UNIVERSITY OF BERGAMO

Department of Management, Information and Production Engineering

Doctoral Degree in Technology Innovation and Management

XXXVIII Cycle

Advancing Air Travel Demand Forecasting through Data-Driven Methodologies

Thesis Supervisor

Prof. Renato Redondi

Thesis Co-Supervisor

Prof. Sebastian Birolini

Author

Andrea Signori

PhD Candidate ID: 1075526

Academic year 2024/2025

Contents

1 Introduction **8**

2 Research Background **13**

2.1 Literature Review 13

2.1.1 Air Demand Determinants 16

2.1.2 Air Demand Modeling 20

2.1.3 Airline Planning Process 26

2.2 Research Outline and Contributions 30

3 Forecasting Non Stop Demand Using Advanced Predictive Analytics **37**

3.1 Introduction 37

3.2 Related Literature 40

3.3 Data and methods 43

3.3.1 Data sources 44

3.3.2 Variables definition 45

3.3.3 Econometric formulation 48

3.3.4 Supervised machine learning formulation 53

3.4 Results and validation 55

3.4.1 Econometric results 55

3.4.2 Supervised machine learning results 58

3.4.3 Validation experiment 64

3.5 Conclusion 68

4 Air Demand Impacts of China's 2030 High-Speed Rail Expansion Plan **71**

4.1 Introduction 71

4.2 Related Literature 75

4.3 Data and methods . . . 79
4.3.1 Data sources . . . 79
4.3.2 Variables definition . . . 81
4.3.3 Methodology . . . 85
4.4 Econometric results . . . 87
4.5 Case study: 2030 railway network expansion . . . 93
4.5.1 Comparative travel time assessment . . . 93
4.5.2 Air demand future trend . . . 98
4.6 Conclusion . . . 106

5 Airlines Network Development: A Worldwide Empirical Investigation 109

5.1 Introduction . . . 109
5.2 Related Literature . . . 111
5.2.1 Route entry decision . . . 111
5.2.2 Strategic network planning . . . 113
5.3 Data and methods . . . 115
5.3.1 Data sources . . . 115
5.3.2 Variables definition . . . 118
5.3.3 Methodology . . . 124
5.4 Results . . . 130
5.4.1 Econometric results . . . 130
5.4.2 Airline case study . . . 137
5.5 Conclusion . . . 138

6 Conclusion 142

List of Figures

1 Comparison of Demand Generation and Allocation Models 14

2 Demand models based on decision type vs time horizon (the darker the color and the bigger the circle dimension, the more frequently employed is the mentioned approach in the related decision type - time horizon segment). 14

3 Predicted vs. Observed demand values. 61

4 Feature importance of Random Forest computed using SHAP plot. . . . 63

5 Stimulation effect ratio patterns by frequency distribution with a real growth ranging from 1% to 5%. 65

6 Stimulation effect ratio patterns by frequency distribution with a real growth ranging from 5% to 10%. 66

7 Correlation between real stimulation effect ratio and predictions from OLS and ML models. 66

8 Stimulation effect ratio patterns by simulating a frequency increasement of 5%. 67

9 Stimulation effect ratio patterns by simulating a frequency increasement of 10%. 67

10 M2 aggregated model specification results using test set (2018-2019) for measuring the predictive performance (results has been linearized for the representation). 91

11 Variables importance level across the four models configurations (colors correspond to the relative importance of each variable, with lighter shades indicating higher importance). 92

12 Backbone HSR lines planned to be opened in 2030. Subplots divide the network lines into vertical connections (a) and horizontal connections (b). 94

13 Total travel itinerary for a passenger supposing to start and end the trip to the city center, either by HSR or air with the support of car or public transport for the segment city center-station, and vice versa. 96

14 Comparative travel times along the backbone HSR lines 2030, with different penalty time scenarios. Subplots (a) and (b) show: (a) No dwell time applied; (b) Dwell time of 30 minutes for HSR and 90 minutes for air. . . 97

15 Comparative travel times along the backbone HSR lines 2030, with different penalty time scenarios. Subplots (a) and (b) show: (a) Dwell time of 60 minutes for HSR and 90 minutes for air; (b) Dwell time of 60 minutes for HSR and 120 minutes for air. 98

16 The plots show the historical evolution of frequency and market concentration (HHI) over the period 2015-2019 in two different scenarios: blue lines display the growth of city pairs also connected with non stop HSR line, red lines display the growth of city pairs also connected with at least 1 stop connecting HSR line. 100

17 Chinese population (plot A) and GDP per capita in $ thousands (plot B) projections from 2025 to 2050, under SSP scenario. 102

18 City pairs connection with color differentiation based on the air demand increase in 2025, without the new set of HSR lines opening. 104

19 City pairs connection with color differentiation based on the air demand decrease and disappearing during the interval 2030-2050, following the HSR lines opening. 104

20 Relation between origin-destination distance and year experiencing the total demand shift in air-HSR. 105

21 Heatmap of the correlation between variables (train dataset balanced through oversampling). 122

22 Coefficient deviation (stability) from their mean over 100 iterations: oversampling vs. undersampling. 128

23 Forecasting performance KPIs: oversampling vs. undersampling. 129

24 Model results: heatmap with normalized coefficients. 132

25 KPIs across different threshold levels. 135

26 Promising routes for three selected airlines as identified by the classification model applied to 2019 data. Darker lines indicate higher predicted probabilities. 139

List of Tables

1 Summary of contributions to the literature by chapter. 31

2 Descriptive statistics 49

3 Variable definition and formulation. 50

4 Econometric results 59

5 Key performance indicators ML vs econometrics 62

6 Summary of literature for air-HSR competition. 76

7 Data Sources 81

8 Descriptive analytics 82

9 Results from linear and semi-log formulation 90

10 Demographic assumptions in China under SSPs. In SSP4, assumptions for education depend on the provincial development level. 101

11 Number of observations by year and type for the top 20 carriers by number of long-haul routes operated in 2019. 116

12 Sample by carrier and type of observation: new or potential. 118

13 Variable definition and formulation. 120

14 Model KPIs considering a threshold of 0.5 to define promising routes. . . 132

15 Carriers metrics by threshold. 136

Chapter 1

Introduction

Since its inception, the air transport industry has been growing at a fast pace to establish itself as one of the major industries in today's global economy. According to the International Air Transport Association (IATA), by 2019 - just before the outbreak of the COVID-19 pandemic that affected the industry for about 4 years[1] - the aviation sector had reached impressive milestones, with airlines accommodating approximately 4.5 billion passenger journeys and operating over 22,000 unique city-pair routes worldwide (International Air Transport Association 2020). This represented a remarkable increase of almost 1,000 new routes compared to the previous year and more than doubled the number of routes since the early 2000s. Particularly, the sector experienced a compound annual growth rate (CAGR) of 5.3% since 2000, with the average interval between air flights for individuals halving from 44 months in 2000 to just 20 months in 2019. However, the outbreak of the COVID-19 pandemic in 2020 severely impacted the industry, resulting in unprecedented declines in passenger traffic and route operations. Despite this, today the aviation sector has shown strong signs of recovery. As reported by International Air Transport Association (IATA) (2024), by 2024 airlines had carried approximately 5 billion passengers, surpassing pre-pandemic levels, and the number of city-pair routes remained stable, similar to those seen in 2019. Furthermore, global demand, measured in revenue passenger kilometers (RPK), rose by 7.1% in October 2024 compared to the same month in 2023, underscoring the recovery momentum. Looking ahead, the industry's future growth prospects are equally promising, with IATA Global Outlook (International Air Transport Association (IATA) 2021) predicting that global passenger numbers will reach

[1] IATA reports that by November 2023, global air travel demand had reached 99.1% of November 2019 levels, while in February 2024, the airline industry achieved full recovery in total passenger traffic, surpassing February 2019 levels by 5.7%.

10 billion by 2050, driven by emerging markets and the continued expansion of aviation networks, particularly in Asia and Africa.

The utmost relevance of air transport industry is not only due to its direct economic impact but mostly to its boosting effect on global trades and socio-economic development. According to the Air Transport Action Group's (ATAG) (Air Transport Action Group (ATAG) 2024), the global aviation industry significantly contributes to the world economy. In 2023, air transport supported approximately 86.5 million jobs worldwide, encompassing direct, indirect, induced, and tourism-related employment. This marks an increase from the 65.5 million jobs estimated in 2016. Additionally, the industry's total economic impact reached approximately US$4.1 trillion in 2023, accounting for about 3.6% of global GDP, up from US$2.7 trillion in 2016. Moreover, the economic impact of aviation is projected to grow substantially, with the sector expected to contribute $4.1 trillion to global GDP by 2036, supporting over 87 million jobs worldwide. These projections, along with the industry's recovery to pre-pandemic levels, further support the thesis that aviation is one of the leading industries in today's global economy, facilitating unprecedented levels of international trade, tourism, and economic integration. The aviation industry not only plays a vital role in global connectivity but is also a significant driver of economic activity and social development, positioning it as a key pillar of modern global economies.

Following the deregulation of the airline industry in the late 20th century, air transport networks rapidly expanded and became more complex, driven by increased competition and the entry of new players into the market. Deregulation eliminated many of the previous government-imposed restrictions on routes, pricing, and capacity, fostering a more competitive environment where airlines had greater flexibility to set their schedules and prices (Çetin and Eryigit 2018). This transformation led to significant growth in the number of routes, the variety of services offered, and the introduction of low-cost carriers, all of which reshaped the dynamics of the industry. However, the removal of regulatory constraints also introduced new challenges, as airlines had to adapt quickly to shifting

market conditions, fluctuating demand, and evolving consumer preferences (Gillen and Lall 2004). In this highly competitive and dynamic landscape, the ability to forecast demand accurately and optimize operations became crucial for maintaining a competitive edge. Predictive analytical methods, with a strong focus on demand forecasting, have become essential tools for airlines and other stakeholders in the aviation industry to navigate these complexities. By harnessing data-driven models, airlines and airports can accurately predict fluctuations in passenger demand, enabling them to optimize fleet management, slot allocation, and pricing strategies. These capabilities help streamline operations, improve resource allocation, and drive cost-efficiency, ensuring that industry players remain competitive and responsive to changing market conditions.

Thanks to its pervasive and critical role in aviation planning, demand forecasting remains a widely studied topic, continuously evolving as researchers and practitioners seek to refine its accuracy and applicability. Given its profound impact on airline operations, market competition, and strategic decision-making, substantial efforts have been dedicated to advancing forecasting methodologies. Scholars and industry experts are actively exploring the integration of machine learning and big data analytics to enhance predictive accuracy and to capture complex demand patterns, enabling airlines to anticipate demand patterns with improved precision (Lin et al. 2019, Firat et al. 2021). Moreover, demand forecasting does not operate in isolation; rather, it is deeply intertwined with a range of economic, operational, and competitive factors. One of the most pressing challenges for airlines today is the growing competition from high-speed rail (HSR) on short- and medium-haul routes, particularly in regions Asia, where HSR networks continue to expand. Many demand models now increasingly incorporate HSR availability and passenger preferences to assess the extent of air traffic diversion and the potential impact on route profitability (Sun et al. 2017, 2024). These models help airlines determine which routes remain viable, where adjustments in pricing and scheduling are needed, and whether strategic partnerships with rail operators could offer competitive advantages. Understanding these interdependencies is crucial for developing robust forecasting mod-

els that can adapt to complex and dynamic environments. Beyond predictive capabilities, the integration of demand forecasting with prescriptive analytics represents another challenges for scholars. While traditional forecasting models provide insights into expected passenger volumes, prescriptive analytics takes this a step further by recommending optimal responses based on these predictions. Advanced optimization algorithms are being increasingly employed to align demand forecasts with dynamic capacity planning, pricing strategies, and operational adjustments. This integration ensures not only greater adaptability to market fluctuations but also the ability to proactively shape demand through targeted interventions, such as demand-based pricing, capacity reallocation, or strategic route expansion (Sismanidou and Tarradellas 2017, Wandelt et al. 2025). As technological advancements continue to push the boundaries of demand forecasting, the aviation industry stands to gain significantly from these innovations. Accurately anticipating passenger demand and adapting strategies accordingly has become essential in an industry shaped by rapid change, intensifying intermodal competition, and shifting consumer preferences.

In this context, the aim of this thesis is to contribute to the rapidly evolving field of demand forecasting by developing models that analyze air passenger demand across different scenarios and time horizons. To achieve this, we first examine the trade-offs between econometric and machine learning techniques in short-term demand prediction, assessing their relative strengths and applicability in dynamic market conditions. Second, we analyze the long-term impact of high-speed rail (HSR) competition on air travel demand, identifying key factors that influence market shifts and modal substitution. Third, we develop a forecasting model to be integrated with network planning algorithms and route development strategies, providing a data-driven framework for long-term decision-making. The final goal is to enhance the predictive accuracy and strategic applicability of demand forecasting models, by exploring their potentialities across diverse contexts and aligning them with the specific needs of industry stakeholders.

The remainder of this thesis is structured as follows. Chapter 2 presents a comprehensive review of the literature on air demand forecasting, with a particular focus on

methodologies, key determinants, and its integration with other aviation-related tasks. This chapter establishes the theoretical foundation and familiarizes the reader with the core concepts that will be recurrent throughout the thesis. In Chapter 2.2, we state in detail the thesis main contributions and provide an outline of the remaining chapters. Chapters 3, 4, and 5 consist of stand-alone academic papers, each addressing a specific challenge related to the research objectives, providing empirical insights and methodological advancements. Finally, Chapter 6 synthesizes the key findings, offering a general conclusion and outlining potential avenues for future research.

Chapter 2

Research Background

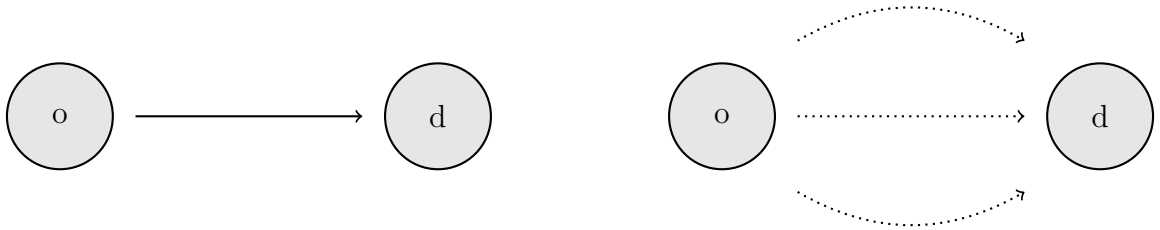
This section offers a broad overview of the evolving research on air demand modeling and forecasting. Although not intended to be exhaustive, it serves to contextualize the research framework underpinning the three subsequent contributions. Each contribution will include a dedicated literature sub-section to introduce key concepts specific to its analytical scope.

2.1 Literature Review

The provision of air transportation services is fundamentally driven by air travel demand. Analyzing air travel market and provide accurate air traffic volume estimates is an important input for a wide variety of economic decisions, including but not limited to, research and development, airplane design, production planning. Accurate demand modeling plays a crucial role in enhancing passenger services by enabling airlines to efficiently allocate capacity, optimize pricing strategies, and improve overall service levels (Carson et al. 2011, Belobaba et al. 2015).

An important distinction to consider when dealing with the estimation of air travel demand is between demand generation and demand allocation models. Demand generation models estimate the total volume of travel demand between two areas (Figure 1a) within a given market or region. These models focus on identifying the underlying drivers of demand, such as economic growth, population, income levels, and airfares. Given their emphasis on structural and macroeconomic factors, they are primarily used for long-term forecasting. They help predict overall market size and potential growth by providing aggregate traffic flow estimates across territories (Grosche et al. 2007, Adler and Hashai 2005). Demand allocation models, in contrast, determine how the total demand

is distributed among available travel alternatives (Figure 1b). They consider factors such as airline competition, flight frequency, fares, travel time, and service quality to model passenger choice behavior. Unlike demand generation models, which estimate the overall demand level, allocation models take total demand as given and focus on short- to medium-term forecasting to get insights on how passengers select between airlines, routes, and itineraries. Airlines heavily rely on these models to inform tactical decisions regarding service design, pricing strategies, and capacity allocation (Button and Drexler 2005, Pita et al. 2013).



(a) Demand Generation **(b)** Demand Allocation

Figure 1: Comparison of Demand Generation and Allocation Models

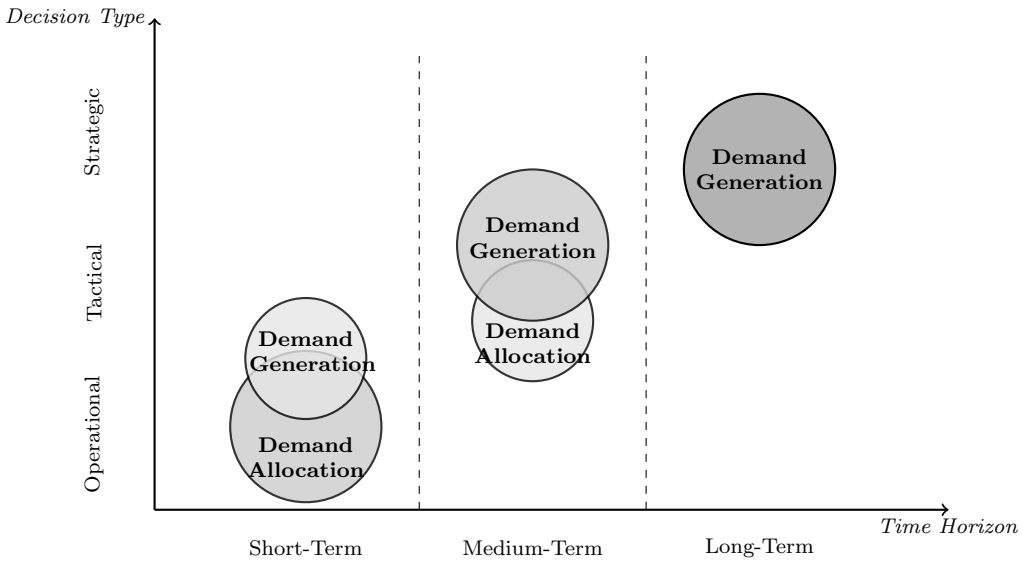


Figure 2: Demand models based on decision type vs time horizon (the darker the color and the bigger the circle dimension, the more frequently employed is the mentioned approach in the related decision type - time horizon segment).

The distinction between demand generation and demand allocation models aligns closely with the different decision-making levels in aviation industry management, each

characterized by its time horizon and objectives (Figure 2)[2]. Strategic decisions, which shape a long-term trajectory over 5 to 30 years, rely primarily on demand generation models to estimate overall market growth, guide fleet acquisitions, infrastructure investments, and evaluate network development opportunities. In this time frame, demand allocation models are typically absent, as their core assumption about endogenous demand becomes unrealistic over such extended time horizons. Instead, demand generation models are more appropriate for capturing long-term variations and structural trends driven by macroeconomic, demographic, or technological factors[3].

Operational decisions, which involve short-term adjustments over days to months, such as crew scheduling, aircraft routing, and revenue management, mainly depend on demand allocation models to optimize real-time service execution. Nevertheless, at this stage also demand generation models may be employed. In the short term, demand generation models are less commonly applied, as the relative stability of demand means that assuming a fixed (endogenous) value or estimating it through a model often leads to comparable outcomes. Unlike in long-term contexts—where demand is subject to significant fluctuations and cannot realistically be treated as fixed—the short time horizon makes both approaches similarly valid. Nevertheless, demand generation models can still provide enhanced predictive accuracy, particularly when capturing subtle variations that fixed-demand assumptions may overlook. During these first two stages, demand generation and allocation have been traditionally solved as sequential steps, each handled by distinct models by assessing demand generation at one level of aggregation and then distributing the estimated volumes to different itineraries (Hsiao 2008). For example, in short-term allocation models, total market demand is typically considered fixed, with the focus placed on optimizing how it is assigned among available travel options. In contrast, tactical decisions, typically cover a medium-term period spanning 1 to 5 years, focus on

[2]In the context of this thesis, however, the empirical work relies primarily on demand generation models. The framework is therefore introduced not to suggest that the subsequent studies perform explicit demand allocation, but rather to provide a broader conceptual structure within which the modeling choices can be situated.

[3]Regardless the time horizon, also demand generation models have to deal with the demand endogeneity issue, that usually refers to the demand-supply interaction.

schedule planning, and pricing strategies. At this stage, demand generation and allocation are no longer treated sequentially but are addressed simultaneously using hierarchical demand models. The upper level estimates total market demand as a function of various air travel determinants, while the lower level allocates this demand based on the utility of different travel alternatives (Wei and Hansen 2005, Hsiao and Hansen 2011). This integrated approach overcomes the limitations of sequential modeling, where assuming a fixed total demand for allocation fails to capture the influence of itinerary-specific attributes on overall air traffic volume. By integrating both modeling approaches across different time horizons, airlines can ensure effective long-term planning while maintaining agility in responding to evolving market conditions.

Literature on air travel demand estimation is widespread, with a focus on two central issues: identifying the key determinants of air travel demand and developing model formulations that effectively capture passenger behavior, while accounting for the level of data aggregation and the available information.

2.1.1 Air Demand Determinants

As one of the pioneers in the field, Jorge-Calderón (1997) offers a widely accepted classification of air travel determinants, dividing them into two broad categories: geo-economic characteristics of the regions where travel occurs and service-related determinants, which are influenced by air services and thus partially controlled and managed by industry players such as airlines and airports. For the first category, numerous studies (Grosche et al. 2007, Boonekamp et al. 2018, Javadinasr et al. 2022, Adekunle et al. 2025) have emphasized key geo-economic determinants such as population size, income levels, trade and tourism flows, cultural, ethnic, and political links between countries, and aviation-dependent employment. In terms of service-related determinants, factors like flight frequency, travel time, itinerary routing (with particular focus on the disutility of connecting versus non-stop flights), schedule delays, departure times, and price, among others, have been identified as critical level-of-service attributes that influence passenger choices (Kop-

pelman et al. 2008, Lurkin et al. 2017, 2018, Kristoffersson and Liu 2024).

Among these determinants a first challenge is represented by the interdependence between demand and supply side. Substantial evidence in the literature supports the bi-directional relationship between air travel supply (often represented by flight frequency) and demand. A prominent example of this dynamic is the rise of low-cost carriers, which, by offering low fares and point-to-point services, are able to stimulate demand and serve markets previously considered economically unviable, such as remote or leisure-dominated destinations (Fu et al. 2010, Antunes et al. 2020). Other studies have emphasized the crucial role of additional airline offerings, such as increased frequencies and optimized schedules, in both attracting demand from competitors and generating new traffic (Abdelghany et al. 2017, Cadarso and Vaze 2023). This interplay highlights how the availability of air services not only shapes demand but is also influenced by demand levels, with higher demand prompting airlines to expand and adapt their service offerings. Properly capturing this two-sided relationship requires careful methodological consideration, as the dynamic interplay between supply and demand poses challenges in establishing causality and isolating the direction of effects. Despite extensive research on air travel determinants, the most appropriate empirical modeling approaches for addressing this interdependence remain an open question in the literature (see Section 2.1.2).

Parallel to the well-established body of research on air travel demand determinants, a substantial literature has emerged examining the challenge of competitive and collaborative dynamics between high-speed rail (HSR) and air transportation.

On one side, HSR has been shown to complement air transport. Specifically, in a joint-decision-making context, the two modes can collaborate to offer intermodal services, particularly for long-haul air routes (Jiang et al. 2021). Growing literature highlights the potential of air-HSR cooperation, with increasing support for integrating these services to alleviate hub airport congestion, reduce the environmental impact of air travel, and expand airport catchment areas (Lee et al. 2024, Avenali et al. 2025). Cross-modal collaboration, in this context, has proven to be an effective solution for feeder services to

airports, particularly when hub airport capacity is constrained (Xia and Zhang 2016).

Conversely, numerous studies in this field have examined the degree to which HSR acts as a substitute for air travel, focusing on its impact on airlines. The air-HSR interaction, viewed through the lens of competition, seeks to determine how much HSR can replace air travel and the resulting effects on the aviation industry (Bergantino and Madio 2020, Li et al. 2024). Specifically, most previous research has highlighted significant competition between HSR and air transport in short-haul markets (Milan 1993, D'Alfonso et al. 2016, Sun et al. 2024). In terms of airline responses, empirical evidence indicates that the introduction of parallel HSR services exerts downward pressure on key air transport variables —capacity, frequency, and fares— which not only serve as fundamental determinants of air travel demand but also play a crucial role in shaping the competitive balance between air and HSR. These factors are particularly significant in modeling the dynamic interplay between the two modes, as they influence both passenger preferences and airlines' strategic responses to HSR competition (Mizutani and Sakai 2021, Zhang et al. 2019, Wang et al. 2024). These effects manifest at both the tactical level, influencing airline scheduling optimization, and the strategic level, shaping market coverage and network design decisions (Cadarso et al. 2017, Jiang et al. 2022, Fang et al. 2024). The impact of HSR competition is particularly pronounced in China, where advanced railway infrastructure has led to profound shifts in market dynamics. For instance, air travel demand on the Beijing–Shanghai route dropped by 34% following the introduction of HSR (Chen 2017). On the Wuhan–Xiamen corridor, airfares plummeted by 80% after HSR entry (Jiang and Zhang 2014). While the introduction of HSR often triggers a gradual market adjustment, certain cases demonstrate abrupt shifts. In March 2010, all flights between Zhengzhou and Xi'an were discontinued just seven weeks after HSR service began (Fu et al. 2012). Similarly, the Guangzhou–Wuhan air route experienced a 48% reduction in seat capacity within a year of HSR launch, despite airlines cutting fares by more than 50% (Fu et al. 2012). These cases underscore the disruptive potential of HSR, particularly when its service characteristics—such as frequency, accessibility, and travel time—closely

align with or surpass those of competing air routes. While the competition between HSR and air transport has been so far extensively studied, a significant gap remains in the forecasting of HSR's impact on future air markets. Most existing research focuses on retrospective analyses, assessing the effects of HSR on established air routes and market dynamics. However, relatively few studies attempt to predict the potential consequences of new HSR lines that are still in the planning stages with decades long-term future perspective. Wang et al. (2018c) explored how future HSR expansion could influence ground accessibility, offering insights into the broader transformation of transportation networks. Similarly, Zhou et al. (2018a,b) examined the evolution of China's HSR system over the past decade, proposing a connectivity index to measure city integration within the network. While these studies provide valuable intuition, the literature still lacks a clear research direction on how upcoming HSR developments will reshape air markets in the long-term future, leaving an important gap in predictive modeling and strategic planning.

In conclusion, the determinants of air travel demand are shaped by a variety of socio-economic and service-related factors, with a notable interdependence between the supply and demand sides. Understanding these dynamics presents several challenges, particularly in addressing the bi-directional relationship between air travel supply and demand. This interplay complicates the establishment of clear causality in empirical models, necessitating sophisticated methodologies to capture the dynamic relationship between supply and demand. Furthermore, factors such as population size, income levels, and trade flows remain fundamental geo-economic determinants of demand, while service-related aspects like flight frequencies, travel times, and price continue to play critical roles in shaping passenger behavior. Despite the progress in the field, existing research still tries to improve the modeling of these determinants and their incorporation into mathematical model formulations, especially given their reciprocal interactions. Overcoming these challenges is essential for improving the predictive accuracy of air travel demand models and for informing both industry strategies and policy decisions.

2.1.2 Air Demand Modeling

On the modeling side, most studies in this field rely on econometric or statistical models to analyze the impact of predictor variables on air travel demand. These approaches fall under the category of parametric methods, which offer the advantage of clear interpretability by explicitly defining the functional relationship (e.g., linear, logarithmic) between dependent and independent variables. This structured framework allows for a precise estimation of the marginal effect of each predictor while holding other variables constant (Wooldridge 2010). One of the earliest and most widely adopted parametric models is the gravity formulation (Jorge-Calderón 1997, Grosche et al. 2007, Adler and Hashai 2005, Adler et al. 2018), inspired by Newton's law of gravity and mainly used in the realm of demand generation. These models conceptualize demand as the outcome of competing forces of attraction and impedance[4]. Over time, researchers have refined the basic gravity model by incorporating service-related variables to better capture the role of service quality in stimulating air travel demand. Earlier studies often examined factors such as average flight frequency, ticket prices, or the presence of low-cost carriers in isolation. A more comprehensive alternative is the generalized cost approach, which consolidates multiple travel -related components—such as travel time, flight frequency, and waiting time —into a single measure, providing a more holistic representation of the passenger experience (Koppelman et al. 2008, Lieshout et al. 2016, Zachariah et al. 2023). The models are typically log-linearized and estimated using linear regression techniques and time-series methods and integrated with discrete choice models (Valdes 2015, Cook et al. 2017, Wang and Gao 2021). Beyond gravity-based formulations, air travel demand forecasting employs a range of econometric and statistical methodologies tailored to different analytical needs. Among these, time-series models, such as autoregressive integrated moving average (ARIMA) and its variations (e.g., SARIMA for seasonal ad-

[4]In its simplest form, the gravity model can be formulated as follows: $D_{od} = \frac{M_o^{\beta_1} M_d^{\beta_2}}{T_{od}^{\beta_3}}$, where D_{od} indicates the total demand between o and d in a given period, M_o and M_d represent attraction factors (e.g. population, income) at the origin and destination, and T_{od} includes impedance factors that restrain air travel between the two areas (e.g. distance) and $\beta are the three calibrated exponents (parameters)$.

justments), are widely used for short- to medium-term forecasting, as they effectively capture temporal patterns in demand by analyzing historical trends and cyclic behaviors (Kim et al. 2012, Jungmittag 2016). For long-term forecasting, however, econometric models—particularly panel data regression models and structural demand models—are more commonly employed. These methods incorporate macroeconomic indicators (e.g., GDP growth, population trends) and infrastructure developments (e.g., airport expansions, high-speed rail competition) to provide insights into long-term demand evolution and account for both cross-sectional and time-series variations, making them better suited for strategic planning and policy evaluation. These models allow researchers to control for unobserved heterogeneity across different markets while capturing dynamic changes over time. Fixed-effects and random-effects specifications are commonly used, depending on the nature of the data and assumptions regarding unobserved factors (Rosenfield et al. 2020, Dayioglu and Alnipak 2023).

Another key methodology in air travel demand forecasting in the realm of demand allocation is the discrete choice model approach and the earliest formulation relied on a "S-curve" relating flight frequencies with itinerary market shares (Button and Drexler 2005, Pita et al. 2014)[5]. Discrete choice modeling, particularly the logit model, are used to analyze passenger preferences and decision-making behavior and rooted in random utility theory, which assumes that individuals make choices by selecting the alternative that provides them with the highest perceived utility. In this framework, the utility associated with a given choice is composed of two components: a deterministic (observable) component, which can be estimated based on measurable attributes (e.g., travel time, cost, frequency), and a random (unobservable) component, which captures unknown influences on decision-making[6]. If the random component is assumed to follow an independent and

[5]Let k be a given itinerary operating in market *(o,d)*, demand D_{odk} on itinerary k is estimated as follows: $D_{odk} = D_{od}\frac{N_k^{\mu}}{\sum_{k' K} N_{k'}^{\mu}}$, where D_{od} is the total market demand between o and d, N_k is the service frequency of itinerary k, K is the set of available itineraries in market *(o,d)*, and μ is an empirical coefficient.

[6]For a binary choice situation—such as selecting between air and rail travel—the utility of each alternative U_i is expressed as: $U_i = V_i + \epsilon_i$, where V_i is the deterministic component, typically modeled as a linear function of explanatory variables, such as price, travel time, and frequency; ϵ_i represents the

identically distributed (IID) extreme value (Gumbel) distribution, the probability that a passenger chooses one alternative over another follows the logistic function, leading to the binary logit model (Cascetta and Cascetta 2009). The logit model is widely used to analyze mode choice behavior, policy impacts (e.g., the introduction of high-speed rail), and competition between airlines. However, its independence of irrelevant alternatives (IIA) property—which assumes that the relative probability of choosing between two alternatives remains unchanged if a third option is introduced—can be a limitation in more complex decision-making scenarios (Garrow 2016, Mokhtarian 2016). These models estimate the probability of passengers choosing a specific mode of transport (e.g., air vs. rail) or airline based on factors such as fare, frequency, travel time, service quality, and personal characteristics. Discrete choice models are particularly valuable for understanding travel behavior at different time horizons and evaluating the impact of market or policy changes. For short-term analysis, discrete choice models are often applied to assess immediate responses to changes in airfare, route availability, or service frequency. While in long-term forecasting, discrete choice models are often combined with scenario-based simulations to estimate how major developments—such as high-speed rail expansion, airport capacity increases, or changes in fuel costs—affect modal choice over extended periods.

