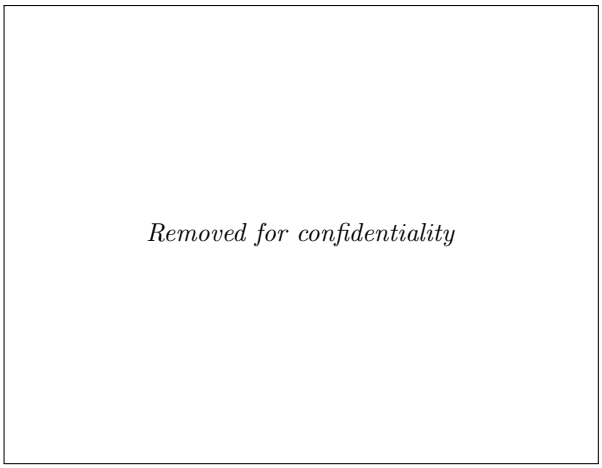
(a) Baseline



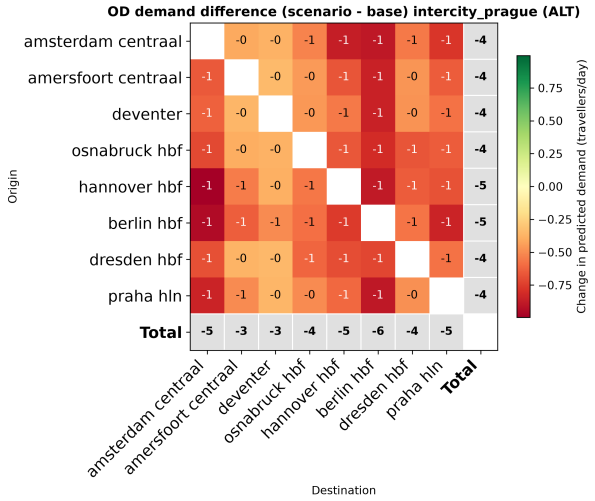
(b) Difference 2 hour travel time reduction

Figure 8 **OD matrices of estimated demand (passengers/day) for corridor 2**

Figure 9 shows the OD matrices of corridor 3, where Amsterdam to Berlin is the busiest OD pair. Doubling the fare changes again shows a very minimal effect which with demand changing by at most 1 passenger on any OD pair.



(a) Baseline



(b) Difference, fare doubling

Figure 9 **OD matrices of estimated demand (passengers/day) for corridor 3**

6.4. Uncertainty and corridor ranking

The forecast should be interpreted as an indication of how the model responds to corridors with different characteristics, rather than precise estimates of demand. The estimated demand is mostly driven by the population and GDP per capita of the FUAs. Distance increases demand while travel time reduces it, so even travel times of over 10 hours do not reduce demand strongly, and the rail fare and air travel variables have a marginal effect. The exception is the transfer, where the demand reduction is substantial and thus

showing the value of a direct service. Since demand is predicted for every OD pair separately, a corridor with more stops always gets more total demand, which is partially why 1A has much more demand than corridor 2 with similar travel times and city sizes.

Several sources of uncertainty remain. The correction for NSI's share of ticket sales is based on a limited set of OD pairs and is likely an overestimate of that share for these corridors, and the smearing factor could not be checked, since there is no observed demand on the corridors. Furthermore the use of the pooled model retains insignificant coefficients, has a low model fit and does not control for unobserved OD heterogeneity, and the fixed-effects model show that a number of variables do change sign due to the use of fixed-effects. Finally, all corridors have travel times close to the maximum in the data. The corridors are therefore ranked on average demand per OD pair and demand between the endpoints, which are less sensitive to the number of stops than total demand. This gives the ranking corridor 1B, 1A, 3 and 2. The ranking only reflects the demand estimated by this model and excludes the effects of the scenarios.

7. Discussion

7.1. Interpretation of the results

Comparing the full set of results across models and datasets, the ticket sales results seem most plausible. They are the closest to the formulated hypotheses, and most variables are statistically significant. However, most coefficient magnitudes are smaller than commonly found in the literature and than formulated in the hypotheses (see subsection 3.1). A likely explanation lies in the nature of the ticket sales data itself, as it contains only tickets sold through NSI, likely mostly representing Dutch travellers rather than the full market of travellers on a given OD pair. The share of travellers booking through NSI is unlikely to be consistent across OD pairs, likely declining the further a destination lies from The Netherlands. If this share itself also responds to explanatory variable changes, the estimated elasticities could reflect changes in the NSI booking share instead or in addition to changes to underlying rail demand.

The results from both volume datasets should be interpreted carefully. Both datasets cover a limited number of services, with substantial differences in demand levels between them, which makes the pooled estimates sensitive to a small number of high-volume services. Furthermore, the limited within-OD variation means the fixed-effects models have little variation to identify effects from. This explains the very large population coefficients, the positive travel time coefficients, which should not be over-interpreted. Together, these factors suggest the volume dataset results should not be considered very plausible, with the ticket sales results providing a better basis for interpretation.

Fixed-effects is expected to yield more reliable estimates because it absorbs unobserved OD-level heterogeneity, which is likely correlated with regressors such as fare and travel time. This is illustrated by the rail fare elasticity estimated on the daily volume data, which is insignificant in the pooled OLS and OD fixed-effects models, but significant in the OD+time fixed-effects model. In the ticket sales model the coefficient becomes more negative and remains significant. This pattern is consistent with endogeneity: fares may be correlated with unobserved OD characteristics, for example higher fares on high-speed or higher quality services that independently attract more demand. Pooled OLS remains necessary, as the time-invariant variables are absorbed under fixed-effects, and only the pooled model can be directly applied to forecast unobserved OD pairs.

Table 12 **Hypotheses and most plausible findings, reported as elasticities. V = passenger volume model, T = ticket sales model, OLS = pooled OLS, TWFE = two-way fixed-effects (OD + time), *: $0.01 < p \leq 0.05$, **: $p \leq 0.01$**

Variable	Hypothesis	Expected	Found
Origin-destination characteristics			
Origin population	H1	+0.1 to 4.0	0.85** (V-OLS)
Destination population	H1	+0.1 to 4.0	0.68** (V-OLS)
Origin GDP per capita	H2	+0.3 to 1.0	0.17 (T-OLS)
Destination GDP per capita	H2	+0.3 to 1.0	0.37** (T-OLS)
Distance	H3	−0.5 to +0.5	−0.01 (V-OLS)
Rail service supply attributes			
Rail fare	H4	−0.4 to −0.8	−0.148** (T-TWFE)
Rail travel time	H5	−0.3 to −0.8	−0.37* (T-OLS)
Rail frequency	H6	+0.1 to 0.5	0.9** (V-TWFE)
Seat capacity	H7	+ (no magnitude)	−0.85** (V-TWFE), not recommended
Transfers	H8	− (no magnitude)	−0.38** (T-OLS)
Air travel characteristics			
Air frequency	H9	−0.1 to −0.5	0.09 (T-OLS)
Air fare	H10	+0.1 to 0.7	0.07** (T-TWFE)

7.2. Comparison to literature

Table 12 summarizes the hypotheses and the most plausible estimate for every variable across all models and datasets. In some cases this is an insignificant result (such as for distance). Since the results come from different models they should not be combined in a forecast, as the coefficients also depend on the other regressors in their respective model.

The OD characteristics are broadly consistent with the literature and hypotheses H1 and H2. Population and GDP per capita turned out to be important drivers of demand, which was also found in the literature (Wijeweera et al. 2014, Wardman and Batley 2022, Wardman et al. 2018), and multicollinearity between them was observed, an issue also identified by Wardman (2006). The rail service attributes from the ticket sales model are mostly consistent with the literature, while those from the volume data are not. The ticket sales fare elasticity matches the expected sign but is smaller in magnitude than typical literature estimates, plausibly reflecting that ticket sales capture only the NSI booking share rather than total demand. The number of transfers has the expected negative sign in all specifications, and the magnitude of rail frequency is much larger than reported in the literature (Kristoffersson and Liu 2024, Paulley et al. 2006) and than formulated in hypothesis H6. Air frequency is never significant, while air fare does show a mode substitution effect in the fixed-effects specifications of the ticket sales model, as formulated in hypothesis H10, although the magnitude is smaller than typically found in the literature (Hortelano et al. 2016, Román et al. 2014, Wardman et al. 2018).

7.3. Case study

The case study relies on the pooled OLS model, as the fixed-effects models cannot produce demand estimates for OD pairs that are unobserved. The uncertainty is amplified by the requirement to include all coefficients, including the insignificant coefficients or coefficients with implausible signs (such as distance and air fare), which means the forecast is partially driven by relationships that are not considered plausible.

A further part of the forecasting methodology is that it predicts demand per OD pair rather than for a full corridor at once, since the model has no sense of multiple OD pairs belonging to a single train service. This means that when adding another stop, the demand always increases. This is partially mitigated by having a maximum of one stop per FUA, but will always result in higher total estimated demand when adding more stops.

Another limitation is that the methodology has no sense of the availability of existing train services, which means demand is both under- and over-predicted. For example, the Amsterdam Centraal-Berlin Hbf pair (part of corridor 3) is predicted to have around 51 passengers per day (see Figure 9), whereas the passenger volume dataset much more demand. This gap is too large to be explained by just the uncertainty in the correction factors. A likely explanation is that the model is estimated on all OD pairs in the ticket sales data, which includes both high- and low-volume OD pairs. The estimated relations are therefore averages across all these pairs and do not capture factors specific to established connections, such as passenger familiarity, brand recognition or tourism attractiveness. The model can also over-predict demand, for example for OD pairs connecting relatively small cities over very long travel times. Because the distance coefficient is positive and the travel time coefficient is negative, these two effects partially cancel each other out. An example is the Lille - Perpignan OD pair within corridor 1A, for which the model predicts 29 passengers per day, which is only half the predicted demand for Brussels - Barcelona, despite Lille and Perpignan being considerably smaller cities with a similarly long travel time and likely much less tourism attractiveness.

The absolute demand forecasts should therefore not be seen as the actual expected level of demand, since they are based on the partial share of the total demand and rely on two correction factors which amplify the uncertainty. The relative ranking is still considered useful as a first indication of which corridors are the most promising. Both correction factors scale the forecasts of all corridors in the same way and therefore affect the level of demand rather than the order of the corridors. In addition, all corridors considered in the case study originate in The Netherlands, which is also the origin of the ticket sales data, and are assumed to be subject to similar biases.