A critically important and often challenging feature of econometric prediction tasks, especially when using gravity formulations, is the incorporation of key supply-related predictors. It is widely acknowledged that demand and supply are not independent but instead are intricately linked in a two-way relationship. This dynamic interaction complicates the modeling process, as it introduces the potential for reverse causality—where changes in demand may simultaneously influence supply and vice versa. From an empirical perspective, this reverse causality can lead to biased estimates and misrepresentations of the true relationship between the variables, distorting the overall predictions and undermining the model's accuracy. To mitigate this problem, researchers commonly turn to instrumental variable (IV) techniques, which aim to isolate the exogenous vari-

random component, which accounts for unobserved factors influencing the choice.

ation in the supply variable that is not affected by demand. The idea is to use instruments—variables that are correlated with the endogenous explanatory variable (e.g., supply) but are not correlated with the dependent variable (e.g., demand) or the error term (Hsiao and Hansen 2011, Boonekamp et al. 2018). By doing so, these techniques attempt to break the feedback loop and provide more accurate estimates of causal effects. However, the practical application of instrumental variable methods in air transport demand modeling is far from straightforward. A significant challenge is the identification of valid instruments. An instrument must satisfy two conditions: it must be correlated with the endogenous variable (e.g., supply) and must be exogenous to the dependent variable (demand). The difficulty in satisfying these conditions stems from the complex nature of the air travel market, where both supply-side and demand-side factors are often influenced by similar unobserved factors (e.g., economic conditions, regulatory policies). Identifying control variables that meet these criteria is not only conceptually difficult but also empirically challenging, leading many studies to acknowledge the problem of weak instruments (Brueckner et al. 2014, Suh and Ryerson 2019). Weak instruments —those that are poorly correlated with the endogenous variable—pose a significant problem. When weak instruments are used in place of a stronger predictor, they fail to accurately capture the variations in supply that are unrelated to demand, leading to biased estimates. They also lead to an imprecise treatment of endogeneity, further complicating model results. If the instrumental variables do not adequately address the endogeneity issue, the estimates of supply's impact on demand will be unreliable (Hahn and Hausman 2003, Keane and Neal 2024). This results in lower prediction accuracy, which ultimately impacts the ability to forecast demand accurately over time. Given that both demand generation and allocation models often rely on accurate supply-side variables (e.g., flight frequencies, airline competition, airport capacity), the failure to properly account for supply-side endogeneity affects not only the estimation of total demand (by over- or under- estimate the passengers flow) but also the distribution of demand across different routes or airlines (by erroneously predict an airline's market share, influencing decisions on competitive

pricing, route development, or expansion strategies). In conclusion, while the importance of supply-related determinants is not under discussion, the reverse causality issue presents a trade-off between methodological rigor and accuracy, ultimately compromising the forecasting reliability of the model (Zhang et al. 2017b, Birolini et al. 2020). This limitation highlights the reduced explanatory power in predicting the dependent variable, as studies often sacrifice accuracy in their attempts to address endogeneity, leading to diminished predictive performance. Such difficulty ultimately limits the applicability of parametric methods in addressing the demand-supply interrelation, as they are easily interpretable, but with the risk of biased estimates that are not able to disentangle the two-way stimulation effect and provide reliable models.

In addition to traditional parametric methods, non-parametric approaches, particularly machine learning (ML) algorithms, have gained significant attention in recent years for their potential in real-world applications, especially in complex fields like demand modeling. Unlike conventional econometric techniques, which often rely on predefined assumptions about the model structure, machine learning algorithms are driven predominantly by the data itself (Maheshwari et al. 2018). This fundamental difference makes ML models inherently more flexible, enabling them to adapt and uncover complex patterns in the data without the constraints of assumed relationships or causal structures. This data-driven approach allows ML algorithms to potentially enhance the accuracy of demand modeling, as they are designed to learn from the data itself rather than being forced to fit a rigid model framework. One of the primary advantages of machine learning in demand modeling is its ability to handle complex, non-linear relationships between predictors and outcomes. Traditional econometric models, by contrast, often assume linearity or rely on simpler functional forms to model the relationships between variables. This assumption can limit their ability to accurately represent real-world phenomena, where relationships may be far more intricate and non-linear. Machine learning algorithms, on the other hand, excel in capturing these complex, non-linear interactions, which enables a more accurate prediction of demand patterns. Furthermore, they are

particularly well-suited for handling large datasets, which is increasingly important as the volume of data continues to grow in the digital age. While traditional econometric methods can struggle with vast amounts of data—often requiring simplifications to maintain model tractability—ML models thrive on large datasets, using the increased volume of data to enhance model performance. As the dataset size increases, ML models can leverage the additional data to refine their predictions and improve accuracy, whereas traditional methods may lose effectiveness due to overfitting or computational constraints (Khanzode and Sarode 2020, Sarker 2021, Dahiya et al. 2022).

However, the flexibility and capacity of ML models come with a trade-off: interpretability. One of the key challenges in using machine learning for demand modeling is the difficulty in explaining how the model reaches its predictions. ML algorithms, particularly ensemble methods, involve complex, multi-step processes where the final prediction is generated by combining outputs from various sub-models. These interactions between sub-models can be highly intricate, making it challenging to trace the decision-making process and understand the influence of specific input variables on the final output. While the ability of ML algorithms to handle non-linear relationships and high-dimensional data enhances their predictive power, it also makes it difficult to articulate the specific contributions of individual features to the model's predictions (Müller 2004, Olmedo 2016, Redyuk et al. 2019, Wanner et al. 2020).

In aviation, machine learning has been increasingly applied to operational tasks such as flight delay prediction. For example, Truong (2021) utilizes both supervised and unsupervised ML algorithms to predict flight delays in the United States, while Yu et al. (2019) apply Support Vector Regression (SVR) to forecast delays at Beijing International Airport. Similarly, Birolini et al. (2023) use ensemble machine learning techniques to predict primary delays using data from Vueling Airlines. However, the application of ML techniques to demand allocation, particularly in the context of itinerary choice models, remains relatively underexplored. In this domain, a few studies, such as Lhéritier et al. (2019) and Acuna-Agost et al. (2023), have applied random forests to model air-

line itinerary choice, and Delahaye et al. (2017) use Support Vector Machines to track consumer preferences. In the specific area of market demand forecasting, the use of ML algorithms is still in its early stages. While ML models offer high predictive accuracy, a key challenge in the aviation sector is not just achieving strong performance but also gaining a deep understanding of the underlying dynamics that drive demand fluctuations. In demand modeling, particularly for strategic forecasting, it is essential to not only predict future trends with accuracy but also to unravel the factors and mechanisms that contribute to changes in demand. This deeper understanding would enable more reliable forecasts and better decision-making for long-term planning in the aviation industry. As ML techniques continue to evolve, their integration into market demand forecasting has the potential to provide more accurate and actionable insights, but efforts to improve model transparency and explainability will be crucial for maximizing their utility in the context of aviation demand modeling.

2.1.3 Airline Planning Process

As outlined at the beginning of this chapter, aviation decision-making is deeply intertwined with demand forecasting across various time horizons. In the short term, forecasts optimize resource allocation, fleet management, and pricing strategies. Medium-term projections support route development, network planning, and competitive positioning, while long-term forecasts inform infrastructure investments and intermodal competition (e.g., HSR vs. air travel). This section focuses on the medium-term planning process, specifically on route planning, that is a key strategic decision shaped by demand projections.

Route planning is a crucial component of the strategic planning process for airlines, as it directly influences their ability to achieve long-term growth and competitive advantage. By determining which routes to operate, airlines can effectively expand their market presence, optimize their networks, and respond to shifts in passenger demand. These decisions ultimately shape the airline's overall success and are central to its positioning in the highly competitive aviation market (Mohri et al. 2022, Wu et al. 2022, Geursen

et al. 2023). In the context of network development, route planning is not just about launching new routes, but also about continuously evaluating and refining the existing network. Airlines must determine when to open new routes, close underperforming ones, or establish strategic connections between cities. These decisions are part of the broader network planning strategy that defines how the airline allocates resources to maximize profitability and market reach. The long-term impact of these choices is substantial, as the route network can significantly influence an airline's market share, operational efficiency, and the overall passenger experience. Even if its importance has been widely acknowledge, academic literature focused on network planning is quite scarce. In this domain, most academic research has focused on hub location and fleet planning problems (Mohammadi et al. 2019, Soylu and Katip 2019, Alumur et al. 2021), but besides hub location and fleet planning, airlines as part of their long-term strategy need to frequently evaluate which routes to open, close, or connect. At the state of the practice, most of the airlines identify which market to enter based on market analysis, basic econometric models of air travel demand, or simplistic route profitability models (Halpern and Graham 2015, Carmona-Benítez et al. 2017). These tools, while useful, tend to focus on short-term financial performance and may not fully capture the broader strategic implications of route network changes. Despite the critical importance of these decisions in shaping airline networks, route planning has received limited attention in the academic literature with only a few models tackling airline network expansion design (Kölker and Lütjens 2015, Birolini et al. 2021). The main reason is the high computational cost of network planning algorithms, which makes it difficult to apply them to real-world scenarios (Carreira et al. 2017, Schosser and Schosser 2020). This follows from the combinatorial complexity inherent in network planning, where the vast number of potential route combinations makes it challenging to develop efficient algorithms capable of identifying the optimal network structure. The few studies that address this issue attempt to tackle the combinatorial complexity, but this presents a major challenge even for relatively small-scale networks (Teodorović et al. 1994, Jaillet et al. 1996).

Different from network planning algorithms, the literature examining airlines' route entry decisions from an empirical perspective is quite substantial, focusing on the identification of factors that influence an airline's decision to enter new markets. These studies differ significantly from network planning algorithms, as they do not seek to optimize a network's structure but rather aim to explain and predict the actual behavior of airlines based on market conditions and competitive dynamics. The primary goal of these empirical studies is to uncover the underlying drivers of route entry decisions by examining market characteristics and the competitive environment in which these decisions occur. This empirical approach provides a deeper understanding of how and why airlines make route entry decisions in the context of real-world markets. Since the seminal study by Morrison and Winston (1990), which laid the foundation for this area of research, scholars have identified key drivers that influence route entry decisions. These factors have been broadly categorized into socio-economic conditions, market competition, and network considerations (Abdelghany and Guzhva 2010, Halpern and Graham 2015, Hanson et al. 2022). These categories of drivers are often closely aligned with the factors considered in demand forecasting, such as market size, population density, income levels, and growth projections, as well as competitive pressures such as the presence of incumbent carriers or alternative transport modes. The integration of these demand forecasting determinants into route entry decision-making reflects the interrelated nature of demand forecasting and strategic network planning. In the field of route entry decision, these factors are often airline- or airline group-specific and are shaped by airline strategy and market coverage (e.g., geographical or market segment focus) subject to fleet availability constraints (Oliveira 2008, Zhang et al. 2017b). Despite the substantial number of studies analyzing route entry decisions, there is a notable lack of consensus on which factors are the most important in shaping these decisions, and how these factors interact within the decision-making process. The diversity of findings in the literature can largely be attributed to the varying contexts in which these studies have been conducted. More importantly, likely due to the limited data availability, previous studies typically adopted a case-based ap-

proach focusing on analyzing single airlines or specific geographical contexts (e.g., single countries) (Fu et al. 2010, Calzada and Fageda 2019, Gaggero and Piazza 2021). These studies - often regional case studies (Dresner et al. 2015, Wang et al. 2017, Zhang et al. 2017b, Wang et al. 2022b) or carrier-specific case studies (Boguslaski et al. 2004, Aydemir 2012, Fu et al. 2015, Zou and Yu 2020) - typically aim to explain specific situations rather than uncovering broader, generalizable results. As a result, they fail to identify overarching patterns or trends that apply across different markets or airlines. Moreover, the case-based approach commonly used in route entry studies is limited by the availability and granularity of data. Most studies examine route entry decisions in markets where airlines have already established a presence, overlooking the crucial dynamics involved in entering new, unserved markets. First-time served markets, where airlines introduce direct, non-stop services for the first time, present a distinct set of challenges and opportunities that have been largely underexplored. These markets often differ in terms of demand characteristics, competition, regulatory environment, and infrastructure readiness. Only recently have studies begun to address the determinants of route entry into first-time served markets, providing important insights into the strategic considerations that guide airlines in these situations (Abdelghany and Guzhva 2022, Wong et al. 2023).

From this landscape it emerges how existing studies, while providing insights into carrier-specific dynamics or specific geographical and market contexts, are denoted by limited generalizability of the results and poor insights on a global scale, particularly in relation to the variables influencing route planning decisions. These studies ultimately fail in capturing the intricate dynamics of the industry and strategic considerations that underpin airlines' entry decisions across heterogeneous market environments. From an analytical standpoint, current methodologies lack a preliminary model capable of systematically narrowing the pool of potential routes before optimization algorithms are applied. By integrating demand-driven insights into network planning, airlines could refine their strategic decisions, ensuring that medium-term route development aligns seamlessly with long-term network expansion objectives.

2.2 Research Outline and Contributions

This thesis is grounded in the literature on air transport demand, addressing the various factors that shape demand across different empirical settings and time horizons. Table 1 summarizes the main contributions of each chapter. Each chapter shows the areas in which the most relevant contributions with respect to the literature have been implemented (whether in terms of dataset construction, methodological innovation, or key findings), as well as the time period covered by the empirical analysis and potential users who may benefit from the findings. The table should be read as a complementary tool with respect to Figure 2: while the figure presents the overarching conceptual framework within which the chapters are situated, the table clarifies the distinct contribution elements of each chapter, illustrating how they address different facets of the thesis as a whole. The main contributions of the present work to the state of the art are twofold, encompassing both methodological advancements and empirical insights. From a methodological perspective, this research makes two key contributions encompassing both demand generation and route selection strategy approach. First, it employs a demand generation model to systematically compare the performance of econometric and machine learning approaches. This analysis critically examines the predictive accuracy and interpretability of data-driven ML methods in air demand forecasting, exploring the extent to which black-box models can offer sufficient explanatory power for future applications. Second, this thesis develops a data-driven model to refine route selection strategies in order to optimize the allocation of new opening routes. The proposed framework provides practitioners with a systematic approach to identifying a targeted subset of potential new routes, facilitating network expansion and integration with airline planning algorithms. From an empirical perspective, this study conducts a comprehensive evaluation of air demand determinants and delve into a case study to analyze competitive dynamics between air transport and HSR. A substantial feature engineering effort has been undertaken, incorporating both well-established factors and a set of novel variables

that have been rarely explored in previous research. In addition, the study provides an in-depth assessment of substitution between air travel and HSR in a multimodal transportation context, along with a thorough evaluation of the projected impacts of HSR expansion plans on air transport demand. The final goal is to provide policymakers and researchers with a comprehensive framework for analyzing air demand dynamics across different time horizons, modeling approaches, and empirical contexts, offering a broad perspective on key challenges and future developments in the field.

	Time line	Dataset	Methodology	Results	Users
Chapter 3	Short	–	(i) ML interpretability (ii) Demand–supply two-way relation	(i) Improved accuracy (ii) Supply-demand stimulation	Operators
Chapter 4	Long	(i) Chinese HSR 2030 plan (ii) Socio-economic projections	–	Long-term pressure on aviation markets	(i) Policy makers (ii) Operators
Chapter 5	Medium	(i) Extensive data on new route candidates (ii) Balancing techniques	–	(i) Innovative features (ii) Practical tool	Operators

Table 1: Summary of contributions to the literature by chapter.

The thesis follows the "three paper format", where each contribution addresses a specific air transport challenge[7]. The first contribution (Chapter 3) focuses on air demand forecasting methodologies, addressing a key challenge in the field: the bidirectional relationship between demand and supply, which often compromises the accuracy of traditional econometric models. This study conducts a methodological assessment of widely used econometric techniques, comparing their performance with various ML approaches. The analysis first evaluate the performance of ML algorithms in presence of demand-supply two-way relationship. Second, it examines and discusses the trade-off between predictive

[7]For the sake of completeness, all machine learning and econometric models were implemented in Python (version 3.14.2) using well-known libraries: scikit-learn for machine learning algorithms, pandas and numpy for data handling, and statsmodels for traditional econometric models, alongside other libraries for minor implementations. Model training was performed on a workstation with an Intel i7 processor and 32.GB RAM, with typical computation times ranging from a few seconds for individual regression models to several minutes for full Random Forest training, depending on dataset size and hyperparameter settings.

accuracy and interpretability in the ML domain, critically examining whether its adoption can be a valuable alternative in air demand forecasting, where explainability is a crucial factor for both research and policy applications. The second contribution (Chapter 4) examines the air-HSR substitution effect, focusing on the impact of high-speed rail expansion over the next two decades within a specific geographical context (i.e., China). The central challenge is to accurately assess how upcoming HSR developments will reshape air travel demand and identify the markets most vulnerable to competitive pressure. To achieve this, a demand model is integrated into a scenario-based analysis, incorporating a tailored set of determinants and socio-economic projections to provide a forward-looking assessment of air transport dynamics in response to HSR growth. The third and final contribution (Chapter 5) tackles a critical challenge in airline planning: the need for carriers to continuously evaluate, refine, and expand their networks in a systematic manner. Existing prescriptive network planning algorithms face computational constraints that limit their ability to assess a comprehensive set of potential route candidates on a global scale. To address this limitation, we propose a model that enhances route selection by acting as a "route-scouting" tool. This model serves as a pre-assessment mechanism, streamlining the identification of promising routes and enabling seamless integration with the prescriptive methodologies used by airlines. The three papers composing this thesis are thoroughly described in the following.

Chapter 3: *Forecasting non stop dominated markets using advanced predictive analytics*

The aviation industry relies heavily on accurate demand modeling to enable effective resource management and planning. Traditionally, econometric techniques, especially gravity-based and time-series models, have been widely used, offering the advantage of producing interpretable results and supporting causal analysis. However, assumptions about their functional forms and limitations in available techniques for addressing endogeneity can significantly undermine predictive performance. In recent years, machine learning (ML) methods have attracted attention due to their flexibility and ability to

identify complex patterns and leverage large datasets. While ML has been extensively applied to support operational tasks within the aviation sector, its use in demand modeling remains relatively limited. The study develops and compares ML methods (random forest, gradient boosting, and KNN) with traditional econometric approaches (OLS, first-difference estimator, and 2SLS). Although ML methods are inherently less interpretable, the results indicate that they deliver substantially higher predictive accuracy than econometric models (around +30%). Furthermore, ML models demonstrate high generalizability, as shown through out-of-sample validation demonstrating the ability to accurately estimate demand-frequency elasticities. This paper contributes to the literature first by investigating the potentiality of data-driven ML methods to enhance forecasting accuracy in the context of the complex, two-sided relationship between demand and supply. Second, using a tailored case study centered on a growing supply scenario, the analysis demonstrates that ML algorithms are more effective than traditional approaches in capturing real-world dynamics and underlying demand–supply elasticities, as evidenced through graphical comparisons. Ultimately, these findings are discussed in order to provide robust insights and increased awareness into the interpretability vs accuracy trade-off and the generalizability of state-of-the art ML, which is key for their broader adoption of ML for aviation demand forecasting.

This paper was co-authored by Sebastian Birolini (Bergamo University). The research was presented at the European Working Group on Transportation (EWGT), 2023.

Chapter 4: *Air markets network design under high-speed railway development: Empirical evaluation of air travel demand affected by the new Chinese High-Speed Rail lines opening in 2030*

High-speed rail is increasingly challenging the domestic air transport operations in China due to its competitive edge in frequency, better accessibility from city centers, and overall convenience for short to medium distance travels. This increasing demand changes will become even more extensive in the next decades with laid out high speed rail development plans. By the end of 2020 the railway network of China had expanded

to include around 40.000 km of HSR lines (equal to 2.2 times the extent of 2013) and the milestone for 2030 until 2050 is to reach more than 70.000 km of HSR lines. Accordingly, airlines are under increasing pressure to adapt their networks in order to be efficient and profitable. In this study, we develop a demand forecasting model with the goal to analyze the historical effect of increased high-speed railway penetration on air markets passengers flows from 2012 to 2019 and predict what will be the impact for the domestic aviation sector for the next two decades with the new railway lines planned to be opened by 2030. We propose and compare different econometric specifications, highlighting the advantages of using a declination combining a semi-logarithmic OLS to predict the total market demand, reaching a predictive accuracy of around 80% when validating the model on an unseen set of test data. Specifically, our analysis evaluates the magnitude of air demand shift towards rail in city pairs that will benefit from new HSR lines opening in 2030. Our results show that for the majority of cities connected below 1000 km, the air demand will likely disappear by 2050, while for cities connected between 1000 and 3000 km of distance the air demand will substantially decrease but without completely disappearing. This paper makes a novel contribution to the literature by incorporating new data on the upcoming HSR lines in China, as outlined in the Long-Term Railway Network Plan published by the Chinese government. It also analyzes and integrates the most recent socio-economic projections for the next two decades, as presented in existing literature. Additionally, the paper offers innovative policy decision-making support by identifying aviation markets most likely to face pressure from the introduction of new HSR lines, and by estimating the timing of these shifts in demand. Our research contributes to the discourse on sustainable transportation planning, highlighting the potential implications for policymakers, urban planners, and transportation industry stakeholders.

This paper was co-authored by Xiaoqian Sun (Beihang University, China) and Sebastian Birolini (Bergamo University, Italy). The research has been developed during the visiting period at the School of General Engineering at Beihang University. The early version of the paper was presented during the Air Transport Research Society -

China Chapter in 2024, while the complete version of the paper was presented during Air Transport Research Society Main Conference in 2025. The manuscript is under revision to Transportation Research Part A: Policy and Practice (R/R).

Chapter 5: *Airlines network development practices: An extensive worldwide empirical application*

This paper aims to empirically examine how airlines strategically decide on entering new long-haul markets. The proposed model -rooted on a comprehensive and innovative dataset- not only sheds light on key determinants driving the market entry decision, but it also can be used by aviation operators as practical tool. In this context, it plays a pivotal role in evaluating the network development potentials for each carrier, allowing scalability to be integrated with network planning algorithms to promote refined route selection strategies. We employ classification methods, distinguishing between pooled and carrier-specific formulations, and analyze a comprehensive data set spanning six years (2014-2019) on a global scale. To deal with a highly unbalanced data set (i.e., operated routes vs potential routes), various balance sampling procedures are tested and compared, demonstrating significant benefits on the results. We engineer features to capture essential drivers, including market potential, competition, network dynamics, and more, to reveal the fundamental factors influencing the airlines' network development decisions. Ultimately, the model is validated with a tailored out-of-sample procedure to demonstrate its capability to inform network planning strategies in practice and support refined route selection algorithms. To summarize, the primary contribution of this study to the existing literature is the construction of an extensive dataset that incorporates a broad range of potential new connections, ensuring a balanced training modeling phase to produce robust and informative coefficients. The second key contribution lies in providing a comprehensive analysis of the complex factors influencing carrier entry decisions. This includes the development of cutting-edge network-related determinants -among others- and a critical exploration of the similarities and differences across carriers. Lastly, this study demonstrates how data-driven decision support systems can effectively refine po-

tential route candidates within a carrier-specific framework. While previous research on entry decisions has predominantly focused on specific contexts, long-haul markets and potential new route developments have received comparatively less attention. This work aims to fill that gap by leveraging an extensive dataset, combining econometric analysis with robustness checks, and validating results through tailored out-of-sample tests.

This paper was co-authored by Nicolo' Avogadro (Bergamo University) and Sebastian Birolini (Bergamo University). The seminal version of the paper was presented during the Air Transport Research Society Main Conference in 2022, while the complete version was presented during the Air Transport Research Society Main Conference in 2023. Currently the paper is under revision in Journal of Air Transport Management.

Chapter 3

Forecasting Non Stop Demand Using Advanced Predictive Analytics

3.1 Introduction

The provision of air transportation services is fundamentally driven by air travel demand. Comprehensive analysis of past, present and future drivers of air travel demand is critical for the aviation industry to effectively guide capital investments, anticipate workforce needs, budget accurately, inform tactical and operational decisions, such as resource scheduling and allocation. To meet these needs, researcher and practitioners have continually advanced demand prediction methods (Wang and Gao 2021, Nicholas 2021), integrating them with prescriptive optimization models to support diverse decision-making across industry stakeholders, including manufactures (Silva et al. 2019), air traffic controllers (Starita et al. 2020), airlines (Lundaeva et al. 2024) and airports (Amadou et al. 2021). A critically distinguishing feature of this prediction task is the incorporation of key supply-related predictors. More precisely, demand and supply are intricately linked in a two-way relationship because of the dynamic interplay between the number of passengers (demand) and the availability of flights and routes (supply)[8]. When demand for air travel increases, airlines respond by adding more flights, expanding routes, or increasing the capacity of existing services, thereby enhancing supply. Conversely, when airlines introduce new routes, increase flight frequencies, or offer improved services, they can stimulate additional demand by making air travel more accessible and attractive.

Most studies in the literature, also explored in Section 2, have traditionally employed

[8]For a more comprehensive discussion regarding the complex relationship occurring between demand and supply and the related methodology to handle it, please refer to Section 2.

econometric techniques, often relying on a gravity formulation(Grosche et al. 2007, Adler et al. 2018, Tirtha et al. 2023) combining geo-economic characteristics of the origin and destination – such as population, GDP - and service-related determinants which can be influenced by the airlines – such as frequency, aircraft size -. However, from an empirical standpoint, the interplay between demand and supply creates a reverse causality problem, potentially leading to biased estimates. A common approach to overcome this issue is to resort to instrumental variables techniques (Hsiao and Hansen 2011, Boonekamp et al. 2018, Birolini et al. 2021). Despite the theoretical soundness of this approach, its practical implementation is not straightforward and practically challenging. The main complexity is to identify suitable instruments, namely control variables that are correlated with the variable to be instrumented (e.g., supply) but not with the dependent one (e.g., demand). Many studies in the literature have indeed acknowledged the problem of weak instruments in air transport demand modelling (Brueckner et al. 2014, Suh and Ryerson 2019). This difficulty -most of the times- ultimately translates into an arguably inadequate treatment of endogeneity and lower prediction accuracy due to the fact that a strong predictor is replaced with a combination of weaker proxies.

In comparison with the above-mentioned methods, in more recent times machine learning techniques have received increased attention due to their ability to handle complex relationships and hidden dependencies within data[9]. Such models offer an advantage over traditional econometrics methods by flexibly adapting to data patterns. This adaptability makes machine learning a valuable tool for predictive tasks while mitigating some critical aspects of parametric methods. Regarding aviation industry, they have been employed for modelling operational tasks. They have been mostly applied to allocation problems and flight delay predictions (Yu et al. 2019, Xu et al. 2022, Birolini and Jacquillat 2023). Some recent contributions have explored the use of ML for itinerary choice and airline market share studies (Lhéritier et al. 2019, Acuna-Agost et al. 2023, Birolini et al. 2023). Nevertheless, their application in demand modelling is less common and the

[9]An overview discussion on the topic can be found in Section 2.1.2

broader use in the aviation context is somewhat limited. ML models exploit advanced algorithms that construct more complex structures that can effectively fit the training data while retaining their ability to generalize to unseen data. Those algorithms represent a relatively recent and rapidly evolving approach, demonstrating advantages in enhancing prediction accuracy and handling with large multidimensional dataset. They are not tied to specific patterns, giving them the flexibility to uncover and explore more complex, non-linear relationships within the data, bypassing the constrained functional forms and distributional assumptions inherent in parametric models. This adaptability allows machine learning algorithms to better approximate intricate connections among variables during the training process. On the other hand, the greater complexity of machine learning models represents a challenge in terms of interpretability. Those models are often referred to as "black box" due to the difficulties in evaluating and visualizing the relations behind their predictions. Therefore, while machine learning algorithms excel at capturing complex relationships within data, they may lack transparency regarding the causal mechanisms governing these relationships. The lack of interpretability has thus far constrained the utilization of ML models, especially in contexts where a comprehensive understanding of supply-demand dynamics is a key component, such as demand prediction. In various demand modelling applications within the aviation sector, it is indeed not sufficient to pursue strong predictive performance; it is also important to validate and get a deep understanding of the underlying factors influencing demand and how these determinants are affecting the outcome factor.

In this framework, our paper aims to develop and explore the potential of data-driven methods for predicting total market traffic demand incorporating the supply-side stimulation effect. To achieve this goal, we assemble an extensive global dataset encompassing monthly data for over 20.000 non-stop city-pair routes from 2013 to 2019. This huge sample allows us to concentrate on the stimulation effect of frequency and demand over a large window time. First, the paper implements and compares three different econometric approaches (OLS, first difference estimator, 2SLS), discussing how different variables

combinations (i.e., by considering only time-variant factors or also time-invariant) can provide different overviews of the phenomena. In particular, we demonstrate that by incorporating time-invariant socio-economic factors the model is able to capture the heterogeneity effect within markets, without losing fitting accuracy. Second, we propose three supervised ML algorithms and compare their fitting performance against econometric models through several KPIs. Our findings reveal that ML models achieve higher prediction accuracy, around +30%. After evauating our best ML algorithm, we apply a SHAP feature importance method in order to assess which factors affect the most the final prediction, thus contributing to the model's overall explainability. It reveals that ML can unveil meaningful supply-demand patterns, being able to characterize the role of key variables as expected. Third and last, we develop a tailored validation and investigation of the mechanisms underlying the predictive ML process. By testing econometrics and ML models in a context of growing supply we measure how much the two approaches are aligned with real-world trends in capturing the elastic response of demand. Findings reveal that ML model is more accurate in predicting the stimulation effect of the supply over the demand, while econometrics model tends to overestimate the predictive pattern. Based on that, ML impose itself as valuable approach in providing highly accurate demand forecasts, modelling the stimulation effect between variables aligned with the benchmark level (i.e., real-world observations).

The paper is structured as follows: Section 3.2 provides an overview of the literature on passenger forecasting models and methodologies employed so far, Section 3.3 describes the data used and empirical models estimated, Section 3.4 presents the results of the models and a tailored validation experiment, Section 3.5 concludes.

3.2 Related Literature

Forecasts on air travel demand are important inputs for a wide variety of economic decisions, including but not limited to, research and development, airplane design, production planning and the accurate modelling of demand plays a crucial role in enhancing the over-

all level of passenger services (Abed et al. 2001, Carson et al. 2011, Trapero et al. 2011). For this reason, identifying contributing factors and understanding their effect in causing the variation of air travel demand have been one of the key focus areas in air transport research. Specifically, it is essential to consider the intricate relationship between demand and supply. It is widely acknowledged that demand and supply establish a two-way causal relationship (Hsu and Wen 2003, Hsiao and Hansen 2011), posing challenges for traffic volume estimates.

The majority of studies resort to econometrics or statistics-based analytical models to investigate the effect of predictor variables on air travel demand. Generally speaking, all the econometric models employed belong to the field of parametric methods. The main advantage is to have a clear understanding of the stimulation effect of the predictor variables, making them easy to interpret. Parametric methods, including econometrics formulations, have a well-defined specification of the functional form (e.g., linear, logarithmic) that represents the relationship between the dependent and independent variables[10].

One of the most successful formulations is the OLS gravity model (Casey Jr 1955, Grosche et al. 2007). All the studies employing a gravity model formulation include a set of attractor factors depend on geographical and socio-economic characteristics (demand-side variables) such as income and population (Jorge-Calderón 1997, Valdes 2015, Tirtha et al. 2023), whereas supply-side variables are used to model the difficulty of traveling between markets, such as frequency, air fares, market shares (Mueller 2015, Scotti and Dresner 2015, Hazledine 2017). Few other studies also embrace tourism-related factors (Koç and Arslan 2018), monthly salary (Li and Sheng 2016), airline yield (Gosling and Ballard 2019), even if their inclusion is scanter as often accurate data are unavailable.

However, while the significance of supply-related variables, especially flight frequency, is acknowledged, the presence of a two-way relationship in parametric models complicates the resolution of reverse causality. This challenge creates a trade-off between methodolog-

[10]This explicit structure aids in understanding the expected impact of predictor variables on the outcome, as they correspond to the marginal impact of a unit change in an independent variable while holding other variables constant (Wooldridge 2010)

ical rigor and accuracy, ultimately undermining the forecasting reliability of the models (Zhang et al. 2017b, Birolini et al. 2020). Departing from this limitation, some contributions to the air transportation literature have analysed methods to approximate the demand-supply relationship without incurring in an endogeneity problem. The most used approach is to rely on instrumental variables formulation, by modelling the supply-side with factors that are correlated with the level of service (the independent factor) but uncorrelated with the volume of passengers (the dependent factor). Hsiao and Hansen (2011) and Hazledine (2017) use as instrument the fuel price, Boonekamp et al. (2018) the average aircraft size and feeder value, Birolini et al. (2021) the average frequency operated on flight legs of similar distance. Nevertheless, a first complexity is that these studies sacrifice accuracy in their attempts to correct for endogeneity, resulting in diminished predictive performance. The second complexity with this approach is to find suitable instruments; indeed, often weak instruments compromise the reliability of estimates (Hahn and Hausman 2003, Keane and Neal 2024). This limitation shed light on the low explanatory power in predicting the dependent variable, affecting the instrument's ability to effectively isolate the variation in the endogenous component. Such difficulty limits the applicability of parametric methods in addressing the demand-supply interrelation, as they are easily interpretable, but with the risk of biased estimates that are not able to disentangle the two-way stimulation effect and provide reliable models.

In addition to parametric methods, non-parametric approaches, particularly machine learning algorithms (ML), have garnered increased attention in recent years for real-world applications. The primary advantage of employing non-parametric methods in demand modeling is their flexibility. While econometric methods often rely on fixed model forms and may struggle with reverse causality issues, machine learning algorithms can more effectively capture complex patterns in the data without being limited by initial assumptions about the causal structure. As a result, these techniques have the potential to enhance accuracy and provide valuable insights in demand modeling, prioritizing learning from the data over fitting a specified model framework. The absence of a functional

mathematical form allows for more flexibility in the modelling, but the main drawback is that these algorithms typically require a huge dataset to create the model. The higher flexibility let those algorithms able to explore complex patterns within data, by capturing intricate relationship. While traditional econometrics models often assume linearity, ML models can handle non-linear pattern, enabling more accurate representation of the phenomena. These characteristics have demonstrated the better accuracy performance of ML techniques (Khanzode and Sarode 2020, Sarker 2021, Dahiya et al. 2022). Nevertheless, it comes at a price of increase in the complexity, which makes it difficult to interpret the models in comparison to the conventional statistical approaches.