7.4. Limitations and implications

This research is reliant on the quality of the data. The passenger volume data is limited to a small number of services with very unevenly distributed volumes, and the ticket sales data represents only tickets sold by NSI, which is biased toward OD pairs reachable via direct services from The Netherlands and toward travellers who are familiar with NS. As a result, the elasticity estimates likely reflect NSI booking behaviour instead of total rail demand, and they should not be assumed to generalize to other operators, countries, or rail markets. Future research would benefit a lot from a more complete demand dataset covering a large, balanced set of OD pairs across multiple countries and multiple years, and capturing full passenger volumes. In practice, no single operator holds this kind of dataset, as rail operators generally do not share commercially sensitive data with each other. Overcoming this would require data-sharing agreements between operators or a centralised data source, such as one coordinated by a European regulatory body.

Endogeneity is also a problem in the applied methodology. Omitted variables such as tourism attractiveness or weather are not absorbed in the pooled OLS model, and rail fare is assumed to affect demand while operators also use estimated demand to set fares. This can be tested and mitigated with an instrumental variable (Wooldridge 2016), for example track access charges set by infrastructure managers for rail fare, but this needs sufficient variation across OD pairs and over time. The methodology also did not examine heterogeneity, while elasticities are known to differ between business and leisure travellers (Brons et al.

2002, Wardman and Batley 2022), and possibly across regions and distance bands, which would require a substantially larger and more evenly spread dataset.

In this model demand is found to be driven mostly by population and GDP per capita, therefore NSI should treat market size as a first indication when evaluating candidate corridors, using rail and air service characteristics as secondary considerations. The value of a direct service is clear from the estimated demand with and without a transfer, even for the same travel times. Forecasts for new corridors should be treated as a starting point, as induced demand means realised demand may grow over time as the market develops. Finally, obtaining more complete demand data, for example through data sharing agreements with partner operators, would help the most and directly address several of the limitations identified in this research.

8. Conclusion

This research aimed to answer the following research question: *To what extent do origin-destination characteristics, rail service attributes, and competing air services influence international rail passenger demand, and how can these estimated relationships be used to forecast demand for new international rail services?*

Origin-destination characteristics, rail service attributes and competing air services all influence international rail passenger demand, but the strength and direction of these relationships could only partially be estimated, and only using the ticket sales dataset. The variation in the passenger volume data was not sufficient to identify demand responses: the within-OD variation in supply is very limited, with only five specific changes to the supply over the study period, and the number of OD pairs is too small. This resulted in implausible and inflated coefficients for the capacity, travel time and population variables, while many other variables remained insignificant. The ticket sales data provided a more even distribution of the explanatory variables and demand, and therefore resulted in more plausible results. The model fit remained low, which limits how much weight can be placed on these estimates, and only the significant and plausible results are reported here.

The most important drivers of demand are the population and GDP per capita of the destination, of which GDP per capita has the largest influence, with elasticities of 0.0843 ($p = 0.0076$) and 0.3670 ($p = 0.0003$) respectively. Rail fare has a negative elasticity of -0.1479 ($p = 0.0000$) in the two-way fixed-effects model, so higher rail fares reduce demand. Similarly, for rail travel time an elasticity of -0.3684 ($p = 0.0152$) was found in the pooled OLS model, so increased travel time reduces demand. Transfers also negatively influence demand, with an elasticity of -0.3755 ($p = 0.0000$) in the pooled OLS model and -0.1013 ($p = 0.0000$) in the two-way fixed-effects model. Air fare shows a small substitution effect between air and rail, with an elasticity of 0.0706 ($p = 0.0002$) in the two-way fixed-effects model.

The estimated relationships can be used to forecast demand for new international rail services using the developed methodology. Pooled OLS was the only suitable model, as it is the only model that can predict absolute demand levels for OD pairs that are currently unserved. By constructing the explanatory variables for each OD pair along a candidate corridor, using assumed service characteristics, demand can be predicted at the OD level and aggregated to the corridor level. Applied to four candidate corridors originating in Amsterdam (Barcelona (HSR), Milan (HSR), Copenhagen and Prague), the forecast showed similar demand levels per OD pair, with Amsterdam - Milan having the highest average demand per OD pair and Amsterdam - Barcelona having the highest total demand. The scenario analysis showed that improving travel time does substantially increase demand while increasing fare has very little influence. Introducing a transfer has a large impact, with demand reducing by around 30% in the scenario with a transfer for corridor 1A. This highlights that there is value in running a direct service compared to itineraries with transfers, even when the travel times are equal.

However, the accuracy of these forecasts is limited. It depends on the reliability of the estimated coefficients, which come from a pooled OLS model with a low fit that includes insignificant variables, and on the quality of the input data and the assumed service characteristics. Furthermore, the predicted tickets sold had to corrected to reach absolute demand levels, which introduces further uncertainty. The forecasting results should therefore be interpreted as a relative ranking of corridors rather than as absolute demand forecasts.

References

, a. OECD Data Explorer • Economy - FUAs. URL: https://data-explorer.oecd.org/vis?lc=en&df[ds]=dsDisseminateFinalDMZ&df[id]=DSD_FUA_ECO%40DF_ECONOMY&df[ag]=OECD.CFE.EDS&dq=AT001F%2BAT002F%2BAT003F%2BAT004F%2BAT005F%2BAT006F.A.LAB_PROD..&lom=LASTNPERIODS&lo=5&to[TIME_PERIOD]=false&vw=tb.

, b. OECD Definition of Cities and Functional Urban Areas. URL: https://www.oecd.org/en/data/datasets/oecd-definition-of-cities-and-functional-urban-areas.html.

, . Train Stations in Europe. URL: https://www.kaggle.com/datasets/headsortails/train-stations-in-europe.

, . Treinarchief. URL: https://www.rijdendetreinen.nl/treinarchief.