The difficulty in explicability of ML models stems from their inherent complexity. The algorithms consist of multiple steps (i.e., in the application of ensemble methods, the final result is generated by combining predictions from various sub-models) and the interactions between them can be intricate, making it challenging to trace the decision-making process leading to the final prediction outcome. In general, the advantages of handling non-linear relationships and high dimensionality of data enhances their predictive power, but also makes difficult to articulate how specific input features contribute to the output (Müller 2004, Olmedo 2016, Redyuk et al. 2019, Wanner et al. 2020).

Some studies have gradually implemented machine learning techniques to aviation problems, but in market demand forecasting's domain the employment of ML algorithms has been still under-investigated.

Based on the contributions outlined in Section 3.1, our paper proposes a systematic attempt in investigating the potentiality of data-driven approaches, powering the use in market demand forecasting field. It shows how applying ML algorithms can help in providing more accurate demand estimations in a context of growing supply.

3.3 Data and methods

In this section, we begin by outlining the sample collection procedure and providing an overview of all the variables employed in our empirical analyses. Afterward, we explain

the model specifications, covering both econometrics and machine learning approaches.

3.3.1 Data sources

To develop our study a comprehensive dataset has been assembled from different data sources with worldwide information spanning from 2013 to 2019, retrieved at monthly level. Data on origin-destination (OD) passenger flows were collected from OAG Traffic Analyser[11]. Observations in our dataset are grouped by city pairs, where each city may contain multiple airports (multi-airport areas). Aggregating airports within the same metropolitan area is crucial, as it simplifies the analysis by avoiding the need to account for the overlap and competition effects between different airports in the same region. This aggregation allows for a more streamlined analysis without sacrificing the significance of the results. We consider only non-stop itineraries and select those markets in which the level of connecting passengers is lower than 5%, which represents about 8% of total OD pairs served during the aforementioned time period but around 75% of the total volume of non-stop passengers. This sample identification allows us to narrow down the stimulation effect of the total supply over the non-stop demand. Each observation is completed with detailed data on transport supply sourced from the OAG Schedule Analyser.

Additional filtering rules are applied to ensure that the empirical analysis is based on reliable and economically meaningful observations. Very short-haul itineraries (i.e., distances below 500 km) are excluded. While competition from high-speed rail is one consideration (Lieshout et al. 2016), the main rationale is methodological: on very short distances, air travel volumes are often highly volatile, dominated by irregular operations, and influenced by non-price competitive factors (surface transport accessibility, airport congestion, or ground travel substitution), making them unsuitable for stable frequency–traffic modelling. Similarly, we retain only observations for markets serving at least 100 monthly passengers with a minimum frequency of 10 flights, to avoid noise from

[11]Official Airline Guide (OAG) is a leading provider of aviation data and analytics, widely used by airlines, airports, and travel-related businesses. Their dataset provide comprehensive information on airline schedules, flight status, passengers flows, among others.

marginal or irregular services.

After filtering and cleaning, the resulting sample consists of 753.085 monthly OD observations corresponding to 22.977 unique market pairs worldwide distributed. Based on the partition between long haul (LH) and short haul (SH) with a threshold of 3000 km, 82.2% and 17.8% are designed as SH and LH (respectively, divided in 18.270 and 4.707 unique city pairs). The choiche to set the long haul threshold at 3000 km follows standard industry practice and prior literature, as this cutoff generally reflects the transition between wide-body and narrow-body operations, network structure differences, and distinct competitive environments[12].

Ultimately, the sample has been finalized with geo-economic data. They are computed using a Geographic Information System (GIS). Starting from airports coordinates we define a catchment area with a fixed radius of 100 km and calculate the population and GDP values based on global spatial dataset taking 2019 as reference year.

3.3.2 Variables definition

In our empirical models, the dependent variable is denoted by D^t_{od}, where we index origin and destination market by o and d, time periods by t. Forecasting air passenger flows is a challenging prediction problem. The potential demand on each route depends on the level of service provided, but it is also subject to temporal fluctuations determined by trend and seasonality components. As shown in Table 13, variables used for the model(s) development can be grouped into socio-economic factors, geographical and service-related.

Socio-economic variables characterize the passenger pool at each origin and destination area. GDP per-capita is considered as a representative of economic activities, which are the fundamental driving force behind air travel demand. Intuitively, travel demand from an area is also associated with the population of that area, showing a direct relation-

[12]This segmentation is applied not only descriptively but also in model estimation. Although ML algorithms can handle heterogeneity, combining SH and LH routes in a single model would require jointly learning fundamentally different demand–supply mechanisms, potentially obscuring relationships and reducing interpretability. Segmenting the dataset allows each model to capture the dynamics specific to its operational context.

ship. Consequently, the probability of people travelling between two densely populated and economically strong regions is higher than between more peripheral regions. These variables are a proxy for the market potential and they have been estimated by taking the product of these two magnitudes. We define a variable to capture the degree of seasonality at time t as well as the presence of peak months. It is expressed as the ratio between the number of passenger travelling in a given itinerary during a month and the total number of passengers travelling in the same itinerary during the whole year, at time *t-1*.

Distance is expected to be an impedance factor for air travel. Within shorter distances there is more travelling demand but also more competition from other modes of transport; on the other hand, when distance increases, demand decreases. Meaning that, for longer distances people are less willing to travel. The variable has been defined with the great circle method (thousand of kilometers) between the origin and destination points, considering also the squared distance measure.

Service-related variables play a crucial role in shaping the supply decisions of airlines and subsequently influencing the quality of service provided. Two primary variables, frequency and prices, are commonly utilized to assess the level of service, and extensive research in the literature has explored these aspects (Chen et al. 2012, Scotti and Dresner 2015, Hsu et al. 2013, Hakim and Merkert 2019). Frequency is defined as the number of flights operated along the OD segment at a specific time (t). Developing an integrated demand-supply model necessitates the inclusion of frequency, as it allows not only for the capture of interrelations on the service side, but also for the dimensions of different markets. Indeed, with the frequency variable we are also able to investigate the heterogeneity across markets supply. Prices also hold a pivotal role in influencing demand (Deese et al. 2013, Mueller 2015); however, their consideration requires a not straightforward evaluation due to the challenges associated with obtaining reliable data sources. In our analysis, we chose to omit prices from the model due to the unavailability of appropriate data. Nonethless, we directly include a proxy of market concentration

defined as the Herfindahl-Hirschman Index (HHI). It provides a means to measure the market structure by summing the squared market share percentages of individual airlines and is particularly important in the aviation context as it provides insights into the level of competition within a given market. It assumes values ranging from 0 to 1 where 0 is the perfect competition and 1 is a monopolistic market; a higher HHI indicates a more concentrated market with fewer dominant airlines, which can influence pricing, service quality, and market entry barriers, in contrast, a lower HHI suggests a more competitive market, where passengers might benefit from more choices and competitive pricing. Using HHI helps us understand the competitive dynamics that can impact air travel demand and the overall efficiency of the market.

Notably, the absence of pricing information within the model significantly affects the interpretation of the coefficients. In particular, assuming that prices depend on the level of market concentration and frequency (Vowles 2006, Wang et al. 2019), the effect of prices on demand will be implicitly reflected into the HHI and frequency coefficients[13].

We consider the growth rate such that it captures the percentage increment (or decrement) of available seats in a given itinerary between time *t-2* and *t-1*. It helps in identifying markets in which a growing supply pattern could denote a potential demand not completely satisfied, offering opportunities for airlines to expand their service. Additionally, we consider two variables that could be used as proxies for the level of frequency

[13]To clarify this concept, let's define the demand as:

$$D^t_{od} = f(.) + \beta_1 P^t_{od} + \beta_2 \text{freq}^t_{od} + \beta_3 \text{HHI}^t_{od} \tag{1}$$

And let the price be defined as:

$$P^t_{od} = g(.) + \delta_1 \text{freq}^t_{od} + \delta_2 \text{HHI}^t_{od} \tag{2}$$

Substituting 2 into 1, we obtain:

$$D^t_{od} = f(.) + g(.) + (\beta_1\delta_1 + \beta_2)\text{freq}^t_{od} + (\beta_1\delta_2 + \beta_3)\text{HHI}^t_{od} \tag{3}$$

The frequency stimulation effect has to be interpreted as an indicator of the average demand increment corresponding to an increase in flight supply, under the assumption of an equilibrium pricing response aligned with historical patterns. In particular, considering a potential reduction in prices following an increase in supply ($\delta_1 < 0$), the stimulation effect of the frequency elasticity has to be interpret as the combined impact of higher (lower) supply (β_2) and lower (higher) prices ($\beta_1\delta_1$). Similarly, the HHI, will capture the impact of lower (higher) prices induced by higher (lower) market concentration (i.e., β_3).

(that is, as instruments in the 2SLS model). First, variable accounting for the average number of seats available on each aircraft in a given itinerary is included. This metric is important because it reflects the capacity airlines allocate to specific routes, influencing both supply and passenger convenience. Larger aircraft with more seats can accommodate more passengers, potentially leading to greater demand. Second, we incorporate the number of spokes (feeder value) to approximate the scale of hub operations at both the origin and destination points. This variable is defined as the number of direct flight connections from a given point a to the origin o or from the destination d to a. In the aviation context, a higher number of spokes indicates a more extensive hub-and-spoke network, which can enhance route connectivity, increase passenger flow through hub airports, and improve the overall efficiency of airline operations. This, in turn, can stimulate demand by offering more travel options and reducing travel times for passengers. We consider the percentage of flights operated by a low-cost carrier over the total on a given route OD. Ceteris paribus, we expect a positive relation between OD demand and the presence of LCC.

Table 8 lists descriptive statistics related to the variables collected. The empirical distribution of OD air passenger flows is highly skewed and varies within the sample. On average, each itinerary connects about 10.000 passengers each month, with about 1300 km as distance and flights supply of 76. From the HHI value we can see that the market is in a competitive and less concentrated environment. The high presence of LCCs is reasonable when dealing with markets dominated by local passengers, as LCCs tend to connect relatively smaller markets in comparison to full-service carriers.

3.3.3 Econometric formulation

Regarding econometric techniques, we employ three distinct models, widely explored in the literature: (i) First difference estimator with log-linearization; (ii) Ordinary Least Squares (OLS) gravity model with log-linearization; (iii) Two-Stage Least Squares (2SLS) gravity model with log-linearization.

	Mean	St dev	[min, median, max]
Passengers (monthly)	9.854	15.643	[101, 4.544, 36.637]
GDP per capita ('000, $)	39,249	11,278	[14,256, 37,114, 67,118]
Population ('000)	11,700	13,110	[1,265, 6,999, 68,490]
Distance (km)	1.370	1.090	[500, 1.091, 14.000]
Seats (monthly)	12.216	18.075	[145, 5.953, 338.245]
HHI	0,025	0,024	[0,003, 0,023, 1]
LCC share	0,42	0,28	[0,02, 0,5, 1]
Feeder value	136	143	[4, 76, 490]
Frequency (monthly)	76	97	[145, 5.953, 1.259]
Load factor	0,6	0,74	[0,1, 0,71, 1]

Table 2: Descriptive statistics

Category	Variable	Description and formulation
Distance	Route distance	Route great-circle distance between origin and destination point
Market Potential	Population	Let pop_o be the total population within the origin airport area and pop_d the total population within the destination airport area, both assumed with a static 100 km radius area. We define the population on route OD as $pop_{od} = \sqrt{pop_o \cdot pop_d}$
	GDP per capita	Let gdp_o be the GDP per capita within the origin airport area and gdp_d the GDP per capita within the destination airport area, both assumed with a static 100 km radius area. We define the GDP per capita on route OD as $gdp_{od} = \sqrt{gdp_o \cdot gdp_d}$
	Seasonality	Let $pax_{od}^{(m,t-1)}$ be the number of passengers in given month and time, and let $pax_{od}^{(y,t-1)}$ be the number of passengers in a given year such that $m \in y$, then we defined the seasonality component as the ratio between the number of passengers travelling in a given month and the total number of passengers travelling during the year: ${seas_{od}}^t = \frac{pax_{od}^{(m,t-1)}}{pax_{od}^{(y,t-1)}}$
Service related	HHI	Herfindahl-Hirschman Index based on the no-stop seating capacity. Let $\mathcal{A}$ be the set of airlines offering no-stop services on a route and ϑ_a the number of allocated seats by airline a, we define the index as $HHI = \sum_{a \in \mathcal{A}} MS_a^2$ where $MS_a = \frac{\vartheta_a}{\sum_{a \in \mathcal{A}} \vartheta_a}$ is the market-share (in terms of offered seats) of airline a
	Growth rate	Percentage of the total seats growth on a given itinerary between time $t-1$ and $t-2$ defined as: ${growth_{od}}^t = \frac{{seats_{od}}^{(t-1)} - {seats_{od}}^{(t-2)}}{{seats_{od}}^{(t-2)}}$
	LCCs share	Let $carr_{LCC,od}^t$ be number of low cost carriers in a given time and $carr_{FC,od}^t$ the number of main carriers in a given time, operating a given OD route. Then the share of LCCs is defined as: $lcc_{od}^t = \frac{carr_{LCC,od}^t}{carr_{LCC,od}^t + carr_{FC,od}^t}$
	Frequency	Number of flights offered on OD route
	Number of spokes	For a given route from hub h to origin o or from destination d to hub h, let S_h be the set of spokes s ($s \in S_h$) connected to/from the hub with at least one weekly flight. We define the number of spokes as $S_{ho/dh} = \sum_{s \in \mathcal{S}_h}$
	Average size	Average number of seats available on each aircraft in a given itinerary, defined as: ${AvgSize_{od}}^t = \frac{{seats_{od}}^{(t-1)}}{{freq_{od}}^{(t-1)}}$

Table 3: Variable definition and formulation.

The first difference estimator is used in the context of time-series analysis to examine the change in variables over time (Wooldridge 2010). By calculating the difference between consecutive observations and the impact of changes over time, it helps in mitigating the potential biases originating from fixed factors. In this context it's important to note that excluding time-invariant factors can lead to a loss of information, as we are not considering fixed heterogeneity across observations. Therefore, in using this estimator is essential to evaluate the trade-off between time-trends identification and fixed relationships within the data. Following the form:

$$\Delta(\ln D_{od,t}) = \Delta(\ln x_{od,t}) + \Delta(\ln u_{od,t}),$$

we adopt the specification[14]:

$$\begin{aligned}\ln\left(\frac{D_{od}^{t}}{D_{od}^{t-1}}\right) = \beta_1 \ln\left(\frac{HHI_{od}^{t}}{HHI_{od}^{t-1}}\right) + \beta_2 \ln\left(\frac{seas_{od}^{t}}{seas_{od}^{t-1}}\right) + \beta_3 \ln\left(\frac{growth_{od}^{t}}{growth_{od}^{t-1}}\right) \\ + \beta_4 \ln\left(\frac{LCC_{od}^{t}}{LCC_{od}^{t-1}}\right) + \beta_5 \ln\left(\frac{freq_{od}^{t}}{freq_{od}^{t-1}}\right) + \left(\epsilon_{od}^{t} - \epsilon_{od}^{t-1}\right) \end{aligned} \tag{4}$$

The OLS gravity model with log-linearization states that the demand between two points (origin and destination) is given by the presence of factors of attractiveness and impedance and it has been widely employed in the literature (Jorge-Calderón 1997, Grosche et al. 2007, Birolini et al. 2021). The gravity model used to describe passenger demand implies a non-linear multiplicative specification of the form $y = \prod_{k=1}^{n} x_j^{\beta_j}$, which is estimated by taking the logs of both sides, i.e. $y = \sum_{k=1}^{n} \beta_j \ln(x_j)$, and fitting linear regression models. The following model specification is estimated:

[14]Following the logarithm's property the first difference of the log-transformed variable is:

$$\Delta \ln(y_t) = \ln(y_t) - \ln(y_{t-1})$$

This can be interpreted as:

$$\Delta \ln(y_t) = \ln\left(\frac{y_t}{y_{t-1}}\right)$$

$$\ln(D_{od}^t) = \beta_0 + \beta_1\, dist_{od} + \beta_2\, dist_{od}^2 + \beta_3 \ln(gdp_{od}) + \beta_4 \ln(pop_{od}) + \beta_5 \ln(HHI_{od}^t) + \beta_6 \ln(inc_{od}^t)$$
$$+ \beta_7 \ln(growth_{od}^t) + \beta_8 \ln(LCC_{od}^t) + \beta_9 \ln(freq_{od}^t) + \epsilon_{od}^t \quad (5)$$

Geographical fixed effects are considered in the models. This statistical technique helps us in controlling for location-specific factors that may affect the dependent variable. This ensures that the analysis is not biased by variations in the data caused by differences in locations, where the location is determined by the continent[15] of origin and destination.

As highlighted in Section 3.1, there is a two-way causal relationship between demand and frequency, which leads to a reverse causality problem. To correct for this endogeneity, a two-stage least square (2SLS) model with instrumental variables is estimated (Wooldridge 2010). The main challenge in using this model is to find suitable instruments that are correlated to frequency but uncorrelated to passenger demand. The purpose of including the 2SLS model is not to improve predictive accuracy or to compete with ML models in forecasting performance. Rather, it is to provide a framework for obtaining less biased estimates of the average causal effect of frequency on demand. In other words, 2SLS is used to achieve causal inference, correcting for endogeneity, whereas ML models are designed to maximize prediction accuracy. Consequently, it is expected that ML models outperform 2SLS in prediction. The comparison between ML and 2SLS should therefore be interpreted carefully: ML highlights patterns and associations in the data, while 2SLS allows for inference about causal effects, and the two approaches serve complementary, but fundamentally different, purposes. Following Boonekamp et al. (2018), we use as instruments the feeder value and the average aircraft size. The feeder value is defined as the number of connections (spoke flights) per direct flight. This is a proxy of the magnitude of the hub operation at origin and destination areas. It is a valid instrument because flight frequencies to and from hub points are likely to be higher than to and

[15]Defined as region by OAG database.

from non-hub points, ceteris paribus since transfer passengers fill up a significant share of the available seats. The other instrumental variable used is the average aircraft size, which is expected to be negatively related to frequency, other conditions equal. Depending on specific business models, smaller aircraft can be deployed at a higher frequency, or larger aircraft may be used at lower frequency to serve the same passenger demand. For the 2SLS model, first the log form of frequency is estimated, and then the estimated frequency is used to estimate the log demand.

$$\ln(freq^t_{od}) = \beta_0 + \beta_1 dist_{od} + \beta_2 dist^2_{od} + \beta_3 \ln(gdp_{od}) + \beta_4 \ln(pop_{od}) + \beta_5 \ln(HHI^t_{od}) + \beta_6 \ln(seas^t_{od}) + \beta_7 \ln(growth^t_{od}) + \beta_8 \ln(LCC^t_{od}) + \beta_9 \ln(FV^t_{od}) + \beta_{10} \ln(AvgSize^t_{od}) + \epsilon^t_{od} \tag{6}$$

$$\ln(D^t_{od}) = \gamma_0 + \gamma_1 dist_{od} + \gamma_2 dist^2_{od} + \gamma_3 \ln(gdp_{od}) + \gamma_4 \ln(pop_{od}) + \gamma_5 \ln(HHI^t_{od}) + \gamma_6 \ln(seas^t_{od}) + \gamma_7 \ln(growth^t_{od}) + \gamma_8 \ln(LCC^t_{od}) + \gamma_9 \ln(FreqIV^t_{od}) + \epsilon^t_{od} \tag{7}$$

3.3.4 Supervised machine learning formulation

The objective of a supervised machine learning algorithm is to infer the function *(f)* that maps the relationship between N predictors, $x = (x_1, x_2, \ldots, x_N)$ and a continuous outcome vector *(y)* based on M training examples (e.g., historical cases) (Sarker 2021). To ensure robust model building and evaluation, the processed data is partitioned into training (data from 2013 to 2018) and test sets (data from year 2019). To avoid overfitting and estimate the generalization error of the model, k-fold cross validation resampling procedure is adopted during the training phase, where the training sample is partitioned into k equal-sized subsamples in which *k-1* subsets are used for fitting the model, and the single retained subset is used for validation (Srinivas and Ravindran 2018). This procedure is repeated to allow each of the k subsets to be used exactly once for validation.

The parameters measures reported by k-fold cross-validation is then the average of the values computed in the loop. Once the model is trained, its predictive performance is evaluated using new unseen dataset (i.e., test data). We apply three different algorithms: K-Nearest Neighbors (KNN), Random Forest (RF) and Gradient Boosting (GB).

KNN is a technique that predicts a continuous numerical value for a data point based on the average of the k-nearest neighboring data points. It works by finding k data points in the training set with the most similar features values to the input, and then calculates the average of their target values as final prediction. On the other hand, RF and GB are ensemble method. The former combines multiple decision trees to make predictions. It averages the outputs of individual trees to provide more accurate and stable prediction. The latter sequentially builds decision trees, with each tree correcting the errors of the previous one. It iteratively refines the prediction, leading to a highly accurate regression model by minimizing the residual errors.

Along these three algorithms, KNN is a valuable tool for dealing with noisy data in a context where no assumptions about the shape of the decision boundary are made. Random forest exploits bagging technique, which helps in reducing variance and overfitting, leading to a robust modelling estimation. Ultimately, gradient boosting relies on boosting optimization procedure for iteratively reducing errors and excelling in fine-tuning complex relationships with data.

Nevertheless, based on the context of analysis, these three algorithms can lead to a different performance. The gradient based optimization method works by capturing the change of a function with respect to its parameters, performing well when there's a strong gradient signal (i.e., the function output changes significantly with small changes in its parameters). On the other hand, random forest can successfully model small changes in the function output with changes in its parameters, potentially leading to a better result compared to gradient boosting. Also KNN, by capturing local relations between subsamples, is well-suited for working with data where the gradient signal is weak or inconsistent.

Consistent with prior literature (Rajendran et al. 2021), a 10-fold cross validation procedure is used during training. The algorithms are trained using a grid search method where a search of a specific parameter space is conducted to establish the best set of trees and neighbors. For KNN the number of neighbors is varied from 5 to 105 in increments of 10. For both RF and GB, the number of trees is varied from 100 to 1000 in increments of 100. The performance of the algorithms does not improve significantly beyond 20 neighbors and 300 trees. After finalizing the algorithm for each method, we evaluate the out-of-sample performance to select the best algorithm.

3.4 Results and validation

In this section, the empirical results relating to both econometrics and machine learning methods are presented. First, we present the coefficients of econometrics modelling and then the key performance indicators of both approaches are discussed, highlighting the gain in terms of performance when coming to machine learning algorithms and their explainability. As last, we provide a validation experiment showing how much the demand elasticity predicted by the two different approaches is aligned with the real-world trend.

3.4.1 Econometric results

To analyse the impact of including frequency we developed three different models delineations: M0 in which the only regressor included is the frequency, M1 including all the variables except the frequency and M2 including all the factors. In addition, each model specification has been developed segmenting the sample in short-haul and long-haul distance (as cutoff value 3000 km is assumed).

Table 4 presents the results of first difference estimator (FD) as specified in Eq. 4, OLS model in Eq. 5 and 2SLS in Eq. 7. For the 2SLS model, the first-stage regression results as specified in Eq. 6 are presented as well. The model coefficients are statistically significant, and their signs are consistent with expectations.

From the results, the geographical distance between origin and destination has a neg-

ative effect on the number of passengers. This implies an inverse U-relationship between distance and demand volumes, which aligns with the fact that for short distances – where higher traffic volumes are expected – more alternative travel modes (such as high-speed trains) are available, though these modes are less important competitors for longer travel distances (Jorge-Calderón 1997). On the other hand, if the distance increases further, traffic demand decreases again. GDP and population have both positive signs, showing a very small demand growth. Finally, growth rate and LCC share share have both positive signs, demonstrating that the lagged growth is a good proxy for the presence of underserved demand in the market and the additional share of LCC operating in each itinerary is able to stimulate new demand. Consistently, the HHI variable has a negative sign, denoting that the higher the concentration the lower the traffic demand, meaning that a more competitive environment leads to a higher traffic demand. This coefficient has to be interpreted as the compound effect of prices and market competition, as specified in Section 3.3.2. Notably, its magnitude encapsulates both aspects. In our model, prices are excluded, but they are intricately linked to market concentration levels. By incorporating the HHI variable, we effectively capture also the impact of prices, clarifying that the coefficient's significance stems from its ability to account for both effects, which are inherently intertwined and not easily distinguishable.

Among all the models delineations, the accuracy performance is consistently increasing, ranging from 40% to 80%. More interestingly, the models perform with higher accuracy when more variables are considered, as expected. The distance segment considering long haul itineraries performs systematically worst through all the models delineations. In model M0, considering only the frequency as regressor, we cannot capture the markets heterogeneity based on socio-economic features discrimination. It results in a fitting accuracy around 50%, on average. On the other side, model M1 includes only socio-economic features. Without considering the supply side, this model specification is not able to capture the dimensionality and magnitude of the market, as well as the frequency stimulation effect. As model M0, it reaches an accuracy performance around 50%, on average. As last,

model M2 including all the variables demonstrates a substantial accuracy improvement in the range of 70-80%. Particularly, the OLS specification and FD specification yield statistically similar frequency elasticity values. This aspect underscores the pivotal role of incorporating socio-economic determinants to capture the heterogeneity effect within markets. These determinants, which are constant over time, are then excluded from the FD specification and the robust accuracy is reached by tracking the temporal variation of year *t* with respect to year *t-1* for time-variant determinants. Clearly, a noteworthy limitation of FD specification (and correspondingly one of the main advantages of OLS specification) is its inapplicability in a new markets demand forecasting context where no historical data are available.

The correction for endogeneity of the frequency variable has a valuable impact on the regression coefficient. In the OLS model, a 10% higher frequency leads to an about 8% increase in passenger demand, similar also to the increment registered with first difference estimator (7.5-7.9%). When correcting for reverse causality with 2SLS specification the regression coefficient becomes smaller, around 5%. This is aligned with the real-world dynamic since the effect of demand on frequency is expected to be stronger than the effect of frequency on demand. Moreover, in the first stage regression the coefficients for the two instruments have the expected signs and are significant: flight frequencies are higher between markets with a higher feeder value, while frequencies are inversely proportional to average aircraft size.

It's important to note that previous studies (Boonekamp et al. 2018) have typically reported lower frequency elasticity values, primarily because they have isolated and examined the effects of frequency and prices separately. In contrast, the coefficient in our current study necessitates a more nuanced interpretation, by assuming that the response pattern of prices follows the historical pattern and the observed elasticity is therefore the combined effect of both price and frequency coefficients. Nevertheless, when correcting for reverse causality, our model exhibits reduced accuracy, with performance falling within the 50-60% range. While employing a 2SLS model effectively addresses endogene-

ity concerns, it comes at the cost of decreased fitting accuracy due to substituting the frequency variable with an instrumented regressor. This highlights a significant challenge: the trade-off between the interpretability and consistency of elasticity estimates versus the model's fit and predictive accuracy. Even though this approach is methodologically sound, the resulting low accuracy introduces a considerable risk of error. In real-world aviation operations, relying on forecasts with such high error margins is impractical, as it could lead to unreliable decision-making and operational inefficiencies. This underscores the need to explore whether integrating machine learning techniques might offer a solution, potentially balancing causal accuracy with high prediction precision to better meet the industry's forecasting needs.

3.4.2 Supervised machine learning results

Table 5 presents a comprehensive overview of the key performance indicators (KPIs) for all the models tested, specifically focusing on the scenario where all variables are considered (M2), as outlined in Section 3.4.1. The results reveal that the random forest model stands out as the top performer in the test set, exhibiting an out-of-sample R^2 of 0.96 and a remarkably low MAPE of 0.21. K-nearest neighbors (KNN) also delivers good results, demonstrating its reliability in this context, by achieving a R^2 value of 0.96 and MAPE of 0.26. However, the gradient boosting (GB) model falls short, likely due to its inherent capacity to excel with highly elastic data, which is not characteristic of this specific dataset. It registers an accuracy value of 0.87 and MAPE of 0.47. An encouraging observation is that none of the three algorithms exhibit significant signs of overfitting, underlining their robustness and generalization capabilities with regard the test set KPIs.

Comparing machine learning algorithms and econometric modeling performances, several critical distinctions can be further investigated. These distinctions not only shed light on the superior performance of machine learning models but also reveal advantages and disadvantages of each approach. One of the most prominent differences lies in prediction

	M0 (OLS)		M0 (FD)		M1 (OLS)		M1 (FD)		M2 (OLS)		M2 (FD)		M2 (First Stage)		M2 (2SLS)	
	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh
const	4,211***	4,027***			-1,023**	2,955**			-1,744***	4,671***			-1,05**	-0,669**	-7,622***	2,43***
	(0,006)	(0,038)			(0,03)	(0,225)			(0,019)	(0,201)			(0,002)	(0,001)	(0,003)	(0,002)
dist					-0,348**	-0,358**			-0,038**	-0,543**			-0,081**	-0,096**	-0,012***	0,051***
					(0,004)	(0,023)			(0,002)	(0,021)			(0,03)	(0,001)	(0,024)	(0,001)
dist sq					0,211**	0,019**			0,165**	0,342**			0,192**	0,088**	0,278**	-0,014**
					(0,002)	(0,012)			(0,001)	(0,002)			(0,001)	(0,023)	(0,004)	(0,003)
log gdp					0,055***	0,028***			0,003**	0,025**			0,061***	0,004***	0,91**	-0,049**
					(0,001)	(0,002)			(0,001)	(0,002)			(0,045)	(0,001)	(0,002)	(0,001)
log pop					0,037*	-0,031*			0,03**	-0,09**			0,01*	0,075*	0,179**	0,419**
					(0,002)	(0,009)			(0,001)	(0,008)			(0,003)	(0,001)	(0,001)	(0,026)
log HHI					-1,215***	-1,24***	-0,516***	-0,712***	-0,217***	-0,597***	-0,36***	-0,207***	-1,102***	-0,773***	-0,668***	-0,358***
					(0,003)	(0,016)	(0,058)	(0,067)	(0,002)	(0,017)	(0,052)	(0,061)	(0,004)	(0,002)	(0,031)	(0,019)
log seas					0,008*	0,016*	0,022**	0,095**	0,01**	0,015**	0,045**	0,241*	0,014*	0,032**	0,143**	0,586*
					(0,001)	()0,008)	(0,032)	(0,047)	(0,001)	(0,007)	(0,027)	(0,039)	(0,002)	(0,003)	(0,002)	(0,002)
log growth rate					0,026**	0,019**	0,113**	0,025**	0,007**	0,017**	0,094**	0,019**	0,039**	0,002**	0,57***	0,05***
					(0,001)	(0,006)	(0,018)	(0,027)	(0,001)	(0,005)	(0,016)	(0,021)	(0,014)	(0,043)	(0,034)	(0,003)
log lcc share					0,021***	0,034***	0,018***	0,01***	0,043***	0,075***	0,033***	0,025***	0,069**	0,071**	0,081**	0,079**
					(0,023)	(0,003)	(0,001)	(0,002)	(0,002)	(0,011)	(0,003)	(0,045)	(0,056)	(0,041)	(0,001)	(0,035)
log avg size													-0,075***	-0,037***		
													(0,002)	(0,001)		
log freq	1,034***	1,012***	0,894***	0,81***					0,757***	0,79***	0,63***	0,615***			0,55***	0,498***
	(0,001)	(0,011)	(0,051)	(0,041)					(0,001)	(0,012)	(0,045)	(0,04)			(0,001)	(0,002)
log feeder value													0,091***	0,085***		
													(0,001)	(0,021)		
geog. FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		
N	2.589.144	557.020	2.589.144	557.020	2.589.144	557.020	2.589.144	557.020	2.589.144	557.020	2.589.144	557.020	2.589.144	557.020		
R^2	0,67	0,38	0,57	0,39	0,58	0,32	0,58	0,42	0,84	0,51	0,86	0,57	0,46	0,48	0,58	0,59

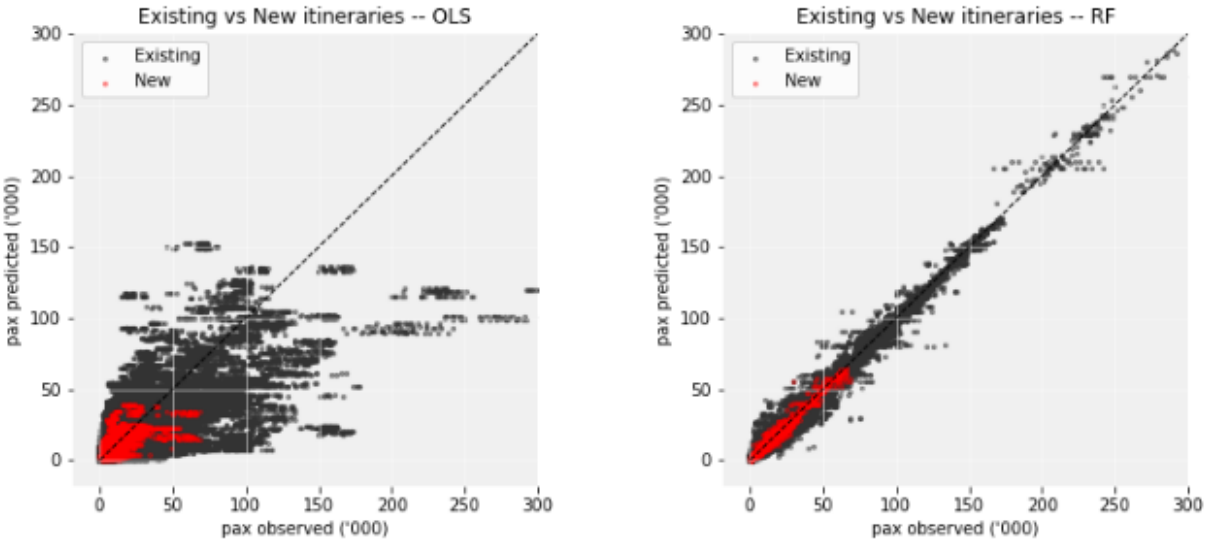
Legend: **sh** = short-haul; **lh** = long-haul; **OLS** = Ordinary least squares; **FD** = first difference.