, . Turin-Lyon Base Tunnel. URL: https://www.fsitaliane.it/content/fsitaliane/en/strategic-projects/turin-lyon-base-tunnel.html.

, . Why we're building the Fehmarnbelt fixed link. URL: https://femern.com/why-we-re-building-the-fehmarnbelt-fixed-link/.

, 2016. Fourth railway package of 2016 - Mobility and Transport. URL: https://transport.ec.europa.eu/transport-modes/rail/railway-packages/fourth-railway-package-2016_en.

Allard, R.F., Moura, F., 2018. Effect of transport transfer quality on intercity passenger mode choice. Transportation Research Part A: Policy and Practice 109, 89–107. URL: https://www.sciencedirect.com/science/article/pii/S0965856415302718, doi:10.1016/j.tra.2018.01.018.

Avogadro, N., Redondi, R., 2023. Diverted and induced demand: Evidence from the London-Paris passenger market. Research in Transportation Economics 100, 101304. URL: https://www.sciencedirect.com/science/article/pii/S0739885923000446, doi:10.1016/j.retrec.2023.101304.

Baltagi, B.H., 2021. Econometric Analysis of Panel Data. Springer Texts in Business and Economics, Springer International Publishing, Cham. URL: https://link.springer.com/10.1007/978-3-030-53953-5, doi:10.1007/978-3-030-53953-5.

Banerjee, N., Morton, A., Akartunalı, K., 2020. Passenger demand forecasting in scheduled transportation. European Journal of Operational Research 286, 797–810. URL: https://www.sciencedirect.com/science/article/pii/S0377221719308677, doi:10.1016/j.ejor.2019.10.032.

Birolini, S., Cattaneo, M., Malighetti, P., Morlotti, C., 2020. Integrated origin-based demand modeling for air transportation. Transportation Research Part E: Logistics and Transportation Review 142, 102050. URL: https://www.sciencedirect.com/science/article/pii/S1366554520307018, doi:10.1016/j.tre.2020.102050.

Brons, M., Pels, E., Nijkamp, P., Rietveld, P., 2002. Price elasticities of demand for passenger air travel: a meta-analysis. Journal of Air Transport Management 8, 165–175. URL: https://www.sciencedirect.com/science/article/pii/S0969699701000503, doi:10.1016/S0969-6997(01)00050-3.

Börjesson, M., 2014. Forecasting demand for high speed rail. Transportation Research Part A: Policy and Practice 70, 81–92. URL: https://www.sciencedirect.com/science/article/pii/S0965856414002444, doi:10.1016/j.tra.2014.10.010.

Castillo-Manzano, J.I., Pozo-Barajas, R., Trapero, J.R., 2015. Measuring the substitution effects between High Speed Rail and air transport in Spain. Journal of Transport Geography 43, 59–65. URL: https://www.sciencedirect.com/science/article/pii/S0966692315000101, doi:10.1016/j.jtrangeo.2015.01.008.

Clewlow, R.R., Sussman, J.M., Balakrishnan, H., 2014. The impact of high-speed rail and low-cost carriers on European air passenger traffic. Transport Policy 33, 136–143. URL: https://www.sciencedirect.com/science/article/pii/S0967070X14000304, doi:10.1016/j.tranpol.2014.01.015.

Diaz-Gutierrez, J., Ranjbari, A., 2025. How effective are fixed-effects models in fixing the transit supply–demand bidirectional interaction? Transportation Letters 17, 1199–1212. URL: https://doi.org/10.1080/19427867.2024.2422713, doi:10.1080/19427867.2024.2422713. _eprint: https://doi.org/10.1080/19427867.2024.2422713.

Duan, N., 1983. Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association 78, 605–610. URL: https://www.jstor.org/stable/2288126, doi:10.2307/2288126.

Fale, M., Kramberger, T., Rupnik, B., Lixin, S., Vizinger, T., 2025. Evaluating Middle-Distance Passenger Transportation: A Comparison of the Time Efficiency, Sustainability, and Break-Even Distances Between High-Speed-Rail and Air Travel in Sweden. Sustainability 17, 386. URL: https://www.mdpi.com/2071-1050/17/2/386, doi:10.3390/su17020386.

Givoni, M., 2007. Environmental Benefits from Mode Substitution: Comparison of the Environmental Impact from Aircraft and High-Speed Train Operations. International Journal of Sustainable Transportation 1, 209–230. URL: https://doi.org/10.1080/15568310601060044, doi:10.1080/15568310601060044.

Gundelfinger-Casar, J., Coto-Millán, P., 2017. Intermodal competition between high-speed rail and air transport in Spain. Utilities Policy 47, 12–17. URL: https://www.sciencedirect.com/science/article/pii/S0957178716300157, doi:10.1016/j.jup.2017.06.001.

Holmgren, J., 2007. Meta-analysis of public transport demand. Transportation Research Part A: Policy and Practice 41, 1021–1035. URL: https://www.sciencedirect.com/science/article/pii/S096585640700047X, doi:10.1016/j.tra.2007.06.003.

Hortelano, A.O., Guzman, A.F., Preston, J., Vassallo, J.M., 2016. Price Elasticity of Demand on the High-Speed Rail Lines of Spain: Impact of the New Pricing Scheme. Transportation Research Record 2597, 90–98. URL: https://doi.org/10.3141/2597-12, doi:10.3141/2597-12.