Table 4: Econometric results

accuracy. The machine learning models, consistently outperform the econometric models in terms of predictive power. This is evident from the significantly higher R^2 values achieved by the random forest and KNN, as compared to the econometric models, as they fall short with lower R^2 values (in the range of 0.5-0.8) and higher error values (in the range of 1.2-0.77). Notably, the first difference estimator is the best performing in terms of accuracy, reaching an R^2 of 0.86. Simultaneously, the 2SLS estimator achieves the lowest MAPE value, around 0.8. This performance highlights the remarkable capacity of ML algorithms to adapt to various data characteristics, even in cases where the relationships among variables are non-linear or complex. In contrast, econometric models, by design, assume specific linear relationships between variables. This rigid assumption limits their ability to capture the intricacies of the data, especially when the relationships are non-linear. In particular, the good performance on the test set demonstrates the strong generalization capabilities of ML algorithms.

Due to continuous changes in the transportation network, a key requirement for a demand model is to predict well on new itineraries. To assess this aspect of predictive performance, we perform a tailored out of sample test experiment in which we exclude from the training set the observations regarding three airports: Amsterdam (AMS), Charles de Gaulle (CDG) and Rome-Fiumicino (FCO). We test the trained models (RF and OLS) only on those three airports. The superiority of RF becomes visible by plotting the predicted versus observed values and differentiating by new and existing itineraries (Fig. 3). For RF the values for new itineraries are scattered appropriately along the 45-degree line, showing the best fitting performance.

Even if RF is the best performing model, it lacks the coefficient readability of econometric models. A common criticism of ensemble and broadly machine learning models is that they are "black box". This critique stems from the trade-off that existing between model accuracy and interpretability. In general, to improve accuracy, models grow increasingly complex, thereby decreasing interpretability. Nevertheless, an interesting characteristic of RF is their capacity to provide a measurement of feature importance.



(a) OLS **(b) Random Forest**

Figure 3: Predicted vs. Observed demand values.

	RF				GB				KNN				FD				OLS				2SLS			
	ws		oos		ws		oos		ws		oos		ws		oos		ws		oos		ws		oos	
	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh	sh	lh
MAPE	0,12	0,24	0,21	0,3	0,43	0,6	0,47	0,66	0,2	0,34	0,26	0,43	0,95	0,91	0,9	0,89	1,2	1,01	1,03	0,99	0,83	0,79	0,78	0,77
MAE	141	193	125	201	294	321	304	392	130	192	152	205	892	843	871	850	1001	997	998	992	838	798	817	796
RMSE	204	302	284	308	481	578	541	606	235	393	293	398	7130	7099	7020	7068	7845	7778	7812	7756	7021	6990	6911	6959
R2	0,98	0,88	0,96	0,85	0,91	0,82	0,87	0,77	0,97	0,86	0,96	0,85	0,88	0,58	0,86	0,57	0,86	0,52	0,84	0,51	0,66	0,69	0,58	0,59

Legend: **sh** = short-haul; **lh** = long-haul; **oos** = out-of-sample; **ws** = within-sample.

Table 5: Key performance indicators ML vs econometrics

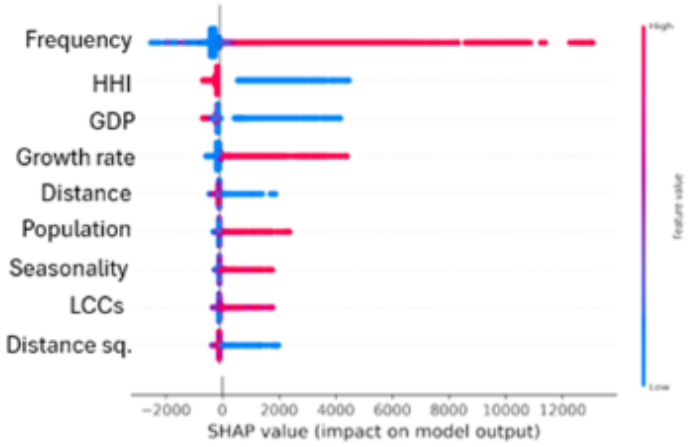


Figure 4: Feature importance of Random Forest computed using SHAP plot.

The field of game theory offers a method to plot the variables importance by calculating Shapely Additive Explanation (SHAP) values (Lundberg and Lee 2017). SHAP values improve interpretability by attributing to each feature the change in the predicted value when conditioning on that feature. We plot the SHAP values for RF in Fig. 4. The features are placed in descending order by their importance to predictive performance. When evaluating the model on the entire test set, frequency is the most important feature, as expected. Particularly, an increase in the LCCs share will increase the predicted passenger count. Conversely, a decreasing in the HHI value (meaning an increase in the competition in the market) will increase the predicted passenger count. Applying a colour scale to the SHAP values provides a reference for feature input values, and the horizontal axis depicts whether the effect is to increase or decrease the target value. For each feature, observations with relatively high input values are depicted in red, and observations with relatively low input values are depicted in blue. This plot does not prove causality, but it provides model interpretability by illustrating the correlation between model input and output values.

This discussion has effectively illustrated the superior fit of machine learning models in comparison to econometric models, aligning with our initial expectations. The application of SHAP plot provides valuable insights into the contribution of individual features to the model's predictions, which is instrumental in understanding the interplay of demand

and supply dynamics. However, while the primary strength of machine learning lies in its predictive power and adaptability to complex data patterns, it is still an open question whether ML models can also yield meaningful insights into the prediction of the frequency's stimulation effect over the demand.

3.4.3 Validation experiment

With the goal of having a deep understanding of the dynamics between demand and supply, this subsection explores the insights unveiled by ML algorithms. We conduct a comprehensive validation study covering the growth trends from 2018 to 2019. The prior discussion has underscored the superior predictive power of ML algorithms over traditional econometric approaches. However, one critical aspect that remains to be examined is the capacity of ML to provide meaningful and reliable stimulation trend patterns concerning the supply side. To address this gap, our objective is to compare the real supply stimulation trend over the demand trend of year t with respect to year *t-1* and compare it with OLS and ML stimulation trend prediction. This comparison aims to investigate which method better aligns with real-world patterns.

Our methodology starts with the collection of a unique dataset encompassing passenger flows for the years 2018 and 2019, along with passenger flow predictions for 2019 generated by both OLS and ML models. This dataset enables us to observe the real growth registered from 2018 to 2019 and compare it with the forecasting derived from the two model specifications. After collecting the sample, we calculate the real percentage growth of demand in 2019, and compare the OLS vs ML growth prediction of 2019, taking the real flow of 2018 as reference benchmark for all the three estimations. For studying the phenomena of increasing demand in a consistent way, we consider also the real percentage growth of supply from 2018 and 2019. The goal of this analysis is, given the real increase of supply, study how consistent is the demand increasing prediction from OLS and ML, and compare it with the actual demand increment. Based on this analysis, we extrapolate two sub-samples: one focusing on routes with a supply increase in the

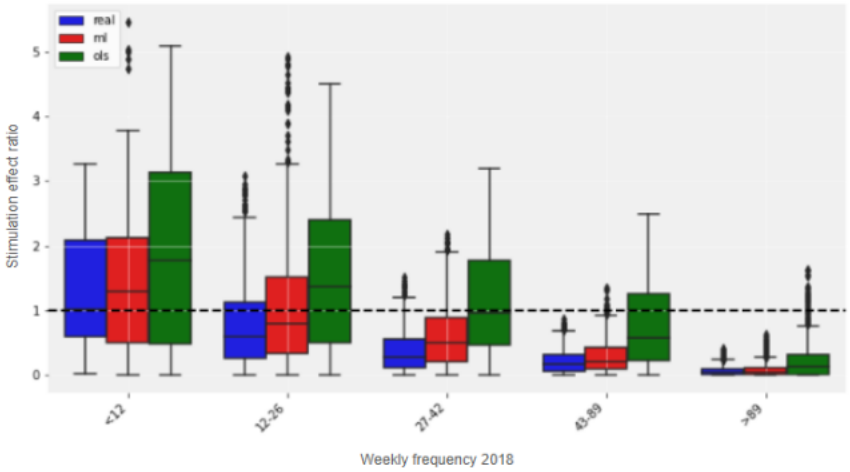


Figure 5: Stimulation effect ratio patterns by frequency distribution with a real growth ranging from 1% to 5%.

range of 1% to 5%, and another considering an increase in the range of 5% to 10%.[16] Using the frequency levels in 2018 as the benchmark level, we classify the weekly frequency distribution on 5 classes, based on quantile distribution. As final step, we calculate what we call "stimulation effect ratio", expressed as the ratio between the percentage variation of passenger flows over the percentage variation of frequency. As percentage variation of passenger flows we consider three different specifications, taking into account the variation of real-world flows registered in 2018 versus those in 2019, as well as versus the predicted flows by OLS and ML models for 2019.

In Fig. 5 and 6 we analysed those routes showing a growth rate spanning from 1% to 5% and from 5% to 10% during the aforementioned period. On the x-axis we have the frequency classes and on the y-axis we measure the stimulation effect ratio values according to real-world trends, OLS and ML predictions. Notably, our findings indicate that ML models consistently provide stimulation effect ratio predictions that are closely aligned with real world trends, outperforming econometric predictions.

With Fig. 7 we further investigate the correlation between actual stimulation effect ratio and those predicted by econometrics and ML models. ML models demonstrate a significantly higher correlation, reaching a coefficient of 0.8, highlighting their accu-

[16]Both sub-samples have a number of unique OD pairs ranging from 4000 to 5000, the class with more observations (around 1200-1600 city pairs) is the one representing more than 90 flights per week.

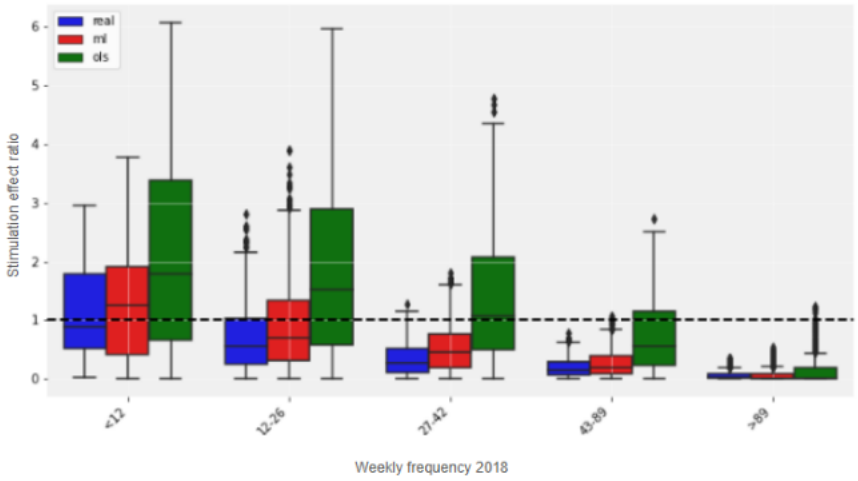
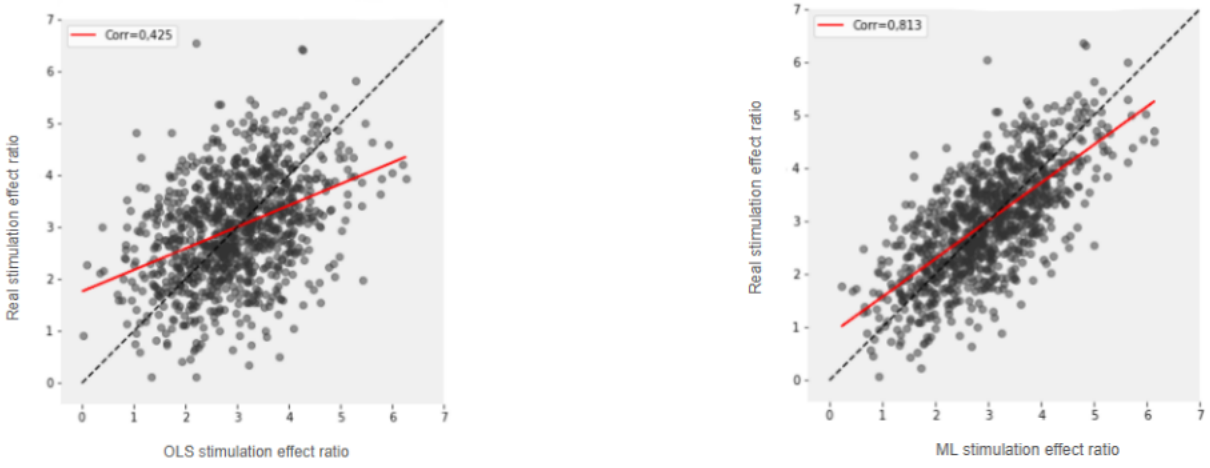


Figure 6: Stimulation effect ratio patterns by frequency distribution with a real growth ranging from 5% to 10%.



(a) OLS predictions **(b) ML predictions**

Figure 7: Correlation between real stimulation effect ratio and predictions from OLS and ML models.

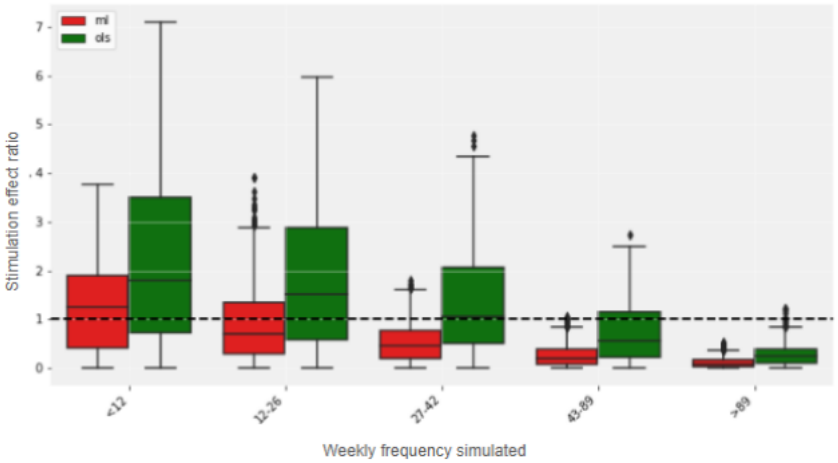


Figure 8: Stimulation effect ratio patterns by simulating a frequency increasement of 5%.

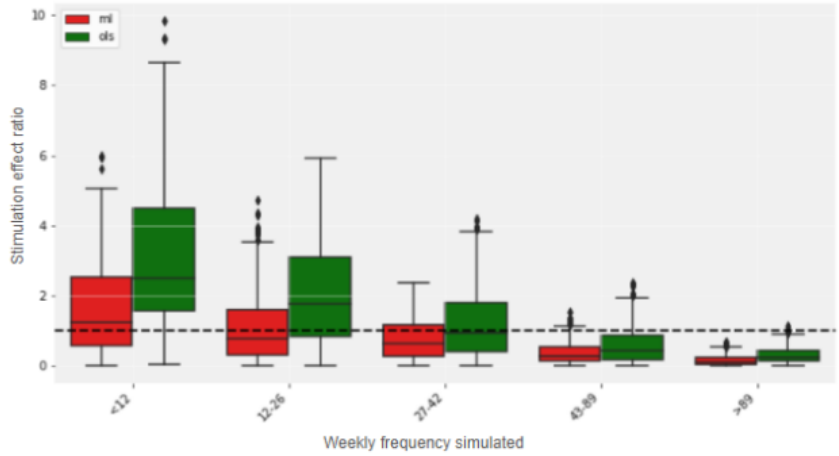


Figure 9: Stimulation effect ratio patterns by simulating a frequency increasement of 10%.

racy in capturing complex real-world dynamics, while OLS correlation falls short with a coefficient of 0.4.

As last, in Fig. 8 and 9 we simulate a 5% and 10% frequency increment, starting from 2018 supply levels. OLS and ML models are trained with the baseline 2018 frequency levels, then they have been tested on the simulated values. We compare the elasticity predictions of both econometrics and ML models when coming to forecasting future scenarios. In this context becomes evident that econometrics method consistently tends to overestimate elasticity values, predicting a higher trend.

These observations underscore the critical role that ML models play in enhancing the

reliability and accuracy of demand-supply interactions. With this analysis we highlight an important shortcoming that is the gain in terms of comprehensibility of ML techniques. The real strength of ML models lies in their ability to not only deliver superior predictive accuracy but also to align closely with real-world trends, thereby preserving the capacity to understand the impact of supply on demand without distortion. This makes ML models an invaluable tool for both accurate forecasting and insightful analysis of demand-supply dynamics.

3.5 Conclusion

In the context of aviation, accurately modeling the interaction between supply and demand is crucial for several reasons. Air travel demand is not only influenced by socio-economic factors such as income levels, population growth, and GDP but also by the availability and frequency of flights, the capacity of airlines, and the overall network structure. The relationship between demand and supply is inherently two-way: while demand influences the level of supply (e.g., more flights are scheduled when demand is high), supply also affects demand by shaping passenger choices and behaviors (e.g., more frequent flights can stimulate additional demand). Failing to properly account for this interrelationship can lead to biased estimates and poor model fit, which can undermine the decision-making process. For instance, overestimating the impact of supply could lead to unnecessary expansion in routes or capacity, resulting in inefficient use of resources. Conversely, underestimating it might lead to missed opportunities for growth and revenue. In aviation, where the stakes are high due to the significant investments involved, capturing the true dynamics between supply and demand, both in terms of variables' interpretation and forecasting accuracy, is essential for making informed decisions on route planning, fleet management, and strategic growth. By carefully addressing these interactions, models can provide more reliable insights, leading to more effective and sustainable outcomes in the industry. To actively contribute to the discussion around robust demand forecasting methodology, we propose a predictive analytic framework that

analyses three supervised ML algorithms and compares their performance with three traditional econometric formulations. We evaluate the models' ability to capture the socio-economic heterogeneity, supply-side variations, and overall market dynamics. The results demonstrate that, while all models perform well when using the complete set of variables, the random forest algorithm consistently delivers the highest quality solutions, accurately modeling the interactions between supply and demand, and providing more reliable and accurate forecasts for strategic decision-making in aviation.

In particular, considerable improvements in terms of forecasting accuracy can be achieved when shifting from econometrics to data-driven approach. For instance, econometrics models present a R^2 values in the range of 50%-80% and they fall short in error indicators, measured by MAPE, in the range of 0.7-1.2, especially when correcting for reverse causality. On the contrary, ML algorithms outperform them measuring R^2 values in the range of 80%-95% and minimizing the MAPE metrics with values around 0.2-0.4, showing an improvement of about 60%. Notably, for each model we calculate both the within sample and out of sample key performance indicators and ensure that none of the supervised ML algorithm show an overfitting issue, demonstrating that all of them are able to capture intricate relationships within training data, being able to generalize them to unseen test data. To better validate our models, we not only provide a variable-importance discussion for our best performing ML algorithm (i.e., Random Forest), but we develop a tailored validation procedure measuring the accuracy in capturing the elastic response of demand in a growing supply scenario (in the range of 1%-5% and 5%-10%). Results corroborate closeness of ML algorithm flow predictions to real-world trend compared to econometrics. In particular, ML model is able to unveil meaningful insights into demand-supply relation and to capture the demand increasing pattern under frequency stimulation effect without overestimating it, as econometric model's coefficient does.

The modeling approach developed in this paper and the encouraging results can thus powering the use of ML in predictive implementations where accuracy and interpretability are equally important, especially in scenario where the accuracy of traditional econometric

models falls short due to inherent characteristics of the data. Beyond faithfully replicating historical trends, the tailored ML model can serve as a strategic decision-support tool for airlines and other aviation stakeholders. By accurately capturing the complex interaction between supply and demand, the model enables the evaluation of tailored "what-if" scenarios, such as the impact of increasing flight frequency on specific routes, introducing new services, or reallocating capacity across markets. Airlines can leverage these insights for route planning, fleet scheduling, and capacity optimization, ensuring that resources are deployed where they are most likely to generate additional demand and revenue. Similarly, regulators and airport authorities can use the model to anticipate market responses to infrastructure investments or policy changes, supporting informed decisions on airport expansion, congestion management, and long-term network development. In this way, the ML framework not only improves forecasting accuracy but also provides actionable tool that helps stakeholders make forward-looking, evidence-based decisions, aligning operational and strategic choices with real-world demand dynamics.

Chapter 4

Air Demand Impacts of China's 2030 High-Speed Rail Expansion Plan

4.1 Introduction

Over recent decades, a consistent increase in the number of air passengers can be observed across major markets in the world. The main factors contributing to this growth include rising incomes, a trend towards globalization, increased affordability of air travel, and an overall improved connectivity (Hofer et al. 2018, Hanson et al. 2022, Ozmec-Ban and Babić 2023). This growth trend has been interrupted temporarily through the COVID-19 outbreak, with a significant disruption of air transportation and times of volatile demands; see Sun et al. (2021b, 2022) for overview studies. Nevertheless, the overall growth trend is projected to continue, with forecasts anticipating a steady rise in air travel demand, particularly in relationship to emerging markets, such as the Asia-Pacific region. According to IATA (International Air Transport Association)[17], in China the airline industry has been experiencing a strong recovery. Domestic air travel surged by 272% in November 2023 compared to the previous year, driven by the lifting of travel restrictions. IATA forecasts that international demand in the Asia-Pacific region, including China, will reach 92% of 2019 levels in 2024, fully recovering to 101% by 2025. Specifically, China's domestic air demand is expected to exceed pre-pandemic levels, reaching 111% by 2024 and 118% by 2025. China however is also strongly committed to the future use of high-speed railway technology in order to prepare its domestic transportation for sustainable future (Sun et al. 2017). China's high-speed railway network is already the

[17]Press Release No. 2 in date 10 January 2024.

world's longest network at about 40.000 km, essentially constructed within less than one and a half decades (Ren et al. 2023). According to plans from the Chinese government, the network will reach more than 70.000 km within the next two decades. The impact of this development on airlines is wide-reaching, forcing Chinese' domestic airlines to reduce airfare and cancel many (earlier highly-profitable) regional flights (Liu et al. 2019). The major advantages of high-speed rail are its competitive edge in accessibility from city centers (Sun et al. 2021a) and frequencies (Zhang et al. 2019), and overall convenience for short to medium distance travel.

Although the expansion of HSR in China was initially aimed at upgrading conventional rail and enhancing inter-city mobility rather than directly targeting air transport, overlaps in medium-distance markets (roughly 500–1000 km) have nonetheless created significant competition between HSR and airlines[18]. Numerous studies have analyzed the impacts and policy implications of air-HSR competition, covering both short-term and long-term effects on airline profits (Jiang and Zhang 2016, Tsunoda 2018, Chen et al. 2023), parallel airline services (**?**Román et al. 2007, Chen et al. 2020), and route-level demand (Park and Ha 2006, Wang et al. 2018b, Liu et al. 2019, Zhang et al. 2019). Studies also examine passenger behavior, such as willingness-to-pay and modal choice, in the context of air-HSR competition. Research suggests that factors like HSR frequency and travel time significantly influence airline performance (Behrens and Pels 2012, Zhang et al. 2019), while HSR can attract passengers from both air transport and car users on long-haul trips (Martín and Nombela 2007). In China, HSR has notably decreased traffic demand for airlines, especially on short-haul routes, and has positively impacted tourism development (Zhang et al. 2017a, Chen et al. 2023). The introduction of HSR has had a substantial negative effect on air demand for short-haul routes (less than 850 km) where air-HSR substitutability is high (Wang et al. 2018b). Additionally, recent studies like (Zhu et al. 2021) have explored how HSR impacts air-HSR competition under conditions of high

[18]In Section 2.1 we discussed an overview of both air-HSR cooperation and competition dynamics. In this chapter we focused on the competition point of view, focusing on its aspects and implications in the Chinese context.

flight delays, particularly in the Beijing-Shanghai corridor.

In response to the introduction of HSR, airlines have implemented various strategies to mitigate its impact, including adjusting flight frequencies, modifying pricing strategies, and enhancing service quality. Studies have explored how airlines adapt by reallocating capacity and revising network configurations to stay competitive (Fu et al. 2010, Bilotkach and Lakew 2014). Research also examines how airlines adjust airfare and traffic scheduling when HSR is introduced (Gu and Wan 2022), as well as the effect of HSR on airline pricing strategies and competition between full-service carriers and low-cost carriers on high-demand routes (Su et al. 2019, 2020).

Nevertheless, while a variety of econometric models in the literature are capable of producing forecasts, relatively few studies have explicitly applied them to assess the potential effects of planned HSR lines that have not yet been implemented. Existing work has largely concentrated on evaluating the realized impacts of current HSR on air routes and market dynamics, whereas ex-ante investigations are more often based on survey data or accessibility indicators. Only a limited number of studies address the potential role of HSR projects that are still under construction. For example, Wang et al. (2018c) examine how new HSR development may influence ground accessibility, offering insights into the broader reshaping of transportation networks. Likewise, Zhou et al. (2018a) and Zhou et al. (2018b) focus on the expansion of the Chinese HSR network over the past decade, proposing a connectivity index to evaluate the degree of city integration within the system. By contrast, research on the relationship between HSR and air transport demand has been predominantly retrospective. Strauss et al. (2021) assess the impact of HSR on air demand up to 2017, Chen et al. (2025) analyze reductions in air demand through 2019, and Yang et al. (2023) investigate how the Chinese city-pair air network adapted to HSR expansion between 2009 and 2019. These studies provide important evidence of realized impacts, but they do not extend to discuss the potential effects of future HSR projects. Besides the meaningful intuition of the studies, there is no prominent research direction regarding the influence that the current HSR infrastructure

development can have for air city-pair markets demand in the long-term future. Given that infrastructure projects require substantial investment and have long lifespans, it is crucial to forecast how ongoing HSR development may influence air city-pair demand in the coming decades. Such forecasts they offer valuable forward-looking insights that can support strategic planning and policy evaluation.

In this paper, forecasting models provide valuable insights into how future HSR developments might reshape air travel demand and network configurations in established air domestic markets. This study represents a pioneering effort in forecasting the timing, extent, and specific corridors where a significant portion of passenger demand is expected to shift from air transport to HSR. By addressing this gap, stakeholders can better anticipate the strategic responses of airlines, thereby enriching the understanding of the evolving dynamics between air and rail transportation modes. In our study, we focus on the Chinese context and develop an aggregated market demand forecasting model leveraging a dataset spanning from 2003 to 2050, integrating different data sources. We use data from 2012 to 2019 for leveraging variables able to capture not only the air market supply development and the socio economic characteristics, but also the presence and the competing influence of HSR development[19]. As core of our research, we conduct a case study investigation on 2030 railway expansion plan with 20 years forecasts - thus producing and discussing results until 2050-[20]. By applying the forecasting models to HSR expansion plan, the study provides valuable insights into the long-term impacts of HSR development on air travel demand, offering a forward-looking perspective that is crucial for strategic planning. For this purpose, we employ socio-economic data projections coming from Chen et al. (2020), Wang et al. (2022a) and Wang and Sun (2022). By employing future socio-economics projections, we ensure that the demand predictions account also for anticipated changes in population, economic growth, enhancing reliability and robustness of the forecasts. The findings of this research will be highly relevant

[19]Data regarding the opening of HSR lines have been collected within the time period 2003-2020.

[20]since the most updated literature provide robust socio-economic projections until 2050. Please refer to Section 4.3 for a comprehensive discussion regarding data sources.

for policymakers, transportation planners, and airline managers. The ability to predict the impact of HSR expansion on air travel demand can inform infrastructure investment decisions, strategic planning for airlines, and policy formulation aimed at optimizing the transportation network.

The rest of the paper is organized as follow. Section 4.2 provides a comprehensive literature review about air demand forecasting and HSR competition. Section 4.3 explains the data and features used, the methodological framework as long as the variables elaborated for the demand prediction purpose. Section 4.4 shows the results, discuss the accuracy of the proposed models and the influence of different types of variables. Section **??** discusses the development of new HSR lines 2030, shows a first empirical explorative analysis of the network impact and finally apply the forecasting model to the new set of city pairs, presenting which cities will be mostly affected by the shift air-HSR. Finally, Section 4.6 concludes the study.

4.2 Related Literature

In this section, we provide a literature review discussion on the relation between air transport and HSR. Table 6 summarizes part of the most relevant literature on air-HSR competition and trade off among each other. It highlights their geographical focus, time period analyzed, level of aggregation, and the type of demand predicted.

Table 6 provides an overview of key contributions in the literature on air–HSR competition, outlining their geographical focus, period of analysis, level of aggregation, and the type of demand predicted. Beyond summarizing these elements, the table also highlights the most common perspectives adopted in existing studies while illustrating the heterogeneity in approaches and findings across different research contexts. Several studies focus exclusively on the Chinese market, reflecting the significant interest in understanding the dynamics between HSR and air transport domestically. For instance, Wang et al. (2015), Zhang et al. (2019) and Zhang et al. (2024) center their analysis on China, examining how HSR impacts the air transport sector within the country. This focus is

Table 6: Summary of literature for air-HSR competition.

Literature	Geography		Time period		Aggregation		Demand predicted	
	China	Others	Ex-post	Ex-ante	Market level	Carrier level	HSR	Air
Givoni and Dobruszkes (2013)		✓	✓		✓		✓	
Albalate et al. (2015)		✓	✓		✓			✓
Wang et al. (2015)	✓			✓	✓			✓
Li and Sheng (2016)	✓		✓		✓			✓
Wang et al. (2018b)	✓		✓			✓		✓
Kroes and Savelberg (2019)		✓		✓	✓			✓
Li et al. (2019)	✓		✓		✓		✓	
Zhang et al. (2019)	✓				✓			✓
Bergantino and Madio (2020)		✓	✓			✓		✓
Gu and Wan (2020)	✓		✓		✓			✓
Wang et al. (2020b)		✓	✓		✓		✓	
Cai et al. (2021)	✓		✓			✓	✓	
Strauss et al. (2021)	✓		✓		✓			✓
Yu et al. (2021)	✓		✓		✓			✓
Wu and Han (2022)	✓		✓			✓	✓	
Nurhidayat et al. (2023)		✓	✓			✓	✓	
Lee et al. (2024)	✓		✓			✓	✓	
Ma et al. (2024)	✓			✓	✓			✓
Zhang et al. (2024)	✓		✓		✓			✓
Our study	✓			✓	✓		✓	

Geography = Which is the geographical context on which the study is focused on.
Time period = If the study is considering an historical analysis (ex-post) or if it deals with future railways lines not already opened (ex-ante).
Aggregation = If the study considers carriers' level specification or not.
Demand predicted = Type of demand predicted in the study (air demand or high-speed rail demand).

crucial given China's rapid railway network expansion and its profound implications for air travel. Very few studies, such as Wang et al. (2015), Kroes and Savelberg (2019) and Ma et al. (2024), attempt to predict near future air demand trends considering future railway network expansions. In particular, in Wang et al. (2015), the author focuses on the hinterland regions in China and the overlapping markets of air transport and HSR. Due to the lack of long-term socio-economic projections data, the authors predict the short-term impact of HSR but acknowledge that the long-term effects remain unclear. The most notable finding of the study is that HSR has the most significant competitive advantage on medium distances. The majority of studies, such as Li and Sheng (2016), Li et al. (2019), Gu and Wan (2020), Strauss et al. (2021) and Yu et al. (2021) conduct their analysis at the market level, examining overall trends in air travel demand without delving into the competitive dynamics of individual airlines. Some studies explicitly focus on predicting HSR demand, such as Givoni and Dobruszkes (2013) and Wu and Han

(2022), where the latter examines the accessibility of HSR stations and its impact on HSR-air competition in the Chinese domestic context. Others, like Wang et al. (2018b), Nurhidayat et al. (2023) and Lee et al. (2024), concentrate on air demand prediction from a carrier level perspective. These studies highlight the trade-offs and competitive interactions between the two modes of transportation, providing insights into how HSR expansion can influence air travel patterns.

However, from the previous studies, it emerges that the competitive worldwide pressure did not always result in the overall reduction in total air travel demand, which in some contexts increased thanks to operators reorganizing their routes (Clewlow et al. 2014, Sun et al. 2024) and to forms of air-HSR integration. In particular in the European context, Albalate et al. (2015), Dobruszkes (2025) and Kampp et al. (2025) find that the presence of HSR stations in airports might compensate airlines for the effects of competition with HSR. A subsequent literature has further explored forms of air-HSR integration from a theoretical (see e.g., Jiang et al. (2017), Jiang and Zhang (2014), Xia and Zhang (2016)) and empirical perspective (e.g., Givoni and Banister (2007), Li and Sheng (2016), among others). On short and medium haul distances, as for the Chinese domestic market, changes in operators' business strategies when facing the introduction of HSR have twofold dimensions: the supply-side and the demand-side. Many empirical studies addressed the former. For instance, some contributions have shown that airlines reacted to the introduction of HSR with changes in the number of flights (Jiménez and Betancor 2012, Dobruszkes et al. 2014, Chen 2017) and seats (Dobruszkes et al. 2014, Albalate et al. 2015, Wan et al. 2016, Chen 2017). Yet, the demand dimension is also relevant. When HSR is introduced, transport demand becomes more elastic as passengers have more option to exploit (**?**Zhang et al. 2017a). Previous papers have shown that the introduction of HSR services can lead to a significant change in demand (De Rus and Nash 2007, Clewlow et al. 2014, Chen 2017, Nurhidayat et al. 2023). Partly, this is dependent on the greater attractiveness of HSR due mainly to the reduced impact of access and egress time (Wang et al. 2018b, Bergantino and Madio 2020, Xu et al. 2023).