Kampp, M., Sedelmeier, J., Schüth, J., Thust, M., Kaiser, D., Scherr, W., Schlaich, J., Senk, P., 2025. Design of a European high-speed rail network and use of passenger demand forecasting to test European policy targets. European Transport Research Review 17, 23. URL: https://doi.org/10.1186/s12544-025-00715-y, doi:10.1186/s12544-025-00715-y.

Kristoffersson, I., Liu, C., 2024. Estimation of demand models for long-distance international travel - Key determinants for Swedes' travel abroad. European Journal of Transport and Infrastructure Research 24, 62–88. URL: https://www.webofscience.com/api/gateway?GWVersion=2&SrcAuth=GetFTR&SrcApp=UA&DestURL=https%3A%2F%2Fct.prod.getft.io%2FY2xhcml2YXRlLHVuZGVmaW5lZCxNVEF1TlRrME9UQXZaV3AwYVhJdU1qQXlOQzR5TkM0MExqWTJPVFksWkRBNFpHRXlaV1V0WXpCbU5TMDBaB-dOOi_Z5ksKbYkCdseVD9kUyfkepZYgFMbdO5F8exs&DestApp=GetFTR&SrcItemId=WOS:001388740200004&SrcAppSID=EUW1ED0DCCuhQmUr6zw4V6CEbpqFJ&HMAC=oMUfMC2AUAQmYDhi%2FGw76UzpuqDbRcTZG5HBnXU7Rk8%3D, doi:10.59490/ejtir.2024.24.4.6696. num Pages: 27 Web of Science ID: WOS:001388740200004.

Lieshout, R., 2012. Measuring the size of an airport's catchment area. Journal of Transport Geography 25, 27–34. URL: https://www.sciencedirect.com/science/article/pii/S0966692312001706, doi:10.1016/j.jtrangeo.2012.07.004.

OECD, a. Economic indicators by territorial typology on access to functional urban areas. URL: https://data-explorer.oecd.org/vis?lc=en&df[ds]=dsDisseminateFinalDMZ&df[id]=DSD_REG_ECO%40DF_TYPE_METRO&df[ag]=OECD.CFE.EDS&df[vs]=1.0&dq=A..AUT%2BBEL%2BCHE%2BCHL%2BCOL%2BCZE%2BDEU%2BDNK%2BESP%2BEST%2BFIN%2BFRA%2BGBR%2BGRC%2BHUN%2BIRL%2BITA%2BJPN%2BKOR%2BLTU%2BLUX%2BLVA%2BNLD%2BNOR%2BNZL%2BPOL%2BPRT%2BSVK%2BSVN%2BSWE%2BTUR%2BUSA..GDP..Q.USD_PPP_PS&lom=LASTNPERIODS&lo=5&to[TIME_PERIOD]=false&vw=ov.

OECD, b. OECD Data Explorer • Population and built-up density - Cities and FUAs. URL: http://data-explorer.oecd.org/s/1e6.

Ortúzar, J.d.D., Willumsen, L.G., 2024. Modelling Transport. John Wiley & Sons. Google-Books-ID: kMn1EAAAQBAJ.

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J., White, P., 2006. The demand for public transport: The effects of fares, quality of service, income and car ownership. Transport Policy 13, 295–306. URL: https://www.sciencedirect.com/science/article/pii/S0967070X05001587, doi:10.1016/j.tranpol.2005.12.004.

Rich, J., Mabit, S.L., 2012. A Long-Distance Travel Demand Model for Europe. European Journal of Transport and Infrastructure Research 12. URL: https://journals.open.tudelft.nl/ejtir/article/view/2946, doi:10.18757/ejtir.2012.12.1.2946.

Román, C., Martín, J.C., Espino, R., Cherchi, E., Ortúzar, J.d.D., Rizzi, L.I., González, R.M., Amador, F.J., 2014. Valuation of travel time savings for intercity travel: The Madrid-Barcelona corridor. Transport Policy 36, 105–117. URL: https://www.sciencedirect.com/science/article/pii/S0967070X1400153X, doi:10.1016/j.tranpol.2014.07.007.

Tsekeris, T., Tsekeris, C., 2011. Demand Forecasting in Transport: Overview and Modeling Advances. Economic Research-Ekonomska Istraživanja 24, 82–94. URL: https://doi.org/10.1080/1331677X.2011.11517446, doi:10.1080/1331677X.2011.11517446. _eprint: https://doi.org/10.1080/1331677X.2011.11517446.

Wardman, M., 2006. Demand for rail travel and the effects of external factors. Transportation Research Part E: Logistics and Transportation Review 42, 129–148. URL: https://www.sciencedirect.com/science/article/pii/S1366554504000596, doi:10.1016/j.tre.2004.07.003.

Wardman, M., 2022a. Meta-analysis of British time-related demand elasticity evidence: An update. Transportation Research Part A: Policy and Practice 157, 198–214. URL: https://www.sciencedirect.com/science/article/pii/S0965856422000295, doi:10.1016/j.tra.2022.02.001.

Wardman, M., 2022b. Meta-analysis of price elasticities of travel demand in great britain: Update and extension. Transportation Research Part A: Policy and Practice 158, 1–18. URL: https://www.sciencedirect.com/science/article/pii/S0965856422000283, doi:10.1016/j.tra.2022.01.020.

Wardman, M., 2025. Investigating demand models with more flexible elasticity functions: empirical insights from rail demand analysis. Transportation 52, 1457–1483. URL: https://doi.org/10.1007/s11116-024-10462-z, doi:10.1007/s11116-024-10462-z.

Wardman, M., Batley, R., 2022. The demand impacts of train punctuality in great britain: systematic review, meta-analysis and some new econometric insights. Transportation 49, 555–589. URL: https://doi.org/10.1007/s11116-021-10186-4, doi:10.1007/s11116-021-10186-4.

Wardman, M., Lythgoe, W., Whelan, G., 2007. Rail Passenger Demand Forecasting: Cross-Sectional Models Revisited. Research in Transportation Economics 20, 119–152. URL: https://www.sciencedirect.com/science/article/pii/S0739885907200058, doi:10.1016/S0739-8859(07)20005-8.

Wardman, M., Toner, J., Fearnley, N., Flügel, S., Killi, M., 2018. Review and meta-analysis of inter-modal cross-elasticity evidence. Transportation Research Part A: Policy and Practice 118, 662–681. URL: https://www.sciencedirect.com/science/article/pii/S0965856418309364, doi:10.1016/j.tra.2018.10.002.

Wijeweera, A., To, H., Charles, M.B., Sloan, K., 2014. A time series analysis of passenger rail demand in major Australian cities. Economic Analysis and Policy 44, 301–309. URL: https://www.sciencedirect.com/science/article/pii/S0313592614000447, doi:10.1016/j.eap.2014.08.003.

Wong, H., Yap, M., 2023. A data driven approach to update public transport service elasticities. Journal of Public Transportation 25, 100066. URL: https://www.sciencedirect.com/science/article/pii/S1077291X23000279, doi:10.1016/j.jpubtr.2023.100066.

Wooldridge, J.M., 2016. Introductory econometrics: a modern approach. Sixth edition, student edition ed., Cengage Learning, Boston, Mass.

Yao, E., Morikawa, T., 2005. A study of an integrated intercity travel demand model. Transportation Research Part A: Policy and Practice 39, 367–381. URL: https://www.sciencedirect.com/science/article/pii/S0965856404001168, doi:10.1016/j.tra.2004.12.003.

C

AI statement

This thesis is written in a time where AI is changing and improving incredibly quickly. I think this brings up some difficult ethical dilemmas and requires careful considerations when using AI for scientific research like a master thesis. Currently the expectations for a master thesis have not changed (much) yet due to the rise of AI, which I think can result in some discrepancy in what is expected and what can be done by the student. It also makes it hard to judge if a student actually understands what he or she has done. Personally I was very surprised myself how well AI (especially Anthropic Claude) is able to interpret scientific literature and highlight any gaps or methodological considerations. Even during the time span of this thesis the quality of the output of AI agents changed considerably, and the difference between AI agents is also very clear. At the beginning I mostly used OpenAI ChatGPT, but I realized it is not able to think critically and has tendency to always say you are correct, even when this is not the case. I also experimented with ChatGPT to find literature, but sometimes it 'hallucinated' papers which did not exist which made me very sceptical that it was able to find relevant literature.

How I used AI for this thesis:

- Writing: I used AI to help draft some of the sections which I used as a starting point. Sometimes I also let AI improve the clarity of a section I wrote myself. I always carefully reviewed any text improved by AI so it still says what I want it to say. (Claude)
- Coding: For coding the help of AI was very useful, and it was the aspect for which I used AI the most. For the more complicated parts of coding I carefully considered what I wanted my code to do, then took the time to write a detailed prompt and review the code it generated (Claude). In VSCode (the IDE I used) I also had the integration of Github Copilot, which I did not write any prompts for but it did help with auto filling and the more repetitive tasks of coding.
- LaTeX: For the formatting of tables I used AI to format them nicely or translate a table from python to latex.
- Flowcharts: I used AI to improve the formatting of the flowcharts after I made them myself.
- Literature: I experimented with using AI to find literature but was not convinced that it was able to find all relevant literature so most literature was found using normal methods (Google scholar, scopus, snowballing etc). AI did help with filling the literature elasticity table which is used to summarize elasticities found in the literature used for the hypotheses (Claude).

D

Elasticity literature overview

Table D.1: Elasticity estimates and related findings from the literature.