The methodologies employed in the literature on air demand forecasting under HSR competition encompass a wide range of quantitative and comparative techniques, each tailored to specific research questions. For instance, Albalate et al. (2015) adopt comparative analysis and case studies to investigate competitive dynamics in the European context, with a focus on pricing, frequency, and passenger volumes. In contrast, Li and Sheng (2016) apply regression analysis and time-series forecasting to historical data to capture demand trends in integrated air–HSR services along the Beijing–Guangzhou corridor. Similarly, Zhang et al. (2019) employ econometric approaches, including fixed-effects and random-effects regressions. Other studies, such as Strauss et al. (2021) and Yu et al. (2021), rely on gravity-model formulations, while Li et al. (2019) use panel regression and Wang et al. (2018b) implement a difference-in-differences (DID) framework. Moreover, Kroes and Savelberg (2019) develop a discrete choice and modal split model, whereas Gu and Wan (2020) integrate econometric modeling with simulation techniques, combining price competition models, travel time difference analysis, and catchment area expansion scenarios to assess the effects of HSR entry on air traffic.

In conclusion, most of the reviewed literature assesses the impacts of HSR retrospectively, focusing on realized effects observed in historical data. While forecasting models are likewise estimated from historical evidence, their application to planned HSR expansions allows us to translate past relationships into forward-looking scenarios. This provides projections of how future air travel markets may evolve under upcoming HSR developments. Such forecasts are particularly valuable for policy makers and industry stakeholders, as infrastructure projects require large-scale investment and have long lifespans. By linking established empirical evidence with future expansion plans, forecasting studies can inform strategic planning and decision-making regarding the interplay between HSR and aviation.

The novelty of our study lies in providing demand projections upon the opening of Chinese new HSR lines in 2030 and highlight which connections are more likely to show a huge air-HSR demand shift, providing evidence of the pivotal role of the distance between

cities. More specifically, we integrate the demand forecasting model under HSR presence and the plan of Chinese HSR expansion (along with robust socio economic projections for the next decades), offering insights into future developments under different demographic development frameworks. Three major contributions of our study are summarized as follows: (i) We empirically investigate and demonstrate the competition aspect between domestic civil aviation and HSR in China, giving evidence of the scenarios in which HSR takes more advantage over air transport; (ii) Taking into account the air-rail competition, we estimate the share of air demand that has been displaced by HSR due to its competitive advantage and in which time horizion (e.g., in the next 5 years, 10 years, 20 years); (iii) Given the new set of HSR market connections for the next decades, we evaluate the competitive travel time advantage and the future demand shift; moreover, under several civil aviation and socio economic scenarios we highlight the set of cities where domestic aviation sector will likely be under most pressure.

4.3 Data and methods

This section details the data sources and forecasting methodology employed in our analysis. Section 4.3.1, provides an overview of the datasets utilized, including their origins, types, and the relevant information collected. It outlines the rationale behind selecting these sources and how they contribute to the study. Section 4.3.2 and Section 4.3.3, describe the variables and models specifications applied to analyze the data. It covers the methodological approach for forecasting air passengers demand under HSR presence, detailing the econometric models used and the criteria for evaluating model performance.

4.3.1 Data sources

This study employs a wide range of data sources to thoroughly analyze the impact of HSR on air domestic demand in China. To ensure a comprehensive examination of the evolving dynamics between air transport and HSR, we integrate data spanning from 2003 to 2050.

Our primary data on flight schedules, including flight frequency, seat availability, carrier information, and travel times, was obtained from the OAG Schedule Analyser for the years 2012 to 2019. This dataset is complemented by passenger numbers and distance data from the OAG Traffic Analyser for the same period. These data sources provide a detailed view of air travel patterns and are instrumental in understanding market dynamics within metropolitan areas (city), which are grouped according to OAG's classification system. HSR entry data, detailing the introduction of HSR lines from 2003 to 2020, was manually collected from government and railway official websites. This dataset helps track the temporal expansion of the HSR network and its potential effects on transportation dynamics. Looking forward, we incorporate data on HSR backbone lines planned for 2030 as outlined in the Medium and Long-Term Railway Network Plan (2016), providing projections of HSR network expansion. Additionally, socio-economic projections segmented by province from 2025 to 2050 (Chen et al. 2020, Wang and Sun 2022, Wang et al. 2022a) are used to forecast long-term transportation demand by considering future demographic trends. Travel time data for public transportation and car travel, sourced from OpenStreetMap (OSM), offers city-based values that highlight accessibility and convenience for itineraries starting and ending in city centers. This data is essential for evaluating the competitive travel time advantage of HSR compared to air transport, particularly in new HSR connections. Population and GDP data from the China Statistical YearBook, covering 2009 to 2019, provide insights into demographic trends and economic conditions across various cities. This information is crucial for understanding the broader socio-economic context influencing transportation demand.

Our dataset comprises approximately 500.000 observations, each detailing an origin city, destination city, carrier operating, and monthly frequency, covering about 5.000 unique origin-destination markets. These comprehensive data sources (see Table 7 for a detailed overview) form the foundation of our demand forecasting model. They enable us to capture the intricate interactions between HSR and air transport, and to predict the potential impacts on aviation market efficiency and profitability as the HSR network

Data Sources	Variables	Usage	Time period
OAG Schedule Analyser	Air scheduling and supply data	Modeling air demand	2012-2019
OAG Traffic Analyser	Air demand data	Modeling air demand	2012-2019
Open Street Map	Public transportation and car travel time	Case study on new HSR lines	-
China Statistical Yearbook	GDP pro capite and population data	Modeling air demand	2009-2019
Chinese Railways Official Websites	HSR lines opening and city connected	Modeling air demand	2003-2020
Medium and Long Term Railway Network Plan	HSR lines to be opened in 2030	Case study on new HSR lines	2016-2030
Chen et al. (2020), Wang et al. (2022a) Wang and Sun (2022)	GDP and population projections	Case study on new HSR lines	2025-2100

Table 7: Data Sources

continues to expand.

4.3.2 Variables definition

In our demand forecasting model, we consider a wide range of variables to capture the multifaceted interactions between air transport and HSR. These variables are critical for understanding the determinants of air passenger demand and predicting how it evolves with the increasing penetration of HSR (see Table 8 for overview descriptive analytics)[21].

[21] Population (*Pop ('0000)*) and GDP (*GDP per capita ($)*) are node-based variables. Distance, Passengers, Frequency, Seats, Travel time ratio, Feeding HSR, and Age HSR are link-based variables measured at the monthly level. "." is used as the thousand separator, and "," as the decimal separator.

Table 8: Descriptive analytics

Variable	Mean	St Dev	Min	Median	Max
Distance (km)	1386	839	300	1892	2373
Passengers	59.701	30.444	50	65.167	345.264
Pop ('0000)	487.67	172.44	17.497	559.45	2119.12
GDP capita ($)	12.899	5.235	12.618	13.115	17.213
Frequency	55	72	5	88	1152
Seats	8659	11788	210	5053	191268
Travel time ratio	0.9	1.0	0.2	0.8	1.2
Feeding HSR	45	57	25	73	112
Age HSR (years)	7	3	1	5	12

Our dependent variable is the number of passengers, it reflects the current demand for air travel on a given route. It is denoted by D^t_{od}, where we index origin and destination market by o and d, time periods by t. Analyzing this variable also with a one-year lag allows us to observe demand trends and persistence over time.

Distance is a fundamental variable that directly influences travel mode choice (defined as distance in kilometer in our model). Longer distances generally favor air travel due to its speed advantage, whereas shorter distances may be more competitively served by HSR. To capture non-linear effects, we also consider the squared form of distance.

Frequency refers to the number of flights scheduled in a market pair. Higher frequency increases convenience and flexibility for passengers, enhancing the attractiveness of air travel. Within the empirical model we also account for the change (delta) between time

periods to understand how past frequency adjustments impact current demand. As we have also data at airline level, frequency of competitors on the same route has been considered, measuring the supply level offered by all competing airlines. This variable, included in the model as delta of competitors' frequency, helps in understanding the competitive landscape and how it influences a carrier's market share and passenger demand.

The Herfindahl-Hirschman Index (HHI) is used to measure market concentration. It is calculated using the number of seats offered by each carrier relative to the total market seats. Higher HHI values indicate less competition, which can affect pricing and service quality, thereby influencing passenger demand[22]. For reducing endogeneity issues with the dependent variable, we consider it with 1 year lag.

Socio economic information are given by population[23] and gross domestic product variables. These indicators are used to capture the underlying drivers of air travel demand and are essential for forecasting future demand by incorporating the anticipated socio-economic and demographic developments of the regions included in the analysis. More in particular, population data helps in understanding the potential market size for both air and rail travel. Larger populations typically correlate with higher travel demand. GDP per capita serves as an indicator of economic activity and the ability of the population to afford travel. Higher GDP per capita usually indicates higher demand for both air and rail travel services[24]. The Compound Annual Growth Rate (CAGR) measures the annual growth rate of a variable over time, providing insights into long-term trends in air travel demand or supply metrics. In our study, it has been applied to both population and GDP per capita.

To model the impact of HSR on the aviation market, we use three key variables. First, we assess the ratio of HSR travel time to air travel time, incorporating additional

[22]It assumes values ranging from 0 to 1 where 0 is the perfect competition and 1 is a monopolistic market.

[23]Please note that the population values in Table 8 should be read, for instance, as 487,67*10.000 = 4.876.700.

[24]Both population and GDP variables within the empirical model have been engineered by taking the product of values at origin and destination cities.

waiting times[25]: 90 minutes for air travel and 60 minutes for HSR. In Section **??** we have reported the rationale of adding waiting times during a travel time itinerary computation, as well as the proposal of four different scenarios that have been tested and evaluated, following Koppelman et al. (2008), Moyano et al. (2018), Wang et al. (2020a). After the sensitivity analyses has been conducted, Figure **??** and **??** show that once the waiting times have been introduced, there's no significant difference in the total travel time itinerary evaluation; therefore, we have adopted the combination of waiting times with the lowest time difference between air and HSR. By selecting the scenario with the smallest waiting time difference between HSR and air travel, we have positioned ourselves in the most stringent scenario. This approach assumes a lower likelihood of HSR being more convenient only thanks to reduced waiting times, thereby setting a higher benchmark for HSR's competitiveness. This decision ensures that any observed advantages of HSR in our analysis are robust, as they occur under conditions that are least favorable to HSR's perceived convenience relative to air travel. We also add 30 minutes to the HSR time if the itinerary involves a single stop, acknowledging that Chinese HSR networks are often designed such that one-stop routes can be faster and more efficient than direct flights offering smooth connection experience. This is crucial because China's HSR network is strategically structured to optimize travel efficiency, with many one-stop itineraries potentially outperforming non-stop air travel due to the network's extensive coverage, central station locations, and streamlined boarding processes. Second, we include an HSR feeding variable, which measures the connectivity of an area by summing the number of cities linked by HSR to both the origin and destination. This variable captures the extent of connectivity and the importance of an area as a hub within the HSR network, reflecting its potential to support and facilitate additional connections. Third, age of HSR variable is represented in the logarithmic form of the number of months since the HSR line opened. If between the origin and destination cities there's no direct HSR

[25]Waiting times includes the time required to travel between city center and airports/HSR stations, as well as time spent at stations or airports before departure, security checks, boarding, and other pre-travel procedures.

connection and the itinerary has to be computed by 1-stop travel connection, then the age HSR for that itinerary is assumed to be the age of the youngest leg between the two. This variable captures the maturation effect of HSR services, where newer lines may gradually attract more passengers over time as they become better known and trusted.

Lastly, trend variable elaborated as continuous time index ranging from 1 to 8 is included, representing the study period's years. This variable helps capture underlying time-related effects that could influence demand in a time-series analysis context.

Incorporating these variables into our demand forecasting model allows for a nuanced analysis of the factors influencing air passenger demand in the context of increasing HSR presence.

4.3.3 Methodology

To analyze the impact of HSR on airline network structures and forecast air passenger demand, we adopt an econometrics methodological framework. This framework involves dataset preparation, model formulation, and evaluation using key performance indicators (KPIs). The dataset is split into two subsets for model training and testing, the training set uses data from 2012 to 2017, the test set uses data from 2018 to 2019 for validate the predictive performance[26].

Following Birolini et al. (2021), we employ a Ordinary Least Square (OLS) with a semi-logarithmic regression formulation incorporating the HSR competition, the formulation follows the specification:

$$\ln(Y) = \beta_0 + \beta_j X_j + \beta_i ln(X_i) + \epsilon \tag{8}$$

[26]Although OLS regression provides unbiased estimates of coefficients under classical assumptions, splitting the dataset into training and testing subsets allows us to evaluate the model's predictive performance on unseen data, rather than only its fit to the sample used for estimation. By training the model on 2012–2017 data and testing it on out-of-sample 2018–2019 data, we can assess whether the relationships estimated from historical observations generalize to future periods, which is particularly important when the model is intended for forecasting air travel demand under new HSR scenarios. This approach also helps identify potential overfitting and ensures the model's predictive reliability beyond the estimation period.

For the linear terms X_j, the coefficients β_j represent the change in $\ln(Y)$ for a one-unit change in X_j, which corresponds to an approximate percentage change in Y. For the logarithmic terms $\ln(X_i)$, the coefficients β_i represent the elasticity of Y with respect to X_i, i.e., the percentage change in Y for a one-percent change in X_i. The semi-logarithmic OLS formulation is particularly beneficial when the relationship between the dependent and independent variables is multiplicative rather than additive. By taking the natural logarithm of the dependent variable, we transform a non-linear relationship into a linear one, thus applying linear regression techniques. Additionally, the semi-logarithmic model is particularly suitable in cases where certain variables are best expressed in logarithmic form, while others are more appropriately kept in their original scale. This is common in economic and transportation studies, where relative changes are often more meaningful than absolute changes[27].

To comprehensively analyze the factors and dynamics influencing air passenger demand, we develop multiple model formulations based on different sets of independent variables and data aggregation levels. First, we consider formulations M1 and M2, where in the former we consider population and GDP per capita as proxies for the demand, in the latter we incorporate lagged demand values to capture the persistence and trend of air passengers over time. For each M1 and M2 formulation, we estimate both a carrier-level and a carrier-aggregated model. The carrier-level version keeps observations disaggregated by airline, meaning that a route-week appears multiple times when served by multiple carriers. Although this specification does not provide airline-specific causal demand estimates—because rival attributes are not included and the estimated effects represent averages across all airlines—it serves an important data-analytic purpose. Specifically, the

[27]When dealing with datasets that include negative values or zeros, directly applying a logarithmic transformation can be problematic, as the natural logarithm is undefined for non-positive numbers. A common workaround is to add a constant (such as +1) to all values before taking the logarithm. However, this approach has significant drawbacks (Webb 1970, Bagwell 2005). First, adding a constant can distort the true relationships between variables. The choice of the constant is arbitrary and can significantly affect the results, leading to biased estimates. Second, the resulting coefficients may become difficult to interpret, as the transformation alters the scale and meaning of the original variables. Third, for variables with a substantial number of zeros or negative values, adding a constant can lead to misleading results. It may artificially inflate the values, leading to overestimation or underestimation of the actual effects. Instead, a semi-logarithmic model can be a more robust solution.

carrier-level dataset contains richer variation in supply variables (frequency, aircraft size, capacity deployment, LCC presence, etc.) because competing carriers typically adopt different strategies on the same route. Aggregating to the market level compresses this variation into a single value per OD-week, reducing the information content the model can learn from. Estimating the model at the carrier level therefore allows us to test whether exploiting this additional intra-route variation improves predictive performance. The forecasting itself is conducted at the route level, but the carrier-level estimation provides an empirical check on the value of supply-side disaggregation.

In this sense, the carrier-level model is not meant to recover airline-specific demand curves; rather, it is a data-treatment choice used to evaluate whether greater granularity in supply inputs enhances model robustness and predictive accuracy relative to the aggregated specification.

Overall, by testing both the approaches we highlight the importance of selecting appropriate model formulations and specifications to accurately capture demand dynamics. The semi-logarithmic approach, coupled with detailed comparisons between carrier-level and aggregated models, provides an analytical framework for analyzing and forecasting air passenger demand in the context of evolving competition and market conditions.

4.4 Econometric results

The results from the semi-log formulation, presented in Table 9[28], provide several insights into the factors influencing air passenger demand in the presence of HSR competition. The coefficient estimates exhibit distinct variations between the carrier-specific and aggregated models. These variations underscore the differing dynamics captured by each modeling approach.

Distance shows a positive effect on demand in all model specifications. Notably, the effect is significantly larger in the aggregated model for M1 (1.27) compared to the carrier-specific model (0.18). This suggests that on an aggregated level, distance plays a more

[28]Variables denoted in bold characters are those without the application of the log transformation.

pronounced role in influencing demand, as expected[29].

When considering M2 specification, the inclusion of lagged demand shows positive coefficients as expected (0.11 for carrier-specific and 0.74 for aggregated), indicating that previous period demand positively impacts current demand, with a stronger effect in the aggregated model. On the other hand, when considering M1 specification socio-economic variables coefficients are taken into account. Population has a positive impact on demand. This aligns with expectations as larger populations generate higher demand for air travel. For population, also the coefficients of the compound annual growth rate are positive across all models, indicating that growing populations drive demand.

We include the Herfindahl–Hirschman Index (HHI) and route frequency with 1 year lag as variables to account for market structure and service intensity, both of which are key determinants of air travel demand. While these variables are potentially endogenous—since low-demand markets naturally support fewer airlines and flights—the use of lagged HHI and frequency mitigates simultaneity bias in the time dimension, ensuring that current demand does not mechanically affect past market structure. The incorporation of these variables improves the model's explanatory power and helps capture realistic cross-sectional and temporal variation in demand. This is particularly important for forecasting future scenarios, such as the introduction of new HSR lines, where ignoring market concentration or service levels could distort predicted outcomes. Including HHI and frequency aligns with established empirical practice, allowing us to control for market characteristics while focusing on the causal interpretation of other covariates.

The presence of HSR negatively impacts air travel demand across all models specifications, underscoring the competitive pressure HSR presence imposes on air travel. The variables HSR travel time ratio and HSR feeding provide further insights into HSR's role in shaping air travel demand. The positive coefficient for the HSR travel time ratio - which incorporates both non-stop and one-stop itineraries- suggests that shorter HSR travel times are linked to lower air travel demand, reflecting the attractiveness of faster

[29]Note that distance is a market related determinant, hence influencing more the total city-pair market demand than the demand captured by individual airlines

HSR journeys as an alternative mode of transport. The HSR feeding variable, which measures the connectivity of an area by the number of cities linked via HSR, primarily reflects enhanced HSR accessibility. In the context of our study, higher connectivity increases the options for travelers to reach destinations by HSR rather than by air, effectively substituting for air travel. This interpretation is consistent with the negative coefficients observed in our models, ranging from -0.22 to -0.37 in M2 and from -0.31 to -0.28 in M1, highlighting how well-connected HSR hubs can amplify the competitive impact on air travel demand. As demonstrated also by the positive coefficient of age HSR variable in M2 models, this "demand stealing" effect is most acute immediately following the introduction of a new HSR route, as early adopters quickly shift to HSR. When a new HSR route is introduced, it often results in a significant shift in consumer preferences. The immediate availability of a faster, more convenient rail alternative leads to a substantial reduction in air travel demand along the same routes. Over time, the market adjusts, and the impact stabilizes, with remaining air travel demand becoming less sensitive to HSR presence. In M2 specification, a positive coefficient for the variable age HSR indicates that the introduction of new HSR routes has a pronounced and disruptive effect on air travel demand.

Understanding the temporal dynamics of HSR's impact on air travel demand is crucial for demand forecasting and strategic planning. The initial negative impact of new HSR routes highlights the need for airlines to anticipate and adapt to these disruptions quickly. Strategic adjustments in scheduling, pricing, and service offerings can help mitigate the initial loss in demand. Over the longer term, airlines can leverage the stabilized demand to refine their network and service strategies, ensuring they remain competitive in a market where HSR is a significant player.

The key performance indicators (KPI) indicate that the aggregated model with M2 specification is the most accurate one in terms of R2, mean absolute percentage error (MAPE) and mean absolute error (MAE).

	M1		M2	
	Carrier	**Aggregated**	**Carrier**	**Aggregated**
Distance	0.18***	1.27***	0.19***	0.95***
	(0.03)	(0.15)	(0.02)	(0.10)
Lag demand			0.11***	0.74***
			(0.01)	(0.08)
Pop	0.05*	0.24***		
	(0.02)	(0.06)		
GDP	-0.01***	-0.87*		
	(0.003)	(0.45)		
CAGR pop	1.98***	2.02**	1.24***	1.96**
	(0.40)	(0.85)	(0.32)	(0.78)
CAGR GDP	-0.69	-1.9	-0.08	-0.53*
	(0.52)	(0.21)	(0.09)	(0.28)
Lag HHI	-3.14*	-4.68***	-2.63*	-1.76***
	(0.06)	(0.08)	(0.04)	(0.05)
Delta freq	0.003***	0.03***	0.004***	0.05***
	(0.0004)	(0.01)	(0.0005)	(0.01)
Delta freq comp	-0.001**		-0.001**	
	(0.0003)		(0.0004)	
Age HSR	0.11**	0.57**	0.08**	0.28**
	(0.05)	(0.22)	(0.04)	(0.13)
Travel time ratio	0.42**	0.58***	0.38**	0.67***
	(0.18)	(0.15)	(0.16)	(0.14)
Trend	0.21*	0.19*		
	(0.11)	(0.10)		
Feeding HSR	-0.31**	-0.28***	-0.22**	-0.37***
	(0.14)	(0.08)	(0.09)	(0.07)
Nr obs	363.327	129.596	363.327	129.596
R2	0.53	0.65	0.67	0.83
MAPE	0.82	0.76	0.74	0.21
MAE	3005	1166	889	701

Table 9: Results from linear and semi-log formulation

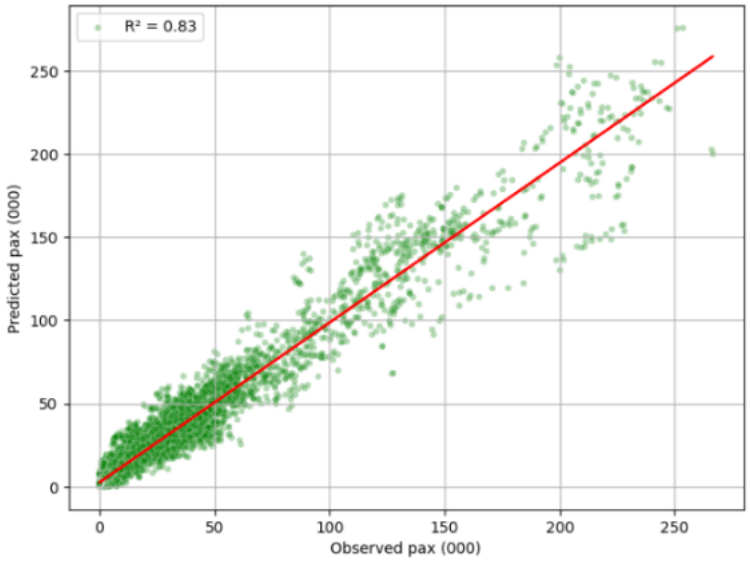


Figure 10: M2 aggregated model specification results using test set (2018-2019) for measuring the predictive performance (results has been linearized for the representation).

By considering the most accurate model with the aggregate-carrier specification (following what we stated in Section 4.3.3), we test the model's performance demand in years 2018-2019 and compare the predicted passengers demand with the observed passengers demand (see Figure 10).

Additionally, Figure 11 illustrates the relative importance of variables considered in the predictive model. The four model specifications (M1, M2 with carrier and aggregated formulations) are compared to assess the significance of each variable. The color intensity reflects the normalized impact of each variable, with lighter shades indicating greater importance.

The importance of each variable in the plot is assessed based on the absolute magnitude of the estimated coefficients, normalized relative to the largest coefficient within each model. This approach allows for a comparison of the relative influence of different predictors on the dependent variable. The heatmap visualization highlights which variables exert the strongest effects, providing an indication of their relative impact[30].

In conclusion, it underscores the substantial influence of HSR presence and competi-

[30]It is important to note that this measure reflects the magnitude of the effect and does not directly indicate the amount of variation explained or the statistical significance of the coefficients. Therefore, the results should be interpreted as a ranking of variables by their relative influence rather than by predictive power or statistical certainty.

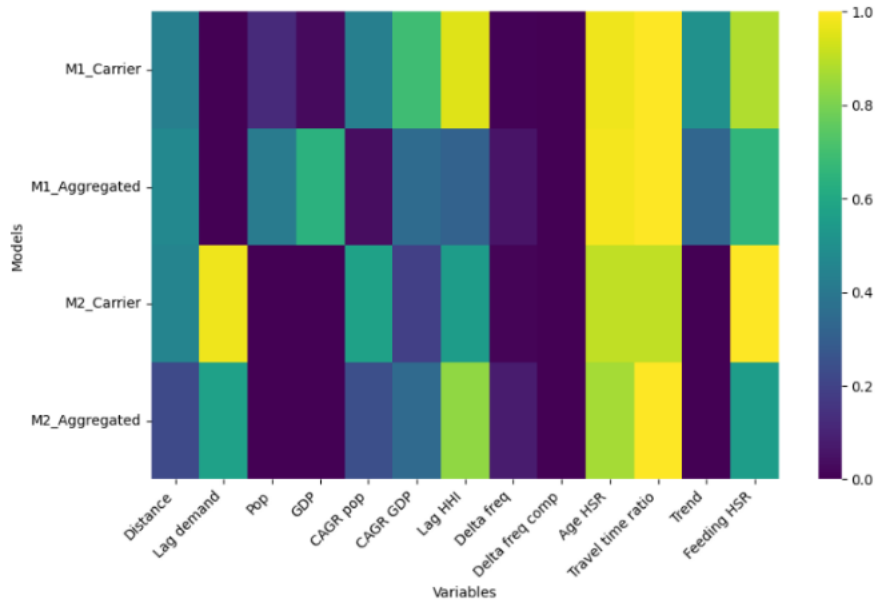


Figure 11: Variables importance level across the four models configurations (colors correspond to the relative importance of each variable, with lighter shades indicating higher importance).

tive dynamics on air travel demand. Variables related to HSR and its competition exhibit high levels of importance across different model configurations. This highlights the need and the pivotal role played by HSR data and competitive elements when modeling Chinese domestic air travel demand. By doing so, they can better anticipate shifts in passenger behavior and adjust their strategies accordingly, ensuring more precise forecasting and effective resource allocation in a rapidly evolving transport landscape.

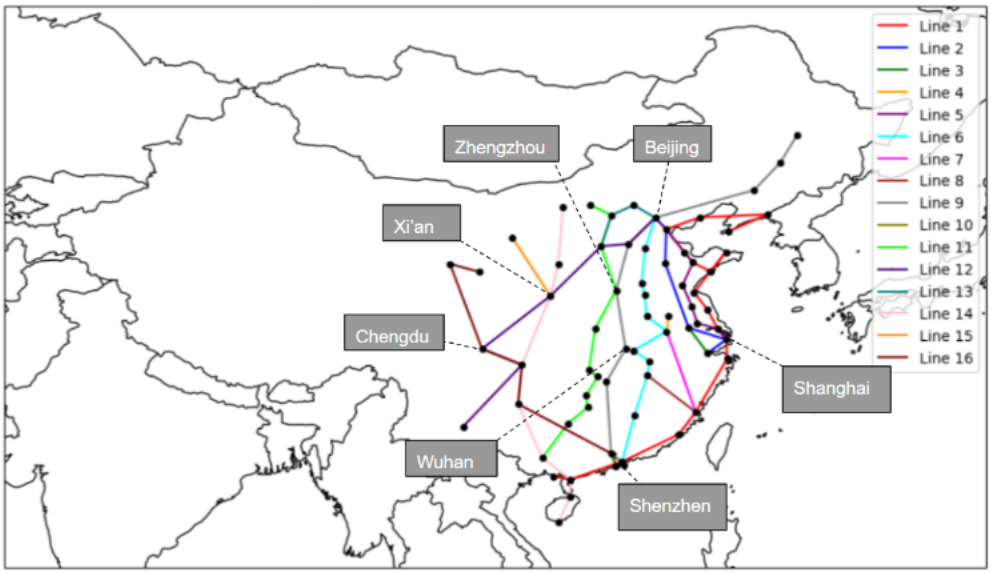
To summarize, the analysis of the semi-log formulation models reveals significant insights into how different factors influence air passenger demand amidst HSR competition. Our analysis revealed a significant negative impact of HSR on air travel demand across our sample. The introduction of HSR routes initially led to a sharp decline in air travel demand, particularly pronounced during the early stages following route implementation. This finding underscores the disruptive influence of HSR on air routes.

4.5 Case study: 2030 railway network expansion

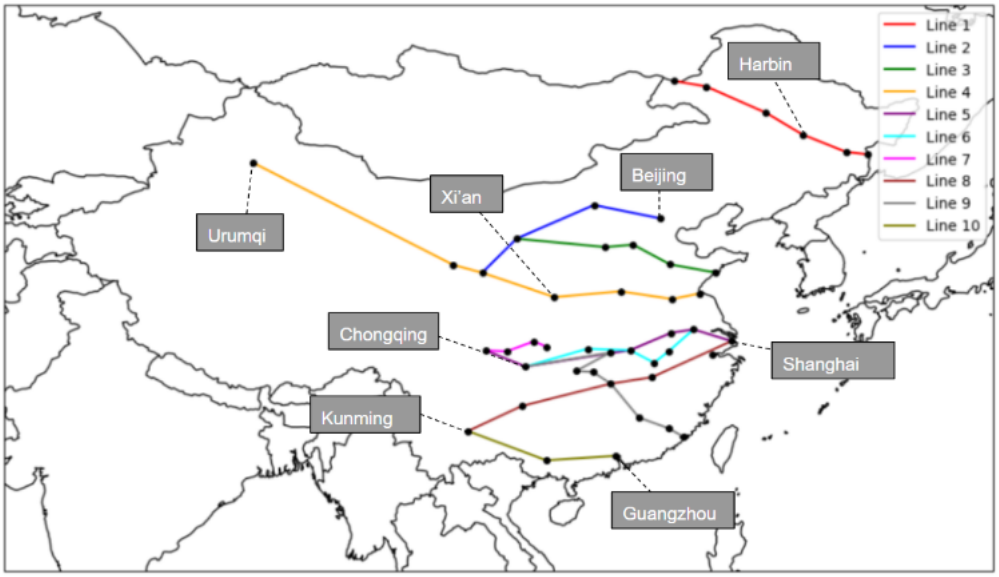
This case study examines the impact of the planned 2030 railway network expansion on transportation dynamics in China. Section 4.5.1 focuses on providing a preliminary evaluation regarding the changes in travel time resulting from the expansion of the HSR network. It compares the new travel times offered by HSR with those of existing air travel routes, highlighting the competitive advantages and potential shifts in transportation preferences. This analysis provides insights into how the expanded network will alter the accessibility and convenience of travel between cities. Section 4.5.2 explores the projected impact of the HSR expansion on future air travel demand, applying the forecasting methodology developed in Section 4.3.3. By analyzing trends and forecasting models, this section assesses how the introduction of new HSR connections is likely to affect air travel volumes in the next two decades. Together, these subsections offer an insights into the anticipated effects of the 2030 railway network expansion on travel times and air travel demand.

4.5.1 Comparative travel time assessment

The case study presented in this section focuses on the Chinese domestic HSR routes scheduled for opening in 2030, as shown in Figure 12.



(a) Vertical connections



(b) Horizontal connections

Figure 12: Backbone HSR lines planned to be opened in 2030. Subplots divide the network lines into vertical connections (a) and horizontal connections (b).

In addition to applying our demand forecasting model to these new market connections, we conduct a preliminary accessibility evaluation. This evaluation compares land-side travel times, which reflect the ease with which potential passengers can reach airports and HSR stations. Given the intense competition between airports and HSR in China (**?**), ground travel time emerges as a crucial factor in passenger decision-making, alongside other considerations such as price, service frequency, and scheduling. The in-

creasing competition of HSR is also related to its very high reliability combined with the fact that airports are often located far from the center of the city. For example, between Beijing and Nanjing (1000 km) the HSR travel time is about 3.5 hours, compared to 2 hours of air; however, the Nanjing airport is 47 km from the city center, whereas the HSR station is only 10 km from downtown. After factoring in this longer access distance, together with airport processing, passengers find the HSR service is very competitive and now on Beijing-Nanjing route it has the 60% of the market share. The reliability, frequency and comfort of HSR service create strong competition for most middle-distance trips.

Building on the work of Sun et al. (2021a), we estimate multi-modal accessibility, which accounts for both automobile driving time and public transit options. This composite indicator reflects the minimum travel time required for passengers to reach their destination, considering either driving or public transit, with the city center serving as the origin and destination in both the departure and arrival cities. By combining these modes of transport, this indicator provides a realistic measure of accessibility, capturing the ease with which passengers can reach an airport or HSR station using the available transportation infrastructure[31]. Figure 13 displays the total travel itinerary that the passenger is supposed to perform. We measure travel time assuming the passenger begins the journey from the city center, based on the premise that city centers typically concentrate the highest population densities and major touristic attractions[32]. This approach recognizes the consumer's utility or disutility in selecting air or rail transportation based on the location and connectivity of stations.