Study	Type	Variable / finding	Altered	Affected	Magnitude	Region	Method	Form	FE	Distance	Run	Source	Notes
Wardman et al., 2018	Cross	Fare	Air	Rail	0.28	Intl. (meta-analysis)	Meta-regression (Model I)	Log-log	Study dummies (20)	Mixed	n.d.	Tab. 5	$n = 5$ studies
	Cross	In-vehicle time	Air	Rail	0.15	Intl. (meta-analysis)	Meta-regression (Model I)	Log-log	Study dummies (20)	Mixed	n.d.	Tab. 5	$n = 5$ studies
	-	SR:LR ratio (general)	-	All pairs	LR = 2.35× SR	Intl. (meta-analysis)	Meta-regression, LR vs. SR	Log-log	Study dummies (20)	-	Both	Sec. 4.3.4	-
Wardman and Batley, 2022	Own	Fare	Rail	Rail	−0.38 to −1.02 (by market)	UK (excl. London)	Direct demand, panel FE	Log-log	Flow FE (annual panel)	20-100 km	n.d.	Tab. 9	
	Own	Generalised journey time	Rail	Rail	−0.68 to −0.69	UK (excl. London)	Direct demand, panel FE	Log-log	Flow FE (annual panel)	20-100 km	n.d.	Tab. 9	Includes in-vehicle-time and weighted interchange penalty
	Own	GVA per capita	Rail	Rail	0.39 to 0.48	UK (excl. London)	Direct demand, panel FE	Log-log	Flow FE (annual panel)	20-100 km	n.d.	Tab. 9	Used as GDP proxy
Clewlow et al., 2014	Cross	In-vehicle time	Rail	Air	4.734	DE, FR, ES, IT, UK	Random-effects panel	Log-log	No (random effects)	Short-haul flights	n.d.	Tab. 2	Effect of rail on air demand (reverse direction)
Wardman, 2022b	Own	Fare	Rail	Rail	−0.87 to −1.11	UK	Meta-regression	Log-log	No	Urban and inter-urban (categorical)	Both	Tab. 6	Companion to Wardman et al., 2018
	-	SR:LR ratio, own-price (rail-specific)	-	-	LR = 1.86× SR	UK	-	-	-	-	Both	Tab. 5	-
Wardman, 2006	Own	GDP (income)	Rail	Rail	0.30 to 2.11	UK	Direct demand, ticket-sales panel	Log-log	No (cross-section of flows)	Long-distance and inter-urban	n.d.	Tab. 2, 3	Rises with distance
	Own	Fare	Rail	Rail	−0.49 to −0.83	UK	Direct demand, ticket-sales panel	Log-log	No (cross-section of flows)	Long-distance and inter-urban	n.d.	Tab. 2, 3	-
	Own	In-vehicle time	Rail	Rail	−0.61 to −1.08	UK	Direct demand, ticket-sales panel	Log-log	No (cross-section of flows)	Long-distance and inter-urban	n.d.	Tab. 2, 3	Useful source on multicollinearity
Brons et al., 2002	Own	Fare	Air	Air	−1.146	Intl. (meta-analysis)	Meta-analysis (37 studies)	Log-log	No (cross-study regression)	Short and long haul	n.d.	Tab. 2, 3	Air only; no cross-elasticities
	-	LR more elastic than SR	-	-	-	Intl. (meta-analysis)	Meta-analysis (37 studies)	Log-log	No (cross-study regression)	Short and long haul	-	-	-
Román et al., 2014	Own	Fare	HSR	HSR	−0.89 to −1.01	Spain (Madrid-Barcelona)	Discrete choice (RP data)	Point elast. of choice prob.	No (RP cross-section)	Single corridor	n.d.	Tab. 13	-
	Own	In-vehicle time	HSR	HSR	−0.49 to −0.83	Spain (Madrid-Barcelona)	Discrete choice (RP data)	Point elast. of choice prob.	No (RP cross-section)	Single corridor	n.d.	Tab. 13	-

continued on next page

Table D.1 (continued)

Study	Type	Variable / finding	Altered	Affected	Magnitude	Region	Method	Form	FE	Distance	Run	Source	Notes
	Own	Headway (frequency proxy)	HSR	HSR	−0.19 to −0.32	Spain (Madrid-Barcelona data)	Discrete choice (RP data)	Point elast. of choice prob.	No (RP cross-section)	Single corridor	n.d.	Tab. 13	-
	Cross	Fare	Air	HSR	0.53 to 0.67	Spain (Madrid-Barcelona data)	Discrete choice (RP data)	Point elast. of choice prob.	No (RP cross-section)	Single corridor	n.d.	Tab. 13	-
	Cross	In-vehicle time	Air	HSR	0.31 to 0.58	Spain (Madrid-Barcelona data)	Discrete choice (RP data)	Point elast. of choice prob.	No (RP cross-section)	Single corridor	n.d.	Tab. 13	-
	Cross	Headway (frequency proxy)	Air	HSR	0.12 to 0.21	Spain (Madrid-Barcelona data)	Discrete choice (RP data)	Point elast. of choice prob.	No (RP cross-section)	Single corridor	n.d.	Tab. 13	-
Kristoffersson and Liu, 2024	Own	Fare	Rail	Rail	−0.479	Sweden	Nested and multinomial logit	Point elast. of choice prob.	No (RP cross-section)	Long-distance international (Swedish residents)	n.d.	Tab. 5, 6	-
	Cross	Fare	Rail	Air	0.012	Sweden	Nested and multinomial logit	Point elast. of choice prob.	No (RP cross-section)	Long-distance international (Swedish residents)	n.d.	Tab. 5, 6	-
	Own	In-vehicle time	Rail	Rail	0.291 to 1.069	Sweden	Nested and multinomial logit	Point elast. of choice prob.	No (RP cross-section)	Long-distance international (Swedish residents)	n.d.	Tab. 5, 6	-
	Cross	In-vehicle time	Rail	Air	−0.008 to −0.043	Sweden	Nested and multinomial logit	Point elast. of choice prob.	No (RP cross-section)	Long-distance international (Swedish residents)	n.d.	Tab. 5, 6	-
	Own	Waiting time outside Sweden (frequency/interchange proxy)	Rail	Rail	0.302	Sweden	Nested and multinomial logit	Point elast. of choice prob.	No (RP cross-section)	Long-distance international (Swedish residents)	n.d.	Tab. 5, 6	-
Paulley et al., 2006	Own	Transfer penalty (not an elasticity)	Rail	Rail	37 min IVT-equivalent	Mostly UK	Meta-analysis (pooled/averaged values)	IVT-equivalent	No	Not specified; not long-distance	-	Sec. 3.7	Not an elasticity
	Own	Frequency (service level)	Rail	Rail	0.75	Mostly UK	Meta-analysis (pooled/averaged values)	Point elast.	No (synthesis)	Not specified; urban/regional	SR	Tab. 4	Only 3 studies
	Own	Income (GDP proxy)	Rail	Rail	Up to 2.0	Mostly UK	Meta-analysis (pooled/averaged values)	Point elast.	No (synthesis)	Not specified; mostly regional	LR	Sec. 5.3	-
	Own	In-vehicle time	Rail	Rail	−0.4 to −1.6	Mostly UK	Meta-analysis (pooled/averaged values)	Point elast.	No (synthesis)	Long-distance inter-urban to short urban	n.d.	Sec. 3.4	-
	Cross	Fare	Long-distance bus	Rail	0.17	Mostly UK	Meta-analysis (pooled/averaged values)	Point elast.	No (synthesis)	Inter-urban	n.d.	Tab. 7	-
Börjesson, 2014	Own	Fare	Rail	Rail	−0.59 to −0.72	Sweden	Nested logit (long-distance model)	-	No	Swedish long-distance trips (≥100 km)	n.d.	Tab. 3	-

continued on next page

Table D.1 (continued)

Study	Type	Variable / finding	Altered	Affected	Magnitude	Region	Method	Form	FE	Distance	Run	Source	Notes
	Own	In-vehicle time	Rail	HSR	−1.5 to −1.0	Sweden	Nested logit (long-distance model)	-	No	Swedish long-distance trips (≥100 km)	n.d.	Tab. 3, 4	-
	Cross	In-vehicle time	Rail	Air	0.16 to 0.20	Sweden	Nested logit (long-distance model)	-	No	Swedish long-distance trips (≥100 km)	n.d.	Tab. 3	-
Hortelano et al., 2016	Own	Fare	HSR	HSR	−0.583	Spain	Nested logit (mode choice)	Point elast.	No (single intercept survey)	Spanish HSR corridors	SR	Tab. 3	-
	Own	Fare	Air	Air	−1.537	Spain	Nested logit (mode choice)	Point elast.	No (single intercept survey)	Spanish HSR corridors	SR	Tab. 3	-
	Cross	Fare	Air	HSR	0.097	Spain	Nested logit (mode choice)	Point elast.	No (single intercept survey)	Spanish HSR corridors	SR	Tab. 3	-
	Cross	Fare	HSR	Air	0.692	Spain	Nested logit (mode choice)	Point elast.	No (single intercept survey)	Spanish HSR corridors	SR	Tab. 3	-
Wijeweera et al., 2014	Own	Fare	Rail (commuter)	Rail (commuter)	−0.3 to +0.5	Australian cities	-	Log-log	No (separate per city)	Commuter rail	Both	Tab. 1	-
	Own	Income (GDP proxy)	Rail (commuter)	Rail (commuter)	−0.3 to −0.5	Australian cities	-	Log-log	No (separate per city)	Commuter rail	Both	Tab. 2	-
	Own	Annual vehicle-km (capacity/frequency proxy)	Rail (commuter)	Rail (commuter)	0.5 to 1.2	Australian cities	-	Log-log	No (separate per city)	Commuter rail	Both	Tab. 3	-
	Cross	Fuel/vehicle price (car cost proxies)	Car	Rail (commuter)	+0.2 to −0.3	Australian cities	-	Log-log	No (separate per city)	Commuter rail	Both	Tab. 4	-

Notes: One row per estimate or finding *Altered*: mode whose attribute changes; *Affected*: mode whose demand (or choice probability) responds. IVT: in-vehicle time; HSR: high-speed rail; SR/LR: short/long run; n.d.: not distinguished; RP: revealed preference; FE: fixed effects; GVA: gross value added; "-": not reported or not applicable.

E

Results of alternative model specifications

Results of ticket sales pooled model without air travel used for four stations in the case study which do not have any air travel alternative available.

E.1. Ticket sales model without air travel variables

Variable	Pooled panel model Parameter	p-value	OD fixed effects Parameter	p-value	OD time fixed effects Parameter	p-value
Model constant						
α	-6.5596	0.0148	-	-	-	-
Service type dummy variables						
$\delta_{ICE,Frankfurt}$	-	-	-	-	-	-
$\delta_{ICE,Basel}$	-	-	-	-	-	-
$\delta_{ICE,Munich}$	-	-	-	-	-	-
$\delta_{Eurostar,Paris}$	-	-	-	-	-	-
$\delta_{Eurostar,Seasonal}$	-	-	-	-	-	-
OD characteristics						
β_{pop_i}	-0.0012	0.9861	-	-	-	-
β_{pop_j}	0.1225	0.0001	-	-	-	-
β_{gdp_i}	0.2165	0.4514	-	-	-	-
β_{gdp_j}	0.3686	0.0003	-	-	-	-
β_D	0.3675	0.0021	-	-	-	-
Rail service attributes						
$\beta_{f_{rail}}$	-0.0227	0.4063	-	-	-	-
$\beta_{TT_{rail}}$	-0.3986	0.0104	-	-	-	-
$\beta_{freq_{rail,ij}}$	-	-	-	-	-	-
β_{cap}	-	-	-	-	-	-
β_{TR}	-0.3639	0.0000	-	-	-	-
Air travel characteristics						
$\beta_{freq_{air}}$	-	-	-	-	-	-
$\beta_{f_{air}}$	-	-	-	-	-	-
Model statistics						
Unit of analysis	OD-day tickets		OD-day tickets		OD-day tickets	
Observations	511743		-		-	
R^2	0.1523		-		-	
Within R^2	-0.0895		-		-	

Table E.1: Results for models for ticket sales data without air travel characteristics (green: $p < 0.01$, yellow: $0.01 \leq p \leq 0.05$)