A HSR efficiency index in terms of accessibility score is calculated as $\frac{HSRtraveltime}{airtraveltime}$. It represents the comparative travel time advantage, with values lower than 1 indicate an

[31]For travel time estimation, we integrate public transportation data from OpenStreetMap (OSM) and driving distance estimation with Open Source Routing Machine (OSRM), which is based on the road network encoded in OSM. These data are used to estimate car driving time and to extract public transit networks. For public transit connections, travel time is modeled as $\frac{distance}{50km/h}$.

[32]Assuming as reference point the city center allow us to minimize the computational time, without affecting the problem's scalability. However, this approach may present limitations, as it does not fully capture the spatial heterogeneity of urban areas. Further works could assume different points of interest and administrative levels (such as the working places) or assume a grid-based approach.

HSR advantage over air travel, while values higher than 1 indicate an air travel advantage over HSR. Additionally, different dwell times (applied as penalization weights on the overall travel time) have been tested to simulate check-in waiting times at both airports and rail stations and they have been summed as part of the total travel time (refer to Koppelman et al. (2008),Moyano et al. (2018) and Wang et al. (2020a) for similar approaches). We tested several scenarios, with waiting times for HSR ranging from 30 to 60 minutes and waiting times for air ranging from 90 to 120 minutes. Therefore, for each travel from a city i to a destination city j and for each transport alternative (air vs rail), we can define the total travel time (city center-to-city center) t_{ij} along the path with the shortest possible time. It consists of two parts, namely the in-vehicle[33] (i.e., in-flight or in-rail travel time) t^V and waiting time at airports/stations t^W. Hence the total travel time t between any given city pair can be calculated as in Eq. 4.5.1:

$$t = t^V + t^W \tag{9}$$

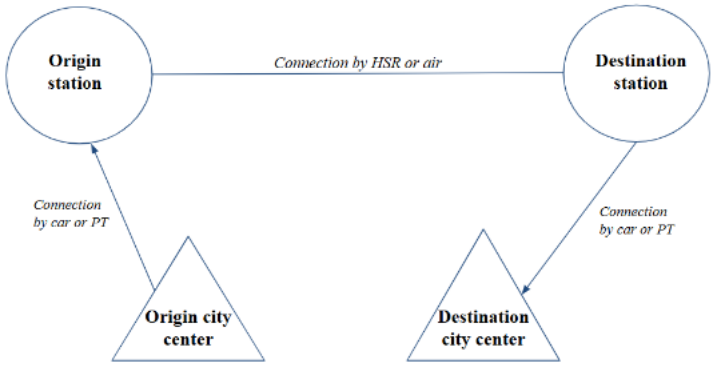
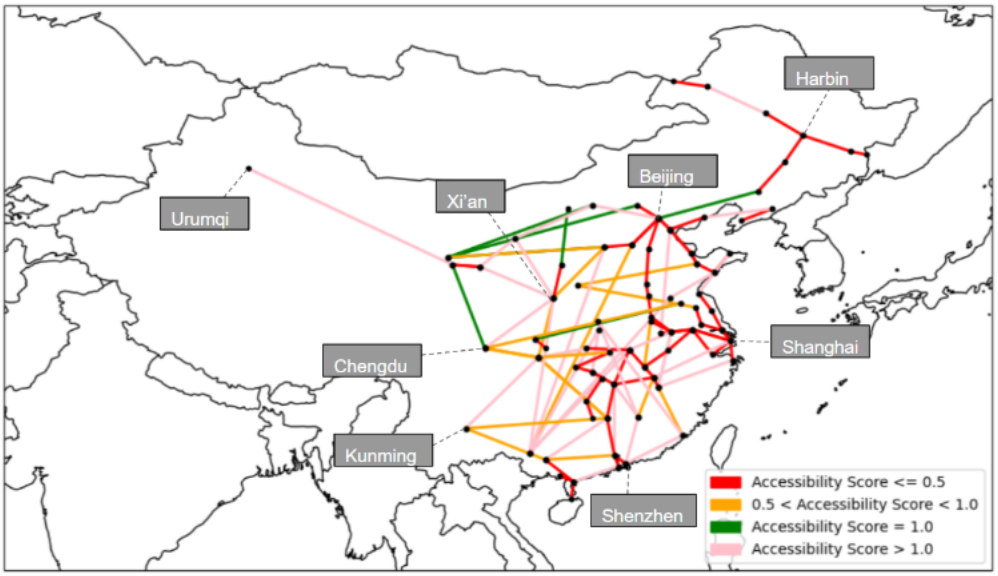


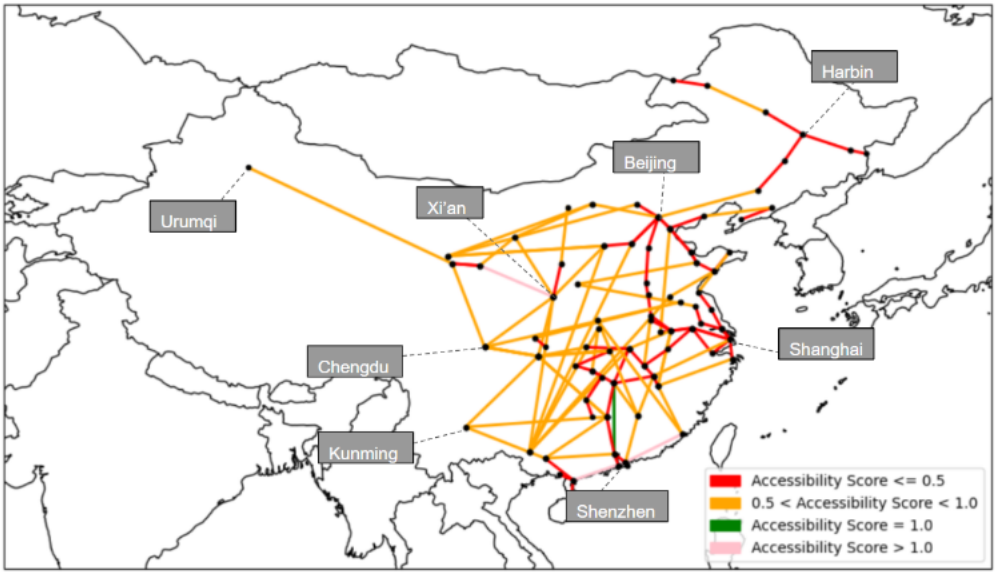
Figure 13: Total travel itinerary for a passenger supposing to start and end the trip to the city center, either by HSR or air with the support of car or public transport for the segment city center-station, and vice versa.

Figures 14 and 15 show comparative travel time scores among market pairs resulting from our analysis.

[33]Flight/HSR travel time t^V reflects the actual travel time spent on flights/trains and primarily depends on the distance between origin and destination as well as the operating speed of the flight/HSR.

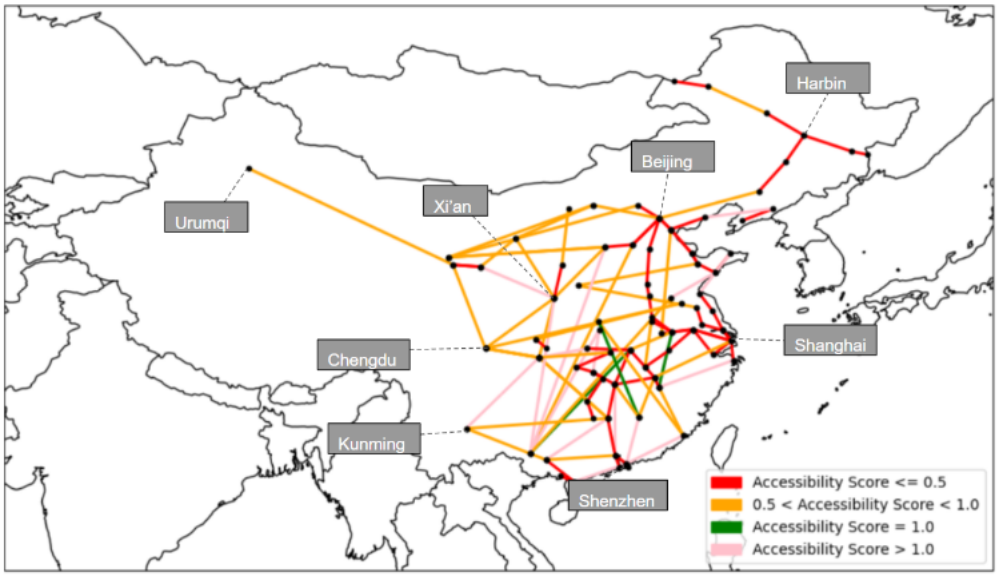


(a) No dwell time

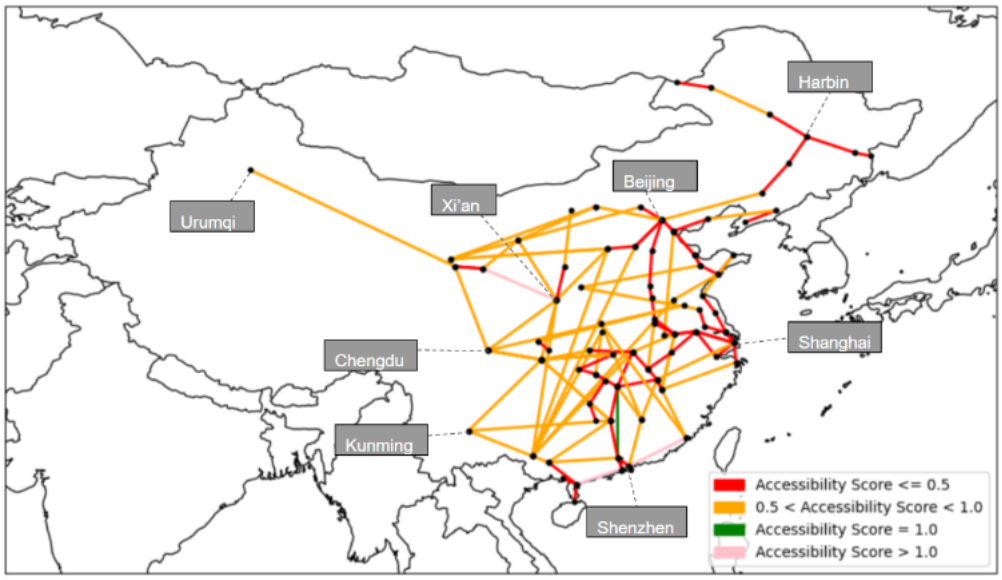


(b) Dwell time 30 minutes for HSR and 90 minutes for air

Figure 14: Comparative travel times along the backbone HSR lines 2030, with different penalty time scenarios. Subplots (a) and (b) show: (a) No dwell time applied; (b) Dwell time of 30 minutes for HSR and 90 minutes for air.



(a) Dwell time 60 minutes for HSR and 90 minutes for air



(b) Dwell time 60 minutes for HSR and 120 minutes for air

Figure 15: Comparative travel times along the backbone HSR lines 2030, with different penalty time scenarios. Subplots (a) and (b) show: (a) Dwell time of 60 minutes for HSR and 90 minutes for air; (b) Dwell time of 60 minutes for HSR and 120 minutes for air.

4.5.2 Air demand future trend

To predict Chinese domestic air travel demand in 2030, we employ the forecasting model developed in Section 4.3.3. This model uses coefficients from the best-performing model, M2-aggregated, as identified in Section 4.4. We apply the model on the subset of cities that in 2030 will be connected by HSR (as shown in Figure 12) and that in the past years

are already connected by air transport.

For the real-world data that will be unavailable from 2025 on, we adopt various elaboration methods regarding the type of data. For data related to aviation scheduling (such as frequency supply and the number of seats offered by carriers), we conduct a scenario analysis, simulating a trend projection under the assumption that the future trend will follow the historical growth trend. For socio economic variables we apply future projections derived from the most extant literature.

Relatively to aviation data, to predict air demand in 2025 we use values of frequency and market concentration observed in 2024[34]. For subsequent years until 2030, we apply annual growth parameters of approximately +5% for flight supply and +2% for market concentration[35]. The rationale for using HSR-exposed routes rather than relying solely on each new route's own historical growth is that several markets in our sample exhibit short or irregular historical aviation series, with limited variability in supply. Using a common benchmark derived from a larger, structurally similar set of markets helps impose consistent trends in the pre-HSR period. Nonetheless, this approach implies the assumption that aviation dynamics in these to-be-connected routes would follow patterns similar to those observed in cities with indirect HSR competition, which should be interpreted with caution. Starting from 2030—the year in which direct HSR services are assumed to become operational—we apply annual decrease parameters of –3% for flight supply and –2% for market concentration[36][37]. By relying on historical patterns—even if imperfectly observed—we aim to construct a plausible evolution of key aviation supply variables before and after HSR introduction. Still, the projections should be viewed as scenario-based rather than precise forecasts, and the associated assumptions are explicitly

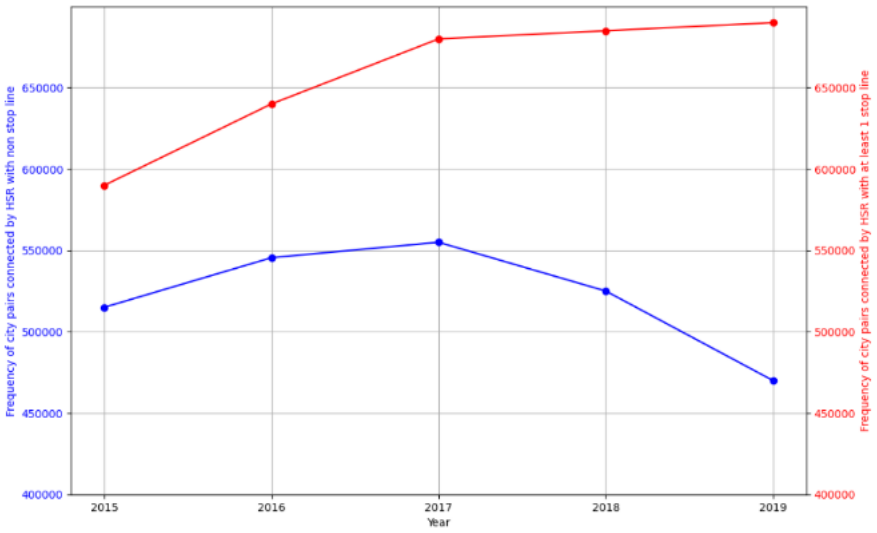
[34]Those variables in the M2-aggregated formulation are used with a one-year lag.

[35]These parameter values are extrapolated from the historical evolution of supply variables in city pairs connected by HSR with at least one-stop itineraries, as shown in Figure 16.

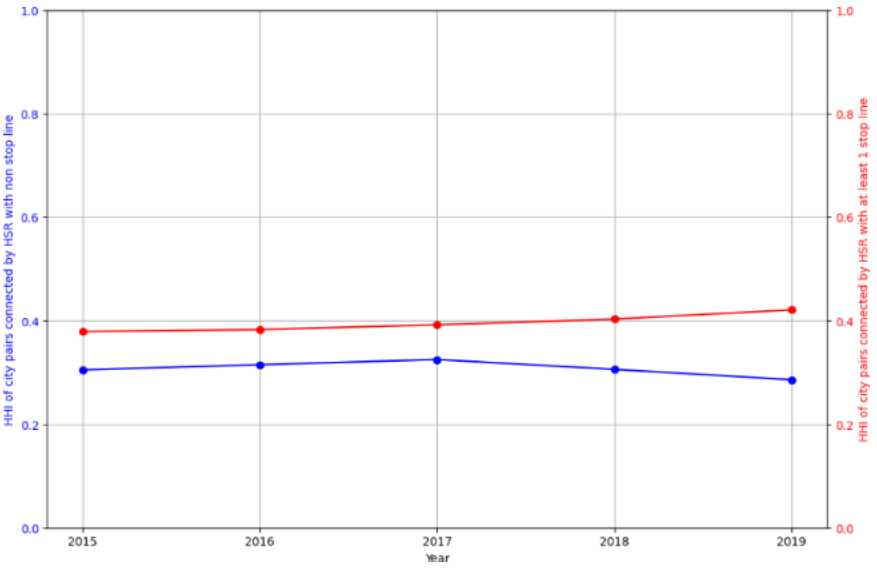
[36]These parameter values are extrapolated from the historical evolution observed in cities connected by HSR through both non-stop and one-stop itineraries, as shown in Figure 16.

[37]However, it is important to acknowledge a limitation inherent to this approach. The sample used to infer these trends contains only periods after HSR has already entered those markets; as a consequence, the immediate post-entry drop in frequency and capacity documented in earlier literature is not fully captured. This limitation reflects a constraint of the available data and should be taken into account when interpreting the magnitude of the projected supply adjustments in the counterfactual scenario.

acknowledged as part of the modelling framework.



(a) Frequency historical growth from 2015 to 2019.



(b) HHI historical growth from 2015 to 2019.

Figure 16: The plots show the historical evolution of frequency and market concentration (HHI) over the period 2015-2019 in two different scenarios: blue lines display the growth of city pairs also connected with non stop HSR line, red lines display the growth of city pairs also connected with at least 1 stop connecting HSR line.

Regarding socio-economic variables projections, we rely on data estimations provided by Chen et al. (2020), Wang and Sun (2022) and Wang et al. (2022a). The data projections are based on the shared socioeconomic pathways scenario (SSPs). The SSPs are defined as a set of future pathways of societal development that describe five alternative

outcomes of trends in demographics and economic development, provided by the International Institute for Applied System Analysis (IIASA). Table 10 provides a comprehensive overview of the five scenarios and their different characteristics.

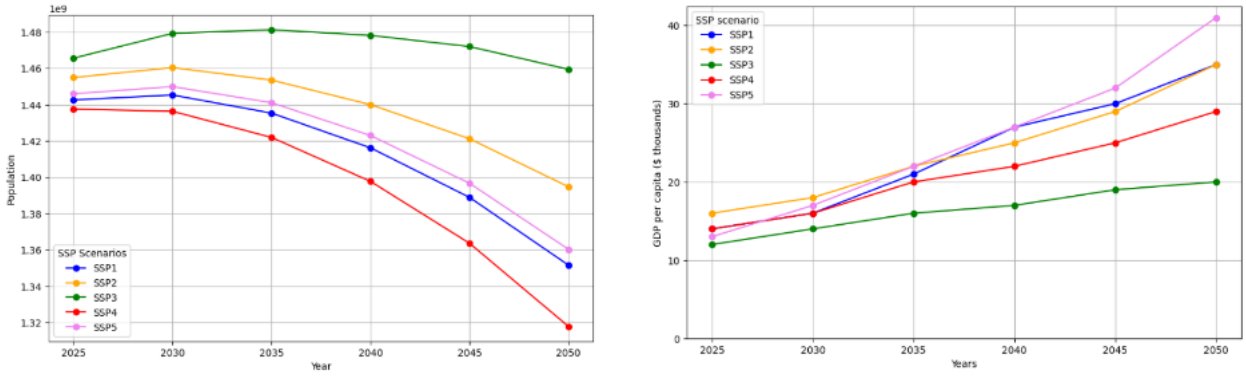
Scenarios	Fertility	Mortality	Migration	Education	Policies
SSP1	Low	Low	Medium	High	Ineffective fertility policy
SSP2	Medium	Medium	Medium	Medium	Effective two-child policy
SSP3	High	High	Low	Low	Effective fully open policy
SSP4	Low	Medium	Medium	H/M/L	Ineffective fertility policy
SSP5	Low	Low	High	High	Ineffective fertility policy

Table 10: Demographic assumptions in China under SSPs. In SSP4, assumptions for education depend on the provincial development level.

The provincial socioeconomic projections take into account fertility promoting policies and population ceiling restrictions of mega-cities that have been implemented in China in the last years. The raw data for each projected year are collected dividing population by gender (2 levels) and by level of education (7 levels), for the scope of this study the data have been aggregated by year and by province. Figure 17 summarizes the total Chinese population and GDP per capita projections from 2025 to 2050, based on the SSPs scenario.

Since our analysis focuses on the city level, we needed to disaggregate the provincial socioeconomic projections. We achieved this by first calculating the percentage of people living in each city of our dataset relative to their province during our last year of observation (2019). Then, using data from the World Bank, we retrieved urbanization increase parameters regarding China, which allowed us to apply the current percentage of people living in cities alongside urbanization projections. As more people are expected to move to urban areas in the coming years, this approach helps us accurately reflect the shifts in population distribution and the socioeconomic dynamics at the city level.

This methodology ensures that our analysis captures the granularity of urban growth and its implications for future scenarios. From the methodological perspective, we elaborate a framework based on the average population and GDP per capita values of the five scenario.



(a) Total Chinese population projections **(b)** Chinese GDP per capita ($) projections

Figure 17: Chinese population (plot A) and GDP per capita in $ thousands (plot B) projections from 2025 to 2050, under SSP scenario.

Figure 18 illustrates the projected increase in air travel demand for the year 2025[38]. The increments reflect typical changes observed in aviation scheduling and market concentration in the absence of non stop HSR itineraries, but with only connecting HSR itineraries available between cities. In this scenario, air demand is predict to exhibit an increase ranging from 5% to 50%, approximately. In particular, long haul itineraries will experience a higher demand increase. Interestingly, even short-haul itineraries, which are typically the most vulnerable to HSR competition, show an increase in air demand in this scenario, albeit at a lower rate than long-haul routes. This may suggests that while 1-stop HSR connections may appeal to some passengers, they do not offer sufficient utility to prompt a widespread shift from air travel. This result can be explained by considering factors inherent to the current travel context. The forecasting model indicates a negative coefficient for the feeding HSR variable, which would suggest a decrease in air demand when there is strong HSR connectivity between the origin and destination stations. However, in this scenario, the origin-destination itinerary for air travel is already a one-stop

[38]Please note that these routes are linked by one-stop HSR in 2025 and then non-stop HSR from 2030.

route, and if one-stop rail itinerary can be competitive with a non-stop air itinerary (even without showing a full-substitution effect of mode of transport), it is less likely that passengers would opt for a two-stop rail journey over a non-stop air journey. In conclusion, the strong negative power of the HSR feeding variable against the air demand is mostly effective regarding non-stop HSR connections.

Figure 19 showcases the decrease in air travel demand following the introduction of non stop HSR connection. The model predicts a substantial shift from air to rail transport as smoother and faster HSR options become available, leading to a complete disappearance of air travel demand in certain city pairs. These city pairs, which are projected to experience a total shift from air travel to HSR, have been labeled according to the year in which this transition is expected to occur. Notably, our analysis reveals that the shorter the itinerary, the sooner the demand shift effect is likely to manifest.

By comparing the predicted demand with the two development scenarios (Figure 18 and Figure 19), we highlight the impact of new rail infrastructure on air travel patterns. It is important to note that the city pairs experiencing a significant drop in air travel demand are the same ones that, in our comparative travel time analysis (see Section 4.5.1), were identified as benefiting most from HSR. In that analysis, these city pairs were projected to gain a clear advantage with HSR across various scenarios of dwell times. This alignment between the time advantage analysis and the predicted demand shifts highlights the reliability of our model and the transformative impact HSR will have on certain air markets.

Both figures offer a clear visual representation of these predicted shifts, providing critical insights into how the integration of HSR will reshape regional transportation dynamics. The prediction of a complete shift in demand from air to HSR in specific markets signals a potential end to air travel in these routes, emphasizing the importance of such predictive analysis in understanding and planning for the future of transportation.

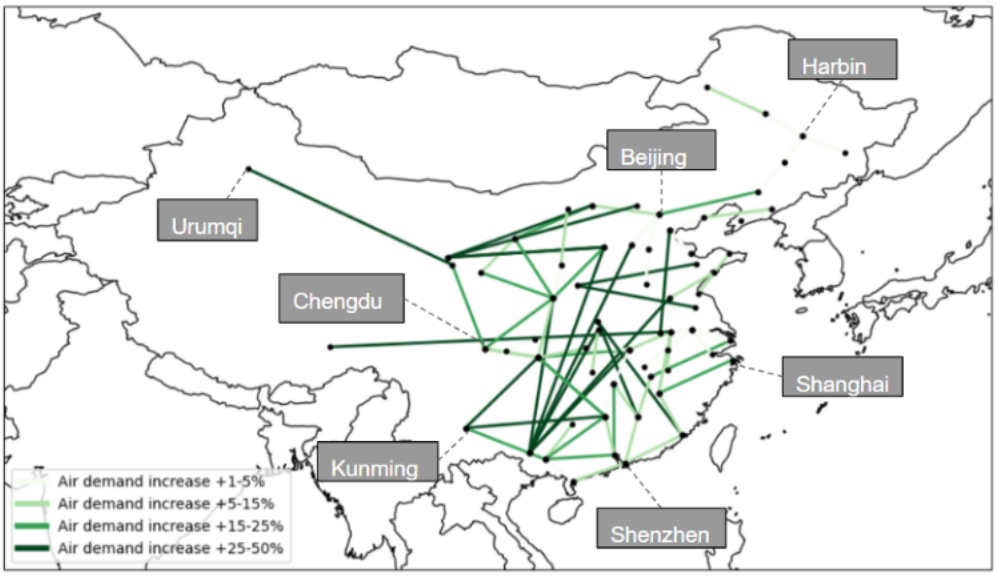


Figure 18: City pairs connection with color differentiation based on the air demand increase in 2025, without the new set of HSR lines opening.

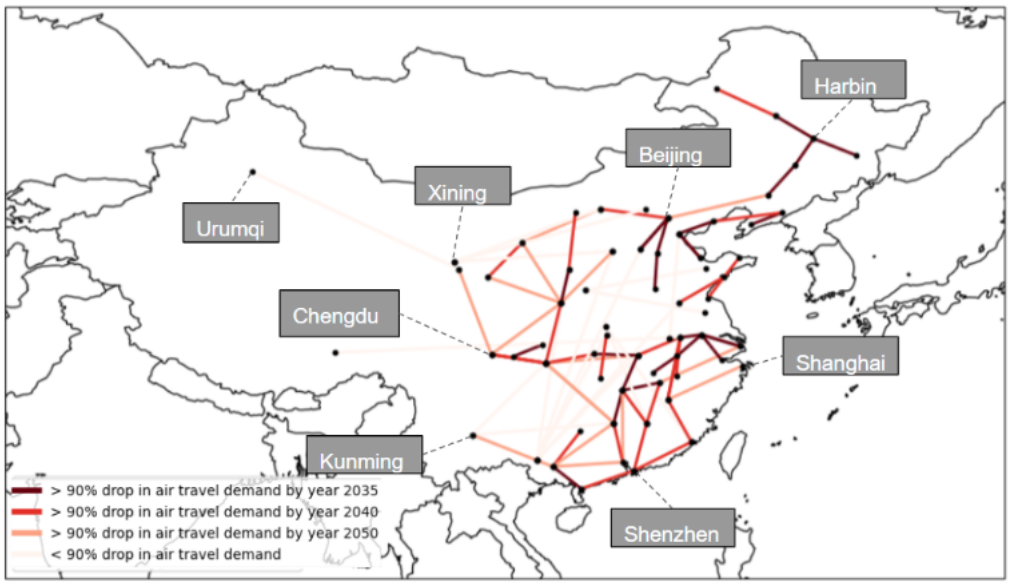


Figure 19: City pairs connection with color differentiation based on the air demand decrease and disappearing during the interval 2030-2050, following the HSR lines opening.

In the context of analyzing the shift from air travel to HSR, a violin plot in Figure 20 displaying the relation between distance and the year of total demand shift provides critical insights into the relationship between route length and the timing of this transition. The plot reveals how city pairs with varying distances are expected to experience a complete shift in travel demand from air to HSR over different years. This relation-

ship is visually represented by the shape and distribution of the violins, where denser sections indicate a higher concentration of city pairs within specific distance ranges. Insights gained from this analysis show and confirm that shorter routes tend to experience a total demand shift earlier, as HSR's competitive advantage is more pronounced over shorter distances. The plot also allows us to observe trends and clusters, indicating which air markets are most vulnerable to disappearing sooner due to the introduction of HSR. This visualization is instrumental in strategic planning, enabling transportation policymakers to prioritize infrastructure development in regions where the shift from air to rail is expected to be most significant.

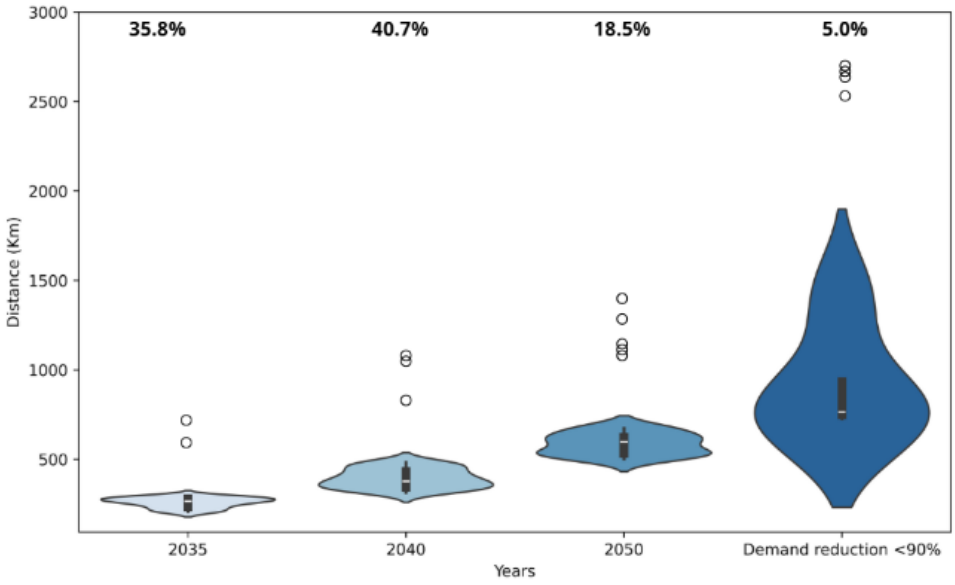


Figure 20: Relation between origin-destination distance and year experiencing the total demand shift in air-HSR.

While the model predicts a substantial decrease in air demand for the majority of city pairs exposed to HSR, a closer examination reveals that some OD pairs—particularly those below 800 km—are projected to retain a non-negligible level of air demand by 2050. These exceptions warrant further discussion to understand and make some assumptions on why certain routes are more resilient to HSR substitution. Several factors may contribute to this persistence. First, the geographic location of the cities relative to HSR stations and rail line connectivity can create local accessibility constraints, making

air travel more convenient or necessary for certain travelers. For example, if the HSR station is distant from the city center or requires additional transfers, the total travel time advantage of rail diminishes, allowing air services to remain attractive. Second, demand composition and passenger preferences play a role. Routes with a high proportion of business travelers or passengers valuing comfort and frequency may retain air demand despite HSR availability, as these travelers may prioritize non-stop, flexible, or premium air services over rail alternatives. Third, hub connectivity effects influence resilience: OD pairs that are linked to major airline hubs provide important onward connections for passengers traveling beyond the origin or destination city. Even if HSR provides a direct alternative, the loss of network connectivity can make air travel necessary for multi-leg itineraries. Fourth, some city pairs may be served by airlines with strategic incentives to maintain service on certain routes (e.g., fleet utilization, marketing presence, or political agreements), which can partially offset the natural shift to rail. In conclusion, while the model predicts a general trend of air-to-rail substitution, the persistence of air demand in certain OD pairs highlights the importance of local accessibility, demand composition, network connectivity, and strategic airline decisions in shaping transport outcomes. Acknowledging these nuances provides a more comprehensive understanding of future modal shifts and enhances the credibility of the forecast, emphasizing that HSR impacts are heterogeneous across routes even within the same distance bands.

4.6 Conclusion

In this study, we developed a demand forecasting model to analyze the impact of HSR expansion on air travel demand in the Chinese domestic context. The analysis particularly focused on the HSR lines opening by 2030 and aims to identify the aviation markets most likely to experience pressure following the introduction of these new rail connections.

By employing a robust econometric approach which integrates highly accurate and robust socio-economic projections and historical scheduling and traffic flight data, allows us to predict both the timing and the extent of demand shifts from air to rail transport

across various city pairs. The methodological rigor of this study plays a crucial role in accurately forecasting these shifts, offering significant insights into the evolving dynamics between air travel and HSR. The forecasting analysis reveals, with an accuracy ranging from 53% to 83%, that HSR exerts significant competitive pressure on air travel, with new rail routes leading to notable reductions in air demand, particularly in markets where HSR offers a more convenient alternative due to the lower travel time. In particular, the negative coefficients of feeding HSR variable, ranging from -0.21 to -0.37, highlight that the higher the HSR connectivity, the lower the air demand, meaning that often a connecting HSR itinerary is perceived as more convenient than non-stop air itinerary. Moreover, the positive coefficients of travel time ratio and age HSR variables, ranging from 0.38 to 0.67 and from 0.08 to 0.57 respectively, show that the lower the HSR travel time, the lower the air demand and the switching demand effect from air to rail is higher during the first phase of rail introduction. The predictive model has been completed also with the inclusion of aviation-related variables, such as the lagged demand, the market concentration index, the percentage increase in frequency and socio economic components, namely population and GDP, all showing positive coefficients on the air demand.

Our application of the demand forecasting model to the set of cities that will be connected by new HSR connections starting from 2030 reveals a profound shift in air travel dynamics, particularly for routes under 1000 km. In the scenario analysis for 2025, the landscape of domestic air travel reflects a period of growth, before the impact of HSR introduction is realized. Our forecasting results show that, in absence of direct HSR lines (only 1-stop rail connection is available), air travel across the subset of city pairs analysed is expected to increase, with different growth rates. Quantitatively, this translates into a projected increase in air travel demand ranging from 5% to 50%, depending on the specific route characteristics and socio-economic conditions. Long-haul itineraries, in particular, are expected to experience the most substantial demand growth. For instance, routes exceeding 1500 km are likely to see demand increases closer to the upper end of this range, around 40% to 50%. This growth reflects the continuing preference for air travel on longer

routes, where the absence of direct HSR competition allows airlines to capitalize on the growing population and economic activities. Nevertheless, by 2050 these shorter routes are projected to lose the majority of air travel demand as HSR becomes the dominant mode of transportation and likely for the airlines there will be not anymore convenient operate those itineraries. The speed, convenience, and city-center accessibility offered by HSR make it a highly attractive option for travelers. This shift not only marks a change in domestic travel patterns but also raises significant strategic challenges for airlines to adapt to this new competitive landscape, which may need to adapt their business or change their network to remain competitive.

While this study provides comprehensive insights into the impact of HSR expansion over air travel demand in China's domestic market, it is important to acknowledge limitations and improvements that can be addressed in future studies. First, from a conceptual standpoint, one limitation lies in the assumptions made about the competitive dynamics within the aviation industry and passenger travel behavior. While the model effectively captures the competitive pressure exerted by HSR on air travel, it assumes that travelers consistently choose the most time-efficient mode of transport. Additionally, it presumes that airlines will not counteract HSR competition through alternative strategies, such as enhancing brand loyalty, offering frequent flyer programs, or other business tactics. Second, from a methodological perspective, the model's heavy reliance on the accuracy of long-term data projections poses a significant limitation. Given that the study involves a prediction spanning two decades, the model does not account for potential disruptions such as economic downturns, unforeseen technological advancements, or other unexpected events that could significantly alter demand dynamics. Future research could address these limitations by incorporating more granular data, exploring alternative behavioral models, and validating findings against real-world developments as new HSR lines become operational. Despite these challenges, this study provides a crucial foundation for understanding the evolving dynamics between air and rail travel, offering valuable insights for policymakers, transportation planners, and the aviation industry.

Chapter 5

Airlines Network Development: A Worldwide Empirical Investigation

5.1 Introduction

The exponential growth of airline networks in recent decades has significantly contributed to the economies of cities and countries. Central to this dynamic is the strategic expansion of airline route networks. By strategically adding new routes and enhancing connectivity, airlines are consistently committed to improving their service quality, ensuring that most passengers can access their desired destinations within the airline's network or through alliance partners (Wong et al. 2023). However, launching new routes represents a strategic investment that extends beyond the mere geographic expansion. By constantly re-designing their network, airlines aim to preserve their competitiveness, strengthen their market position, and augment the overall value of their route portfolios.

As global connectivity expands, the addition of a new route can influence not only the specific market where the connection is introduced but also, through cascading effects, other markets that the new leg can serve. Additionally, a new route may significantly alter current passenger flows by shifting transit options to different hubs or attracting demand previously served via connecting itineraries. These complex network effects, combined with fleet capacity constraints and strategic objectives to pursue, make airline network development anything but trivial, with the prior literature not finding a clear consensus about the set of factors affecting the choice and their respective importance within the airline decision-making process.

Despite its strategic relevance, data-driven tools for systematically analyzing new

route selection remain scarce. In practice, most airlines still rely on case-specific market analyses or simplistic route profitability models (Bannò and Redondi 2014, Halpern and Graham 2015, Carmona-Benítez et al. 2017). In such a context, a promising decision-support tool is route network optimization models, which can assist airlines in determining which routes to launch or discontinue and which markets to prioritize. This approach would provide superior benefits with respect to traditional methods due to the ability to simultaneously consider multiple factors and effectively assess how each route fits within the overall airline network. However, airlines face substantial challenges in extensively employing network planning algorithms due to their high computational demands. The main concern is the combinatorial explosion of potential itineraries: as the number of airports and potential routes grows, the number of possible connections increases exponentially (for details see Section 2.1). This results in a vast search space that is computationally challenging to navigate exhaustively, limiting the applicability of such approaches in many real-world scenarios. At the same time, scholarly attention to analytics-based solutions for identifying new route development is relatively sparse, with most studies attempting to tackle the combinatorial complexity by considering relatively small-scale networks (Hausladen and Schosser 2020, Birolini et al. 2021).

In this paper, we propose a classification model to identify promising new long-haul routes based on worldwide publicly available data. The model provides a comprehensive examination of the multifaceted factors and dynamics that influence airline route entry decisions. Furthermore, it can be used to evaluate the network development potential of individual carriers and allows scalability to be integrated with existing network optimization algorithms to refine route selection strategies.

More in detail, we formulate an analytical classification logit model designed to systematically compare characteristics of newly launched routes with potential ones. We leverage a comprehensive dataset of new routes opened between 2014 and 2019 on a global scale and implement a tailored procedure to build instances of potential routes that could have been opened but were not. We engineer features to capture key drivers

–including market potential, competition, and network dynamics– to uncover the fundamental factors influencing airlines' network development decisions. Given the highly unbalanced nature of the dataset due to the scarcity of new route openings, we test different resampling procedures, ultimately selecting the most robust approach. By proposing carrier-specific formulations, we conduct a cross-carrier comparison to investigate the pivotal factors influencing route decisions. We found substantial heterogeneity among carriers, explaining how different carriers have different route planning strategies. The model's performance is validated through tailored out-of-sample testing, demonstrating variations in key performance indicators as different probability threshold values are considered. Lastly, we demonstrate the real-world applicability of the model to support network planning through real-world case studies where the model is used to predict new route openings based on 2019 data and then compared with actual routes launched by the airlines in subsequent years. The results underscore both the predictive accuracy of the model and the practical utility of the model in selecting promising routes to be further investigated within network planning algorithms.

The rest of the paper is organized as follows. Section 5.2 discusses the relevant literature. Section 5.3 details the sample identification, features engineering, and methodology. Section 5.4 presents the empirical results and introduces evidence from the tailored out-of-sample validation experiment. Lastly, Section 5.5 concludes and provides avenues for future research.

5.2 Related Literature

5.2.1 Route entry decision

The literature examining airlines' route entry decisions from an empirical perspective is quite substantial. The aim of these contributions is to infer drivers underpinning airlines' route entry decisions by ex-post investigating market characteristics and competitive environments where such events occur. As anticipated in Section 2.1, the key

drivers influencing route entry decisions have been categorized into socio-economic factors, market competition, and network considerations (Abdelghany and Guzhva 2010, Halpern and Graham 2015, Hanson et al. 2022). Socio-economic factors include variables such as population size, income levels, and regional economic growth, which ultimately determine the potential market size for a given route. Market competition reflects the competitive pressure within the market and depends on the number of competitors, their fare structure and aspects that collectively influence the attractiveness of entering the market. Lastly, network considerations encompass a broad spectrum of factors related to the role of hub connectivity, route density, and market coverage. These factors are often airline- or airline group-specific and are shaped by airline strategy and market coverage (e.g., geographical or market segment focus) subject to fleet availability constraints (Oliveira 2008, Zhang et al. 2017b).

Although many studies have explored these drivers, there is no consensus on their relative importance. Prior research is often limited by data availability and tends to focus on specific airlines or regions, leading to case-based findings that lack broader generalizability and fail to identify universal patterns across markets (Fu et al. 2015, Calzada and Fageda 2019, Gaggero and Piazza 2021). The limited capability to collect relatively large datasets also constrains most of the studies to the investigation of short-haul markets, where the presence of seasonal connections is higher and the market more dynamic (Boguslaski et al. 2004, Zou and Yu 2020).[39] The ex-post evaluation approach also limits the analysis of factors determining airlines' entry into already established markets rather than also investigating and systematically comparing the characteristics of fist-time served markets (i.e., market not served via no stop services before) against the whole set of potential routes. Only recently, some studies targeted the determinants of first-time served market focusing on their managerial implications (Abdelghany and Guzhva 2022, Wong et al. 2023).

In summary, existing studies, existing studies offer useful insights into specific carriers

[39]The stronger presence of low-cost companies in this travel segment further facilitates the collection of larger datasets due to their practice of rotating leisure destinations.

or regions but lack global applicability and fail to capture the full complexity of airline route planning. They often overlook the strategic factors driving entry decisions across diverse markets and pay limited attention to long-haul routes, which are central to hub-and-spoke systems and global connectivity. (Lee et al. 2014, Cheung et al. 2022).

5.2.2 Strategic network planning

Route planning is a core element of the airline planning process. This process involves a complex interplay of strategic, tactical, and operational decisions aimed at maximizing airline profitability while ensuring the achievement of business goals and required flexibility (as previously discussed in Section 2).

Academic literature devoted substantial effort to developing analytical models to support tactical and operational decisions. Relevant examples are contributions focusing on flight scheduling and fleet assignment, crew scheduling, aircraft routing, and recovery decisions (Santana et al. 2023, Xu et al. 2024, Wen et al. 2024). In contrast, strategic network planning received lower attention. In this domain, most academic research has focused on hub location and fleet planning problems (Mohammadi et al. 2019, Soylu and Katip 2019, Alumur et al. 2021). Research in the former area involves determining the location of hub airports to connect passenger flows from origin to destination nodes by modeling the network architecture (hub & spoke vs. point-to-point) that maximizes profits or minimizes costs (Alumur et al. 2021, Sharma et al. 2024). Despite the important insights from a theoretical perspective, from a managerial perspective this problem assumes less relevance since hub locations have generally remained fixed for most carriers since their inception. Fleet planning has instead been explored by proposing optimization models to support when to dismiss/sell vs. acquire/lease aircraft taking into account factors such as demand uncertainty, purchasing and leasing alternatives, and operating costs (Teoh and Khoo 2016, Baykasoğlu et al. 2022).

Besides hub location and fleet planning, airlines as part of their long-term strategy need to frequently evaluate which routes to open, close, or connect. At the state of the

practice, most of the airlines identify which market to enter based on market analysis, basic econometric models of air travel demand, or simplistic route profitability models (Bannò and Redondi 2014, Halpern and Graham 2015, Carmona-Benítez et al. 2017). Despite the critical importance of these decisions in shaping airline networks, route planning has received limited attention in the academic literature with only a few models tackling airline network expansion design (Kölker and Lütjens 2015, Birolini et al. 2021). The main reason is the high computational cost of network planning algorithms, which makes it difficult to apply them to real-world scenarios (Carreira et al. 2017, Schosser and Schosser 2020). This follows from the combinatorial complexity inherent in network planning, where the vast number of potential route combinations makes it challenging to develop efficient algorithms capable of identifying the optimal network structure. The few studies that address this issue attempt to tackle the combinatorial complexity, but this presents a major challenge even for relatively small-scale networks (Teodorović et al. 1994, Jaillet et al. 1996). In general, current approaches to route planning and network expansion design often lack a preliminary model that can narrow down the pool of potential routes before applying optimization algorithms. Such a model could provide valuable insights to identify the most promising candidates for network expansion, reducing the search space and making the optimization problem computationally more tractable. By integrating this type of pre-processing tool, network planning algorithms would be able to operate more efficiently, significantly cutting down on computational time, paving the way for their widespread adoption within airline decision-making process.

In short, while airlines have extensively leveraged advanced optimization tools for tactical and operational decisions, the strategic aspect of network development, especially route planning, remains an area where analytics-based solutions are still sparse despite their potential to introduce significant benefits and provide helpful insights.

This paper addresses these gaps by proposing a classification model to identify promising long-haul connections to operate based on a set of publicly available variables. The model applied to new routes opened between 2014 and 2019 on a global scale provides a

comprehensive empirical evaluation that sheds light on the determinants of route opening toward explaining airline decisions to develop their network. Moreover, we elaborate on the potential use of this approach in decision-making support tools of the airline, either in conjunction with optimization models or for the initial screening of potential candidate destinations to be better investigated.

5.3 Data and methods

This study applies a classification model to empirically investigate how airlines decide on entering new long-haul markets. The model leverages a comprehensive dataset of no stop long-haul routes opened in the period 2014-2019 on a global scale. The characteristics of these routes are compared against those of routes that could have been potentially opened. We feature engineer different variables to capture essential drivers influencing the airlines' network development decisions, including market potential, competition, and network dynamics. To investigate the effects of such variables, we apply a classification model with carrier-specific formulations to investigate heterogeneous airlines' behaviors in selecting new markets to enter.

In the following, we better introduce the procedure to build the sample, the variables taken into account and the methodology.

5.3.1 Data sources

The data required for the proposed approach is twofold: (i) data about airline entry, that is new long-haul connections started by selected airlines in the period under investigation, and (ii) potential candidates routes to be operated.

To gather data about route entrance, we leverage OAG Schedule Analyser database. We consider the period between 2014 and 2019 and collect route-level data per each different carrier at the global scale. Each route is uniquely identified by the connected airports (i.e., airport pair) and the operating airline. This led to a dataset composed of more than 600.000 route-carrier-year observations. To focus on long haul markets, we

Table 11: Number of observations by year and type for the top 20 carriers by number of long-haul routes operated in 2019.

Year	Canceled	Established	New
2014	108	2,750	306
2015	135	2,921	239
2016	119	3,052	321
2017	149	3,224	281
2018	104	3,401	337
2019	173	3,565	296
Total	788	18,913	1,780

apply a fixed cutoff on route great circle distance of 3000 km. We further restrict our sample only to observations with at least 25 flights per year, assumed as a reasonable threshold to identify active connections (corresponding to a seasonal weekly flight). We focus on full-service carriers (FSCs) due to their higher presence on long-haul routes and their extensive use of hub & spoke connectivity. At this point, we mark each observation (route-carrier-year) as: established, if the route is operated both in year t and in the previous one by the single airline; new, routes starting to be operated by the airline in year t; and canceled, routes that in year t cease to be operated by the airline but they were in the previous one. Table 11 summarizes the number of observations by year and type for the top 20 carriers by number of long-haul routes operated in 2019. As expected, we observe that the majority of long-haul routes are marked as established. This subgroup accounts for about 2900 route-carrier pairs in 2014 and increases to 3643 in 2019. The number of new long-haul routes ranges between 228 and 297 per year in the period under investigation. Lastly, the routes canceled pass from 79 in 2014 to 181 in 2019. For the scope of our analysis, we focus on the route-carrier observations marked as new.[40] These are routes in which the specific airline enter in a given year.

Routes that were subject to an entrance need to be compared against the set of routes that potentially could have been opened by the airline. We build a tailored procedure to generate such observations. Given the focus on FSCs, typically operating long-haul

[40]Due to heterogeneous market entry determinants, we exclude from our analysis flights operated exercising 5^{th} Freedom rights, thus flights between two countries that are not the carrier's home country.

flights from their hub facilities, we first identify the airports that can be considered as hub for each airline. For this classification, we first refer to the historical number of connecting passengers and classify as hub the airports in which the airlines manage the majority of its connecting passengers. Most of the airlines considered have only one airport corresponding to the criteria, which has been identified as a hub (e.g., DXB for Emirates and LHR for British Airways). A few airlines have instead a multi-hub structure (e.g., Lufthansa with FRA and MUC). This ultimately demonstrate the capability of the threshold to identify the network structure of each carrier (single vs. multi-hub structure). Second, we identify potential candidate destination airports considering all the airports worldwide with scheduled long-haul flights, assumed as a proxy of infrastructure technical capability to accommodate such type of traffic (e.g., runway length and terminal/stand infrastructure). To reasonably restrict the set of potential destinations, we apply a filtering criteria excluding airports with number of historical long-haul movements below the 20^{th} percentile for each geographical macro-region. This sample restriction allows to keep only airports with a consistent long-haul traffic and that can be considered as a suitable potential destination. At this point, we identify the set of potential routes for each airline as the routes connecting airline hub(s) to the candidate destination airports, and *vice versa*. To ensure the focus on long-haul connections, we apply a distance cut-off by selecting only potential routes with distance higher than 3000 km. Moreover, to ensure technical feasibility of such flights we restrict the analysis to routes below 12,545 km (99^{th} percentile of distance of existing long-haul routes). Lastly, we exclude from the set of potential connections the observations (route-carrier-year) already observed in the historical dataset (as established, new, or canceled).

The final dataset is composed by the new routes, namely routes that in a specific year have been opened by the airline and the set of potential routes candidates that the airline could have operated but it did not. To preserve results tractability while keeping the global scope of the analysis, for the empirical analysis we focus on observations of the top 20 carriers by number of long-haul routes operated in 2019. Table 12 reports

Table 12: Sample by carrier and type of observation: new or potential.

Carrier	IATA Code	Route-carrier-year observations		
		New	Potential	Total
Delta Airlines	DL	190	5,525	5,715
Air Canada	AC	183	9,544	9,727
United Airlines	UA	128	8,469	8,597
Alaska Airlines	AS	122	458	580
China Southern Airlines	CZ	120	2,818	2,938
China Eastern Airlines	MU	116	6,146	6,262
American Airlines	AA	115	5,350	5,465
Qatar Airways	QR	99	2,466	2,565
Air China	CA	98	3,098	3,196
Turkish Airlines	TK	83	2,501	2,584
Saudi Arabian Airlines	SV	81	6,637	6,718
Lufthansa	LH	68	5,777	5,845
Aeroflot	SU	60	3,094	3,154
LATAM Airlines	LA	59	7,948	8,007
Ethiopian Airlines	ET	58	3,215	3,273
British Airways	BA	52	2,666	2,718
Emirates	EK	48	2,285	2,333
Air France	AF	45	2,536	2,581
KLM Royal Dutch Airlines	KL	37	2,812	2,849
Korean Air	KE	18	2,874	2,892
Total		1,780	86,219	87,999

the number of observations (route-carrier-year) by carrier and type. The final sample consists of 87,999 observations: 1,780 of new routes and 86,219 of potential routes.

5.3.2 Variables definition

In this Section, we introduce and discuss the formulation of explanatory variables designed to capture the essential drivers influencing airline network development. We initially engineered and tested a broad set of features in various combinations within the empirical model. Through a comprehensive evaluation, we assessed the model's performance using different subsets of variables to determine their impact. After carefully considering the results, we retained only the most relevant variables, namely those that contributed significantly to the model's predictive power and did not introduce multicollinearity issues. While the final set of explanatory variables used in the empirical models is specified in Section **??**, this Section presents the entire set of explanatory variables considered.

To systematically capture the most important drivers influencing the airlines' network development decisions, based on prior literature, we consider several groups of factors: route distance, capacity, socio-economic, market competition, market potential and feeding potential. Table 13 summarizes these categories along with the definitions of the variables analyzed within each group[41]. Figure 21 illustrates the correlation between the different variables.

The first dimension considered is distance. Route distance represents the most straightforward measure of travel impedance considered within basic gravity model frameworks, which assume that economic connectivity between regions and travel demand decrease as distance increases. This weaker coupling is also expected to affect the attractiveness of serving a market and, thus, opening a new route. Accordingly, in our analysis, we consider the great-circle distance between origin and destination airports, calculated based on their geographical coordinates.

A second key dimension in assessing the feasibility of opening a new route is fleet availability. In their planning process, airlines determines routes to operate based on the availability of aircraft with different capacity and range characteristics. To account for this factor, our model incorporates variations in fleet capacity by analyzing the changes (at the carrier level) in available seat kilometer (ASK) compared to the previous year. To suitably model the heterogeneous (among airlines) use of aircraft on different flight segments, we compute ASK variations considering three distance categories: 3,000–4,000 km, 4,000–6,000 km, and over 6,000 km. Specifically, for each airline and flight distance, we measure the change in ASK for aircraft typically used by that airline on routes of the given length.[42] Such an approach provides valuable insights into airline capacity dynamics within specific flight segments. An increase in ASK suggests fleet expansion and potential new route openings, while a decrease may indicate capacity reductions or

[41] The table includes a broader set of variables than those shown in the final results. This reflects the full range of variables explored during the model specification stage, although only those demonstrating meaningful explanatory power were ultimately retained in the final model.

[42] To this aim, within each distance category, we compute the ASK variation by considering aircraft models with seating capacities between the 10^{th} and 90^{th} used by the airline for the specific flight distance.

Category	Variable	Description and formulation
Distance	Route distance	Route great-circle distance
Capacity	Δ ASK	Year-over-year change in available seat kilometers (ASK), as a proxy for fleet capacity, for a given airline and flight distance category. It is computed considering the aircraft models (10^{th}-90^{th} percentiles in terms of seating capacity) typically used by the airline for routes of similar length
Market Potential	Population	Total population within the destination airport catchment area assumed as a static 100 km radius area
	GDP per capita	Average GDP per capita for inhabitants living within the destination airport catchment area assumed as a static 100 km radius area
	CAGR population	Compound annual growth rate over a 5 year period of population within destination airport catchment area
	CAGR GDP per capita	Compound annual growth rate over a 5 year period of GDP per capita within destination airport catchment area
	No-stop demand	Number of passengers traveling between the departure city to the destination one using no-stop services
	Connecting demand	Number of passengers traveling between the departure city to the destination one using one-stop itineraries
Market Competition	Nr.airlines	Number of airlines operating the route
	Nr.partner	Number of airlines operating the route within the same alliance of the carrier considered
	Nr.competitors	Number of airlines operating the route not belonging to the same alliance of the carrier considered
	$FTSM$	First-time served market: dummy variable equal to 1 if the market is not served by any no-stop service, 0 otherwise
	HHI	Herfindahl-Hirschman Index based on the no-stop seating capacity. Let $\mathcal{A}$ be the set of airlines offering no-stop services on a route and ϑ_a the number of allocated seats by airline a, we define the HHI as $\sum_{a\in\mathcal{A}} MS_a^2$ where $MS_a = \frac{\vartheta_a}{\sum_{a\in\mathcal{A}}\vartheta_a}$ is the market-share (in terms of offered seats) of airline a
Feeding Potential	Equivalent number of spokes	For a given long-haul route from hub h to destination d, let S_h be the set of spokes s ($s \in S_h$) connected to the hub with at least one weekly flight. We define the equivalent number of spokes for the route from h to d as $SE_{hd} = \sum_{s\in\mathcal{S}_h} \frac{1}{Rf_{shd}}$, where Rf_{shd} is the routing factor, defined as $Rf_{shd} = \frac{\varphi_{sh}+\varphi_{hd}}{\varphi_{sd}}$ where φ represents the great-circle distance.
	Corrected frequency	We define the corrected frequency by considering the equivalent number of spokes and adjusting the weight associated with each spoke s using the feeding frequency f_{sh} (the frequency of connections from the spoke to the hub). Accordingly. the corrected frequency is defined as: $CF_{hd} = \sum_{s\in\mathcal{S}_h} \frac{f_{sh}}{Rf_{shd}}$, where Rf_{shd} is the routing factor.
	Corrected potential demand	Let be $\hat{d}_{sd}$ the number of passengers currently traveling from spoke s to final destination d using one-stop itineraries, we define the corrected potential demand as: $CD_{hd} = \sum_{s\in\mathcal{S}_h} \frac{\hat{d}_{sd}}{Rf_{shd}}$, where Rf_{shd} is the routing factor.

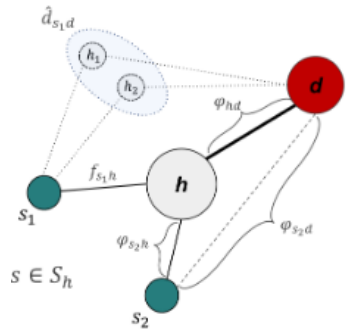


Table 13: Variable definition and formulation.

more stringent fleet constraints.

A third category of variables affecting route opening are socio-economic factors, which ultimately shape the potential market size. In our analysis, we consider the population and GDP per capita of the destination area.[43] To compute these values, we assume a static catchment area with a 100 km radius around each destination airport and derive the population and GDP values based on a high-resolution global spatial dataset (see **??**, for details). These metrics offer insights into the economic activity and consumer demand within the airport's vicinity, serving as key indicators of market size and purchasing power—both critical factors in determining route attractiveness for airlines. In addition to the current socio-economic figures, we also account for long-term trends by incorporating the compound annual growth rate (CAGR) of population and GDP per capita over a five-year period, capturing the prospective regional demographic and economic growth.

Besides socio-economic factors contributing to determining market size, market potential can also be assessed by examining current air demand. Accordingly, we consider both no-stop and connecting demand in the specific market. These two measures provide distinct yet complementary insights into route attractiveness. No-stop demand refers to the volume of passengers using no-stop services on a particular market and results in a twofold interpretation of route attractiveness. On one hand, high volumes of no-stop traffic indicate a strong and well-established market, suggesting that the route serves a thriving market. On the other hand, it may also imply that the market is already well-served by existing no-stop services, potentially leaving less room for new entrants. By contrast, connecting demand represents the number of passengers traveling on the analyzed market using one-stop itineraries. The higher this demand, the greater the number of passengers who could benefit from a more convenient no-stop service. A strong presence of connecting demand indeed is a signal of market potential— if a substantial number of passengers are already traveling between two cities via one-stop itineraries, introducing

[43] Given that the carrier home country and hub airport can be reasonably assumed fixed, we consider only socio-economic variables regarding the route destination.

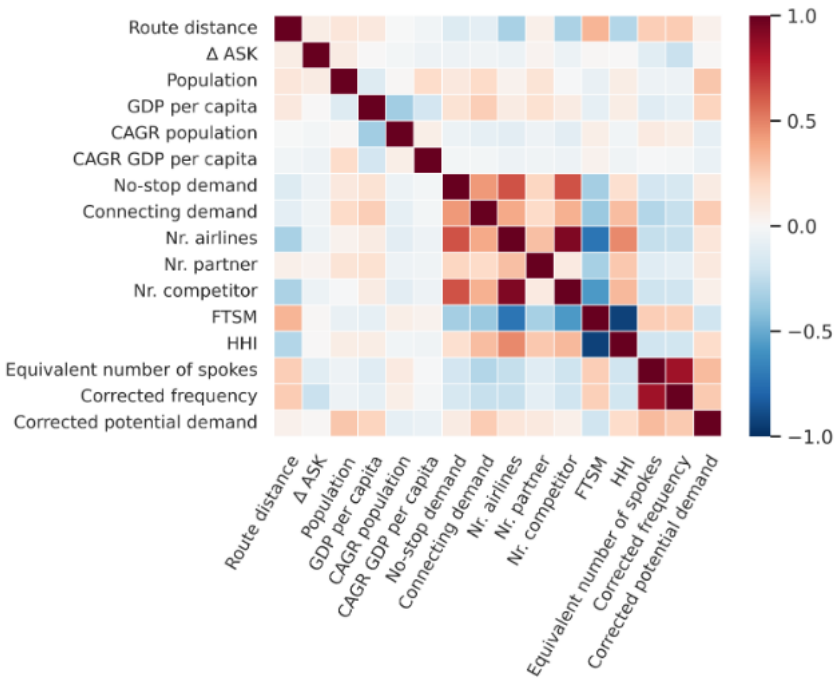


Figure 21: Heatmap of the correlation between variables (train dataset balanced through oversampling).

a more convenient no-stop flight could capture this latent demand, positioning the new route as an attractive alternative to existing connecting itineraries.

The attractiveness of entering a new route also depends on the competitive landscape. Accordingly, we consider the number of competing airlines already operating the route as a variable that may influence its attractiveness. At the same time, especially on long-haul routes, airlines often offer code-share opportunities. To account for this, we incorporate alliance dynamics by considering the number of partners already serving the route. Airlines within the same alliance are classified as partners, while those outside the alliance are considered competitors. This distinction provides insights into the competitive dynamics within the market, as well as the carrier's positioning relative to both its alliance partners and competitors when evaluating the possibility of serving a new route. Besides the number of airlines, we also consider the overall competition in the market using the Herfindahl-Hirschman Index (HHI) computed based on allocated seating capacity. Ultimately, we model the entrance to a fist-time served market by using a dummy variable equal to 1 if the market is not served by no-stop services by any airline.

Lastly, route entry decisions also depend on airline network considerations, including how the potential route fits within the entire airline network. To cover these aspects, we develop three metrics aimed at measuring the strength of the airline's feeder network for serving a specific long-haul route. Since these measures strictly relate to airline hubbing activities and the possibility to connect different flights we define these measures only for routes to and from the airline hub airport. The first measure is the equivalent number of spokes, which is based on the count of the airports (other than the long-haul destination considered) with which the hub airport is connected. These airports, which we define as spokes, constitute the potential feeding base for the long-haul flight considered. For a given long-haul route from hub h to destination d, let's define S_h the set of spokes s ($s \in S_h$) connected to the hub with at least one weekly flight.[44] The equivalent number of spokes is defined by the simple count of the spoke airports, each weighted by the routing factor that passenger would experience if from the spoke would travel to the final destination through the hub. This variable jointly consider the hub airport network scope (cardinality of S_h) as well as its potentiality to serve a given long-haul route due to its geographical positioning (routing factor effect). A major shortcoming of this measure is that it assigns the same weight to each spoke, regardless of the supply (i.e., number of flights) connecting the spoke to the hub, which ultimately underpins the potential connectivity delivered by the hub to the final destination. To address this, we refine the measure by weighting each spoke proportionally to the frequency of flights to the hub(f_{sh}). This metrics, labeled as corrected frequency, prioritizes actual connectivity over theoretical one, ultimately better reflecting potential connectivity offered by the hub for flights to the final destination. Despite these advantages, the corrected frequency still remains a supply-based measures and does not account for actual market size— that is, the number of passenger who could be attracted in case of the opening of the new long-haul route, given the hub feeding structure. To overcome this limitation, we introduce a

[44]While the concept of spoke has been traditional intended from a local perspective, new Middle-East and Turkish hubs heavily rely on long-haul to long-haul connectivity. Accordingly, we do not considered an explicit distance threshold to define the set of spoke airports.

third measure: the corrected demand. This measure, in addition to the routing factor[45], also considers the number of passengers currently traveling from spoke s to final destination d via one-stop itineraries ($\hat{d}_{sd}$). These passengers represents the potential demand that could be attracted in case a new route from hub h to final destination d would enter the market. The new itinerary would be much more attractive the higher the market size and the lower the routing factor. Ultimately, the corrected demand captures the size of the feeder base for a given long-haul route given the airline network. In summary, these three metrics form a hierarchy of "feeding potential": the equivalent number of spokes measures theoretical network scope, corrected frequency captures realized connectivity, and corrected demand reflects the actual market potential enabled by the hub's network structure. Collectively, they provide a nuanced representation of how an airline's network can support new long-haul route openings.

This set of variables captures the key factors airlines consider when evaluating the launch of a new route. One advantage is that these dimensions can be measured not only for existing routes but also for potential ones. More importantly, all the variables presented are publicly available, making them practical tools for investigating network planning and route expansion decisions.[46] As discussed, from the broad set of variables presented, we retained for the econometric model only the most relevant– those that significantly enhance the model's predictive power without introducing multicollinearity issues. The final set of explanatory variables used in the empirical model is detailed in Section 5.4.

5.3.3 Methodology

To investigate the effects of various factors on route entry decisions, we employ a logistic regression model. The model systematically compares the characteristics of newly

[45]The routing factor is defined as the ratio between the total flown distance of an itinerary (origin–hub; hub–destination) and the direct distance between the origin and destination.

[46]To mitigate potential reverse causality issues, we used one-year lagged values for all time-dependent variables. Beyond the econometric benefits, this approach better reflects the information available to airlines during their planning process.

launched routes with those of potential but unserved routes, allowing to estimate the probability of a route being opened based on a set of predictors. The estimated coefficients quantify the relationship between each factor and the likelihood of a route entry, where positive coefficients indicate a higher probability of a route being opened, while negative coefficients suggest a lower probability. A key advantage of the use of logistic regression in this context lies in its ability to provide interpretable coefficients, offering clear insights into the contribution of each predictor in shaping route entry decisions. This straightforward interpretability can facilitate the understanding of the complex interplay of factors -such as demand potential, geographic considerations, network effects, and competitive dynamics- that undermine airline network development strategies. An alternative approach are advanced machine learning classification models (e.g., random forest). We tested such models with limited increase in predictive power in the face of significantly lower interpretability. This is likely due to the limited sample size –stemming from the sporadic nature of long-haul route entries –and the poor availability of big data to be considered as explanatory variables. Therefore, we present only the results based on the logistic model.

Formally, the probability that a route i between origin o and destination d for carrier c is opened is modeled as:

$$P(Y_i = 1 \mid X_i) = \frac{\exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k X_{ik}\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k X_{ik}\right)}, \tag{10}$$

where:

- Y_i is a binary variable equal to 1 if the route is opened and 0 otherwise,
- X_{ik} denotes the k-th explanatory variable for route i, including demand characteristics, distance, GDP of origin and destination cities, market competition indicators, lagged route presence, hub connectivity metrics (e.g., equivalent number of spokes, corrected frequency, corrected demand), and other network-related variables,

- β_0 is the intercept, and β_k are the coefficients associated with each predictor, capturing the marginal effect on the log-odds of route entry.

This specification allows for a quantitative assessment of how each factor influences the likelihood of new route openings, providing a transparent framework to interpret both economic and network drivers of long-haul route development.

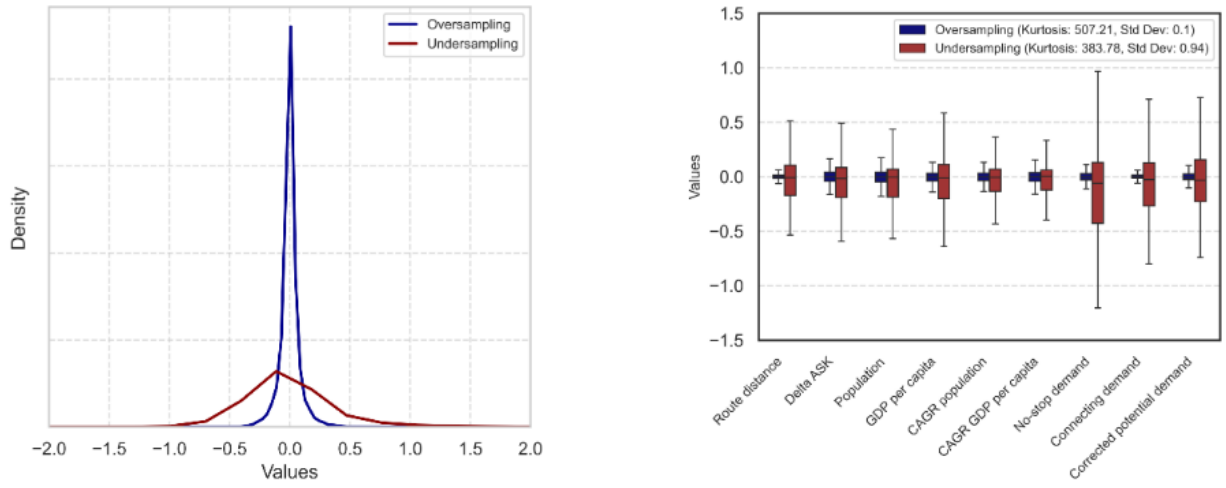
We divide the data sample described in Section 5.3 into two parts: observations from 2014 to 2017 are used for model calibration (train), while observations from 2018 to 2019 are used as a testing bed to evaluate model performance. Since new long-haul route entries are relatively rare events, the dataset is highly imbalanced, with the number of potential routes far exceeding (48x) the number of actual route openings. This imbalance has proven to negatively impact logistic regression, which tends to favor the majority class (Lai et al. 2021, Megahed et al. 2021). To cope with this issue, we train the model on a balanced dataset, while the testing was conducted on the original imbalanced dataset to assess its generalizability and performance in real-world conditions. The balanced dataset for training is constructed by applying two commonly resampling techniques: oversampling and undersampling. Oversampling involves artificially increasing the number of instances in the minority class by duplicating existing observations. This technique ensures that the minority class is adequately represented, providing the model with more opportunities to learn from the underrepresented class (Wang et al. 2018a, Ustyannie and Suprapto 2020). On the other hand, undersampling reduces the number of instances in the majority class by randomly discarding observations, thus obtaining a more balanced dataset. While this approach mitigates bias toward the majority class, it risks losing potentially useful data and may lead to a less robust model (Cartus et al. 2020, Kubus 2020). Both methods modify the original data distribution by either increasing the number of samples from the minority class or reducing those from the majority class. These adjustments aim to mitigate class imbalance, enhancing the model's ability to classify (Oommen et al. 2011, Yap et al. 2014). While both approaches are considered valid, there is no strong consensus on which approach yields superior performance since their

effectiveness is typically context-dependent. Accordingly, the choice between oversampling and undersampling varies based on the dataset characteristics and study objectives (Chawla et al. 2002, He and Garcia 2009).

To determine which balancing technique yields the most stable and reliable results for our case, we repeat the logistic regression 100 times, randomly selecting the sample using both techniques. The optimal approach is then determined by evaluating two aspects: (i) coefficient stability, and (ii) forecasting performance. The former consists of evaluating how consistently each variable's coefficient remains close to its mean across different random extractions, indicating low variability over estimations. Figure 22a presents the distribution of coefficients around their mean under the two sampling methods, while Figure 22b compares coefficient distributions across variables for both techniques. Only statistically significant coefficient estimates were included in the analysis. To ensure comparability within the same graph, coefficients were transformed using a tailored formulation that quantifies their relative deviation from the mean across multiple iterations.[47] This transformation ensures that distributions are comparable across different variables and sampling strategies. The shape of the curve in Figure 22a reveals whether the coefficients cluster around their mean or exhibit high variability. Overall, coefficients obtained through oversampling exhibit lower variability. This evidence is confirmed also when looking to single explanatory variables (22b). Accordingly, we conclude that, in our case, oversampling reduces sensitivity to variations across iterations, thereby enhancing model robustness[48].

[47] We compute coefficient deviation for iteration i, x_i, as $\frac{\beta_i - \bar{\beta}}{\bar{\beta}}$, where β_i is the coefficient estimate in iteration i and $\bar{\beta}$ is the mean of the coefficient over all the iterations.

[48] The selection of the balancing strategy is based on a trade-off between predictive performance and parameter stability. Although sensitivity is a relevant metric in the presence of rare events, the primary objective of the estimation is to obtain well-identified and robust coefficient estimates suitable for inference and out-of-sample forecasting. Undersampling mechanically reduces the effective sample size of the majority class, thereby increasing estimation variance and inducing greater sensitivity of coefficient estimates to random sample composition. Oversampling, by preserving the full set of majority-class observations, yields coefficient estimates that are substantially more stable across repeated estimations, while delivering comparable overall predictive performance. Given that coefficient instability can translate into large fluctuations in predicted probabilities in rare-event settings, parameter robustness is prioritized over marginal gains in sensitivity, motivating the choice of oversampling.



(a) All the variables

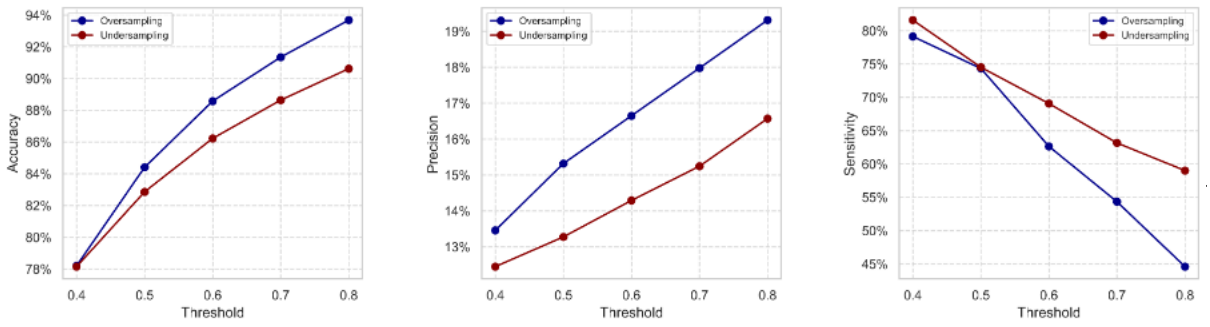
(b) By variable

Figure 22: Coefficient deviation (stability) from their mean over 100 iterations: oversampling vs. undersampling.

A second aspect to consider when investigating the preferred sampling technique is model forecasting performance, which is the ability of the model to classify instances in the testing dataset correctly. We evaluate model performances by focusing on three KPIs: accuracy, sensitivity, and precision. Accuracy is the most intuitive classification metric, representing the overall proportion of correctly classified instances —both positive and negative— out of all predictions. Precision, on the other hand, measures the proportion of positive predictions (i.e., routes identified by the model as new) that are actually correct (i.e., have been actually been opened by the airline).[49] Precision provides insights into the accuracy of the model's predictions when it forecasts a promising new route. Lastly, sensitivity (also known as recall) measures the proportion of actual positive instances (i.e., new routes) that are correctly identified by the model. In our analysis, sensitivity is particularly important as it reflects how effectively the model detects new routes that an airline has opened.[50]

[49] Precision is defined as the proportion of true positive predictions (TP —i.e., routes predicted by the model as promising that have actually been opened) out of all positive predictions made by the model, including both correct (TP) and incorrect (false positive, FP) predictions. Formally, it is computed as $\frac{TP}{TP+FP}$.

[50] Sensitivity is defined as the proportion of true positive predictions (i.e., routes classified as promising and opened by the airline) out of the total number of routes opened by the airline (TP+FN). Formally, sensitivity is computed as: $\frac{TP}{TP+FN}$.



(a) Accuracy **(b)** Precision **(c)** Sensitivity

Figure 23: Forecasting performance KPIs: oversampling vs. undersampling.

Figure 23 reports the three KPIs for the different sampling techniques and various predicted probability thresholds used to classify routes as promising or not promising. In terms of accuracy, both resampling techniques perform well, ranging between 0.78 and 0.94 depending on the chosen threshold. However, oversampling consistently outperforms undersampling. Similarly, oversampling ensures a precision that is consistently consistently 1% to 2% higher than that of undersampling across all the threshold levels. Sensitivity, on the other hand, shows the most heterogeneous pattern among the KPIs. At lower thresholds (0.4 and 0.5), oversampling and undersampling are characterized by similar performance. However, as the threshold becomes more stringent, undersampling demonstrates superior performance.

In summary, our evaluation of balancing techniques based on coefficient stability and forecasting performance suggests that, in our context, oversampling is the preferred approach. It consistently yields more stable coefficient estimates across iterations, indicating lower variability and greater model robustness. Additionally, while both techniques demonstrate good forecasting performance, oversampling achieves higher accuracy and precisions. On the other hand, results for sensitivity are more favorable to undersampling. Overall, these findings indicate that oversampling is the more reliable and effective balancing technique for our analysis, as it enhances model consistency while improving the precision of route predictions. Accordingly, in the following Section, we discuss the empirical results obtained using oversampling as balancing technique.

5.4 Results

5.4.1 Econometric results

In this Section, we discuss the empirical results of the classification model applied to the comprehensive dataset of new route openings and potential routes detailed above. We developed carrier-specific formulations, offering a detailed examination of the coefficients influencing each carrier's network expansion decisions. Additionally, we compare these results with those obtained from a pooled logit model, which does not account for carrier-specific segmentation.

For each carrier in our sample, we perform 100 oversampling bootstrap iterations following the oversampling procedure detailed in Section 5.3. In each iteration, the logit model is fitted using the instances within the training dataset belonging to the selected airline and balanced through oversampling, providing a carrier-specific framework for network expansion modeling. The model captures the key drivers' effect on the airline's network expansion strategy and reflects the importance of each variable within the airline's decision-making process. To obtain single coefficient estimates for each carrier, we aggregate the results across the different iterations. Specifically, coefficient magnitude is computed as the average across all iterations, while the statistical significance of the estimated coefficients is assessed using the z-score, calculated as the ratio of the mean coefficient to its standard error.

Figure 24 reports coefficients of the carrier-specific models as well as those of the pooled model. To facilitate a more intuitive interpretation and comparison, we standardize the variables and display the results using a heatmap.[51] The color intensity indicates the magnitude of the standardized coefficients, with darker shades representing higher values and lighter shades indicating smaller values. The color itself —green or red—

[51]The standardization procedure follows the methods outlined in Adler and Hashai (2005), **?**), **?**, which involve standardizing the coefficients based on the variance in the predictors. This approach ensures that the coefficients are expressed on a comparable scale, facilitating a more meaningful analysis of their impact.

denotes the sign of the coefficient, where green corresponds to positive coefficients and red to negative ones. This approach allows for a more straightforward visual assessment of the relative importance and influence of each variable across the different carriers.

The heterogeneous results, both in terms of coefficient magnitude and direction, highlight the complex interplay of factors shaping the network expansion strategies of different carriers. In contrast, the soft colors of the pooled model indicate its limited explanatory power, suggesting that a generalized approach fails to adequately capture the complexities of individual carriers' expansion dynamics. Looking at single variables, the most influential factor across all carriers is connecting demand, which consistently exhibits a strong and positive effect on the likelihood of entering a route. This result underscores the importance of serving a market with a no-stop connection, potentially attracting passengers who are currently traveling on connecting itineraries. The no-stop demand and the strength of potential demand from the spokes are also critical drivers, though their impact varies across carriers. Considering no-stop demand, some carriers, such as Delta and Air Canada, tend to favor the entry in routes already served by no-stop demand. In contrast, other airlines, including Emirates and Turkish Airlines, exhibit a tendency to avoid opening new routes in markets already served by no-stop services. The potential feeder base and the positioning of the hub airport to effectively serve a given destination considering the existing network are determining factors for many carriers. This is the case of Gulf carriers such as Emirates and Qatar Airways, as well as Turkish Airlines, Korean Air, and LATAM Airlines, which demonstrate a strong reliance on connecting traffic to support new route openings. Consistent with prior literature, distance exhibits a negative effect on route opening across all airlines. Socio-economic variables do not display a homogeneous role in determining route opening. On the one side, some carriers such as Ethiopian Airlines and British Airways place a great emphasis on current market conditions, prioritizing socio-economic factors when making expansion decisions. On the other side, airlines such as KLM and Air France focus on future growth potential, aligning their strategies with long-term oriented demographic and economic projections.

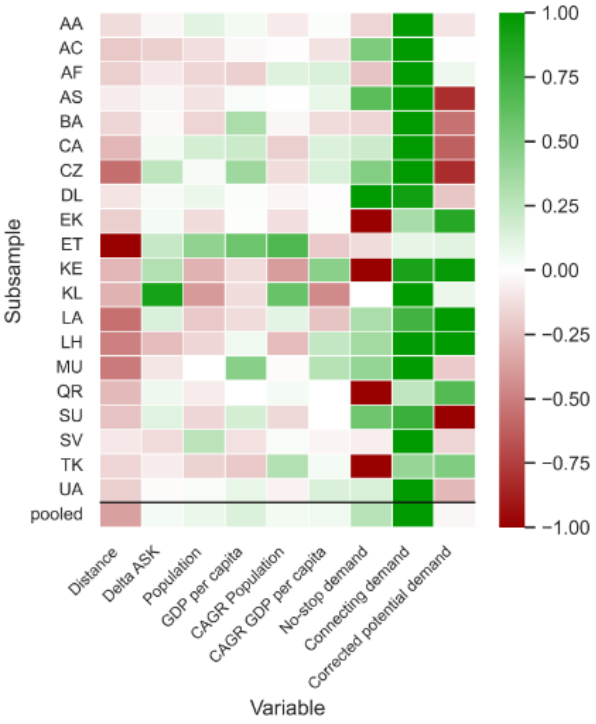


Figure 24: Model results: heatmap with normalized coefficients.

Carrier	Accuracy	Precision	Sensitivity	Screening ratio
DL	97.1%	52.5%	86.7%	5.2%
AC	83.4%	10.8%	84.9%	18.2%
UA	90.6%	12.7%	82.6%	10.4%
AS	93.3%	85.3%	93.5%	32.7%
CZ	82.7%	14.1%	62.5%	18.4%
MU	78.4%	7.1%	75.0%	22.7%
AA	94.6%	29.6%	78.7%	6.9%
QR	80.1%	8.9%	45.2%	19.5%
CA	82.3%	10.7%	70.0%	18.9%
TK	84.1%	10.1%	58.3%	16.4%
SV	80.9%	5.0%	88.0%	19.9%
LH	87.7%	5.0%	54.5%	12.4%
SU	96.7%	14.7%	50.0%	3.3%
LA	88.4%	8.5%	84.8%	12.4%
ET	70.9%	4.6%	83.3%	30.3%
BA	74.5%	7.3%	90.0%	27.3%
EK	81.2%	5.4%	72.7%	19.5%
AF	86.6%	6.5%	35.0%	12.6%
KL	68.3%	3.3%	90.9%	32.6%
KE	86.3%	4.3%	100.0%	14.3%
Pooled	84.5%	9.2%	68.0%	16.3%

Table 14: Model KPIs considering a threshold of 0.5 to define promising routes.

Lastly, some airlines prove to be strongly dependent on fleet and capacity consideration (ASK variable) when determining their route network expansion. This is particularly evident for KLM, which tends to plan network adjustments based on the growth potential (in terms of capacity) available for each type of destination. Overall, this fragmented landscape highlights the heterogeneity in airlines' approaches, further demonstrating the benefits of carrier-specific modeling over a generalized framework.

We evaluate model performance using a customized out-of-sample validation approach applied to the unbalanced testing dataset, considering different KPIs such as accuracy, precision, and sensitivity, as discussed in Section 5.3. To complement these metrics, we introduce the screening ratio, a tailored KPI designed to assess the model's ability to narrow down the set of potential routes and identify the most promising ones. It is defined as the proportion of routes identified by the model as promising relative to the total number of candidates. This metric serves as a proxy for the model's effectiveness in selecting a subset of candidate routes to be considered for further investigation using network planning optimization algorithms. Ultimately, it quantifies the potential reduction

in complexity of the network planning algorithm when fed by the results of the proposed model.

Table 14 reports the results of the out-of-sample assessment of model performance presenting the metrics for the carrier-specific models as well as for the pooled logit model that does not account for carrier-specific variations. As expected, airlines exhibiting lower model accuracy are also those with the fewest new routes introduced during the training period (i.e., from 2014 to 2017). This limited expansion reduces the amount of the historical data used for model calibration, thereby constraining the model's ability to achieve high performance, particularly in comparison to carriers with more extensive route development. The pooled model, which aggregates all airlines into a single framework, yields a significantly lower accuracy (approximately 29%), further demonstrating the importance of carrier-specific dynamics. This underscores the limitations of a one-size-fits-all approach in capturing the nuances of individual airline strategies.

To further explore model performance, the models are tested at varying probability thresholds (ranging from 0.4 to 0.8) to define promising routes rather than the canonical 0.5 threshold (results are shown in Table 15). By varying the threshold, we simulate different risk tolerance levels that airlines might have when making decisions about new route openings. Lower thresholds (e.g., 0.4) may lead to more aggressive predictions, resulting in a higher number of promising routes identified, but possibly at the cost of precision. Conversely, higher thresholds (e.g., 0.8) might lead to fewer predictions, potentially increasing precision but sacrificing sensitivity. This trade-off is clearly depicted in Figure 25 which presents a boxplot of three KPIs—sensitivity, precision, and screening ratio—across different thresholds. As the threshold increases, precision tends to improve (fewer false positives), but this may come at the cost of sensitivity (a decrease in true positives). Conversely, lowering the threshold may improve accuracy but lead to a reduction in precision, as more false positives are generated. By analyzing these results, we gain valuable insights into the model's effectiveness in identifying new routes under different decision-making scenarios. This analysis not only highlights the model's sensitivity to

threshold variations but also helps determine the optimal threshold for identifying new routes while balancing the competing priorities of sensitivity, precision, and screening ratio.

The observed combination of high sensitivity and comparatively low precision, even at higher probability thresholds (0.8), reflects structural features of the route entry process rather than a lack of model performance. New long-haul route openings are rare events and depend not only on observable market characteristics but also on a range of firm-specific, strategic, and institutional factors that are not systematically observable, such as fleet allocation, slot availability, bilateral negotiations, and internal investment priorities. As a result, the model identifies routes that are economically viable or plausible ex ante, but only a subset of these candidates ultimately materialize as actual entries. Ultimately, the limited improvement in precision when increasing the threshold from 0.7 to 0.8 suggests that, conditional on available information, remaining false positives are difficult to distinguish from true positives. Further gains in precision would therefore require additional variables capturing airline-specific strategic constraints rather than adjustments to the classification threshold alone. From a practical perspective, the model should be interpreted as a screening tool that narrows the universe of potential routes rather than a deterministic predictor of route entry. Looking ahead, future research could improve discriminatory power by incorporating richer firm-level information, such as fleet composition and utilization, network-level capacity constraints, alliance membership, or airport-level slot scarcity, as well as institutional variables capturing regulatory and bilateral agreement constraints. Integrating such information would allow future models to better distinguish between economically viable routes and those that are strategically feasible, thereby enhancing precision without compromising sensitivity.

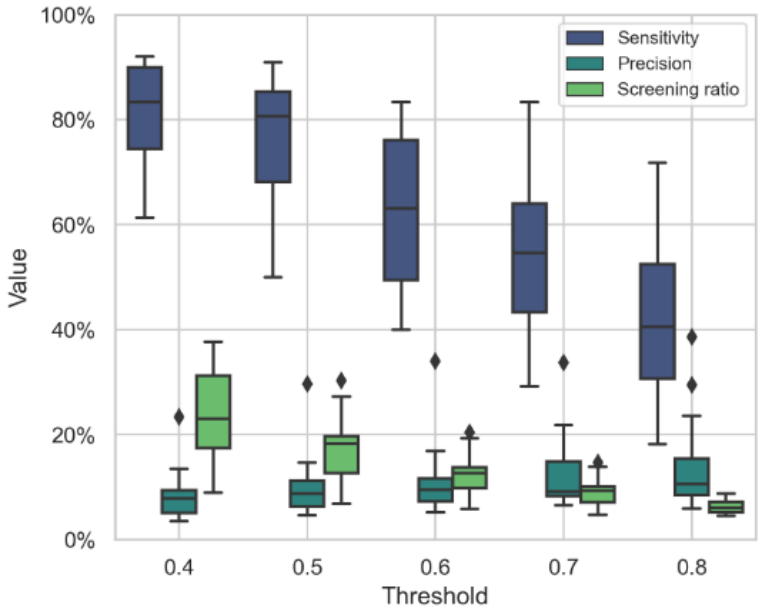


Figure 25: KPIs across different threshold levels.

Threshold	Metric	AA	AC	AF	AS	BA	CA	CZ	DL	EK	ET	KE	KL	LA	LH	MU	QR	SU	SV	TK	UA	Pooled
0.4	True Positive	38	67	9	58	18	25	30	53	8	15	6	10	29	14	38	19	5	23	15	38	514
	True Negative	1.649	2.441	680	134	572	731	731	1.778	523	669	794	494	2.273	1.606	1.345	547	993	1.410	657	2.478	21.191
	False Positive	127	651	146	12	310	283	193	56	226	387	164	433	343	319	677	226	35	794	162	350	6.696
	False Negative	9	6	11	4	2	5	10	7	3	3	0	1	4	8	6	12	5	2	9	8	115
	Accuracy	93%	79%	81%	92%	65%	72%	79%	97%	70%	64%	83%	54%	87%	83%	67%	70%	96%	64%	80%	88%	76%
	Precision	23%	9%	6%	83%	5%	8%	13%	49%	3%	4%	4%	2%	8%	4%	5%	8%	13%	3%	8%	10%	7%
	Sensitivity	81%	92%	45%	94%	90%	83%	75%	88%	73%	83%	100%	91%	88%	64%	86%	61%	50%	92%	63%	83%	82%
	Screening Ratio	9%	23%	18%	34%	36%	30%	23%	6%	31%	37%	18%	47%	14%	17%	35%	30%	4%	37%	21%	14%	25%
0.5	True Positive	37	62	7	58	18	22	26	52	8	15	6	10	28	12	34	14	5	22	14	38	428
	True Negative	1.687	2.580	723	137	655	836	772	1.786	609	750	826	633	2.317	1.692	1.586	628	999	1.804	695	2.564	23.664
	False Positive	89	512	103	9	227	178	152	48	140	306	132	294	299	233	436	145	29	400	124	264	4.223
	False Negative	10	11	13	4	2	8	14	8	3	3	0	1	5	10	10	17	5	3	10	8	201
	Accuracy	95%	83%	86%	94%	75%	82%	83%	97%	81%	71%	86%	69%	89%	88%	78%	80%	97%	82%	84%	91%	84%
	Precision	29%	11%	6%	87%	7%	11%	15%	52%	5%	5%	4%	3%	9%	5%	7%	9%	15%	5%	10%	13%	9%
	Sensitivity	79%	85%	35%	94%	90%	73%	65%	87%	73%	83%	100%	91%	85%	55%	77%	45%	50%	88%	58%	83%	68%
	Screening Ratio	7%	18%	13%	32%	27%	19%	18%	5%	19%	30%	14%	32%	12%	13%	23%	20%	3%	19%	16%	11%	16%
0.6	True Positive	36	53	7	58	15	15	22	51	5	13	5	6	27	10	21	10	4	19	12	38	369
	True Negative	1.706	2.668	754	139	724	892	814	1.790	686	817	850	741	2.361	1.742	1.749	677	1.003	1.964	718	2.628	25.044
	False Positive	70	424	72	7	158	122	110	44	63	239	108	186	255	183	273	96	25	240	101	200	2.843
	False Negative	11	20	13	4	5	15	18	9	6	5	1	5	6	12	23	21	6	6	12	8	260
	Accuracy	96%	86%	90%	95%	82%	87%	87%	97%	91%	77%	89%	80%	90%	90%	86%	85%	97%	89%	87%	93%	89%
	Precision	34%	11%	9%	89%	9%	11%	17%	54%	7%	5%	4%	3%	10%	5%	7%	9%	14%	7%	11%	16%	11%
	Sensitivity	77%	73%	35%	94%	75%	50%	55%	85%	45%	72%	83%	55%	82%	45%	48%	32%	40%	76%	50%	83%	59%
	Screening Ratio	6%	15%	9%	31%	19%	13%	14%	5%	9%	23%	12%	20%	11%	10%	14%	13%	3%	12%	13%	8%	11%
0.7	True Positive	28	35	5	58	11	14	20	50	4	11	4	6	23	10	18	6	4	16	7	37	319
	True Negative	1.720	2.779	771	142	761	934	853	1.796	703	903	870	803	2.394	1.783	1.833	696	1.007	2.067	744	2.678	25.853
	False Positive	56	313	55	4	121	80	71	38	46	153	88	124	222	142	189	77	21	137	75	150	2.034
	False Negative	19	38	15	4	9	16	20	10	7	7	2	5	10	12	26	25	6	9	17	9	310
	Accuracy	96%	89%	92%	96%	86%	91%	91%	97%	93%	85%	91%	86%	91%	92%	90%	87%	97%	93%	89%	94%	92%
	Precision	33%	10%	8%	94%	8%	15%	22%	57%	8%	7%	4%	5%	9%	7%	9%	7%	16%	10%	8%	20%	14%
	Sensitivity	60%	48%	25%	94%	55%	47%	50%	83%	36%	61%	67%	55%	70%	45%	41%	19%	40%	64%	29%	80%	51%
	Screening Ratio	5%	11%	7%	30%	15%	9%	9%	5%	7%	15%	10%	14%	9%	8%	10%	10%	2%	7%	10%	6%	8%
0.8	True Positive	22	30	5	58	8	12	20	50	2	9	4	4	21	7	12	3	2	15	4	33	260
	True Negative	1.740	2.847	787	142	819	975	876	1.798	715	968	882	873	2.432	1.815	1.914	724	1.007	2.107	771	2.710	26.496
	False Positive	36	245	39	4	63	39	48	36	34	88	76	54	184	110	108	49	21	97	48	118	1.391
	False Negative	25	43	15	4	12	18	20	10	9	9	2	7	12	15	32	28	8	10	20	13	369
	Accuracy	97%	91%	94%	96%	92%	95%	93%	98%	94%	91%	92%	93%	93%	94%	93%	90%	97%	95%	92%	95%	94%
	Precision	38%	11%	11%	94%	11%	24%	29%	58%	6%	9%	5%	7%	10%	6%	10%	6%	9%	13%	8%	22%	16%
	Sensitivity	47%	41%	25%	94%	40%	40%	50%	83%	18%	50%	67%	36%	64%	32%	27%	10%	20%	60%	17%	72%	41%
	Screening Ratio	3%	9%	5%	30%	8%	5%	7%	5%	5%	9%	8%	6%	8%	6%	6%	6%	2%	5%	6%	5%	6%

Table 15: Carriers metrics by threshold.

5.4.2 Airline case study

This Section demonstrates how the proposed classification model can be used to identify promising routes through some case studies based on three major airlines: British Airways (BA), Qatar Airways (QR), and Turkish Airlines (TK). To assess the effectiveness of the model, we analyze the network expansion opportunities suggested by the model –defined as routes with high predicted probabilities based on 2019 data. These predictions are then compared to the actual new routes launched by the airlines between 2019 and 2024. Figure 26 illustrates the routes classified as promising by the model, along with long-haul destinations already served by the airline in 2019.

British Airways, the flag carrier of the United Kingdom, operates an extensive global network centered at its main hub in London Heathrow (LHR). In 2019, the airline maintained a strong transatlantic presence, with flights to major North American destinations such as New York (JFK), Boston (BOS), Toronto (YYZ), and Chicago (ORD). Beyond North America, British Airways also had well-established connections to the Middle East, South Asia, and parts of Africa. The classification model suggests that the airline has untapped potential to further develop its network in fast-growing regions (Figure 26a). Southeast Asia, in particular, stands out as a key area for future expansion, driven by a rising middle class, boosting local economy. The model also indicates potential new destinations in Central Asia and the North America. Notably, six of the destinations classified as promising by the model were added to British Airways' network by 2024. These include Islamabad (ISB) and Malè (MLE) in Asia, Paphos (PFO) in Cyprus, Hamilton (BDA) in Bermuda islands, and Cincinnati (CVG) and Pittsburgh (PIT) in the USA.

Similar performances of the model are outlined for Qatar Airways. The rapidly growing Middle East carrier established itself during the last decade as a leading global airline, with a well-developed network linking the Middle East to Europe, Asia, Africa, and the Americas. By 2019, the airline's operations were heavily concentrated around its hub in Doha (DOH), serving as a central gateway connecting long-haul markets. The model

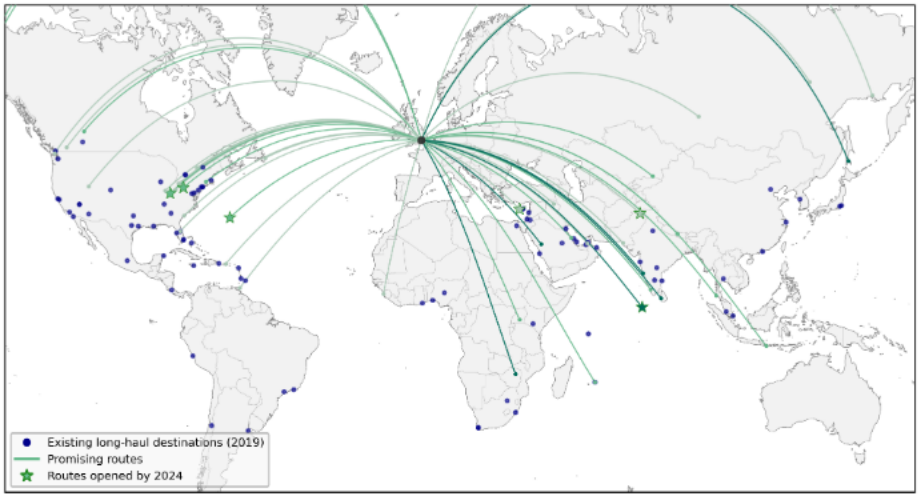
indicates network expansion opportunities for the airline in some destinations in Asia and Africa, as well as strengthening its presence in secondary European and North American destinations (see Figure 26b). These destinations align with the airline's long-haul-to-long-haul hub strategy, which leverages connecting demand across continents. Since 2019, 9 out of the 22 promising routes identified with a probability greater than 60% have been opened as new no-stop destinations by the airline. These include Lisbon (LIS), Hamburg (HAM), and Lyon (LYS) in Europe; Seattle (SEA) and Pittsburgh (PIT) in North America; Accra (ACC) in Africa; Cebu (CEB) and Osaka (KIX) in Asia; and Brisbane (BNE) in Oceania.

The last carrier analyzed is Turkish Airlines (Figure 26c). The airline operates one of the largest and most diverse route networks globally, connecting long-haul destinations across Asia, Africa, and North America to its hub in Istanbul (IST) that plays a crucial role as a strategic transit point between East and West. The model suggests that Turkish Airlines could further expand its network by strengthening its presence in Central Asia, a region with growing population, as well as in Africa, particularly in South Africa and along the West Coast. Notably, ten out of 27 destinations identified as promising by the model have been included in airline network between 2019 and 2024. These are Seattle (SEA), Vancouver (YVR), Marrakesh (RAK), N'Djamena (NDJ), Abidjan (ABJ), Luanda (LAD), Osaka (KIX), Mexico City (MEX), Denpasar (DPS), and Colombo (CMB)."

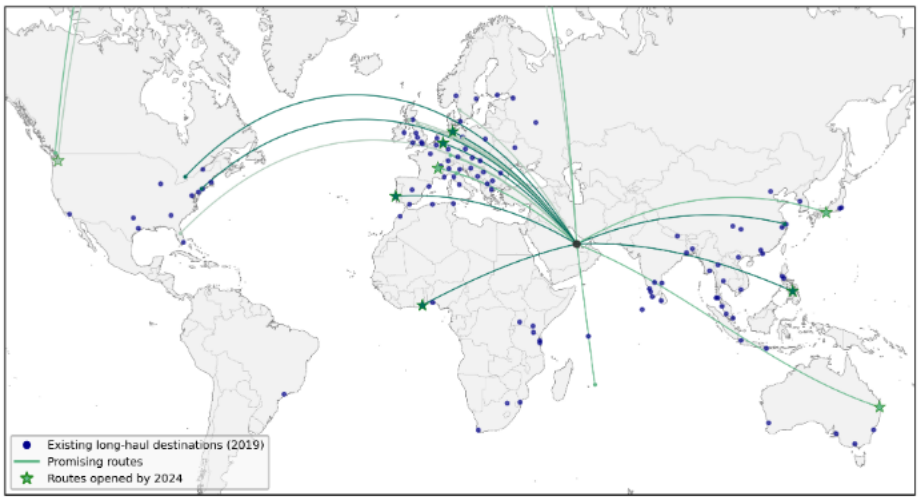
Overall, the application of the classification model to the three selected airlines demonstrated how the model can effectively identify promising routes that the airlines consider for their network development.

5.5 Conclusion

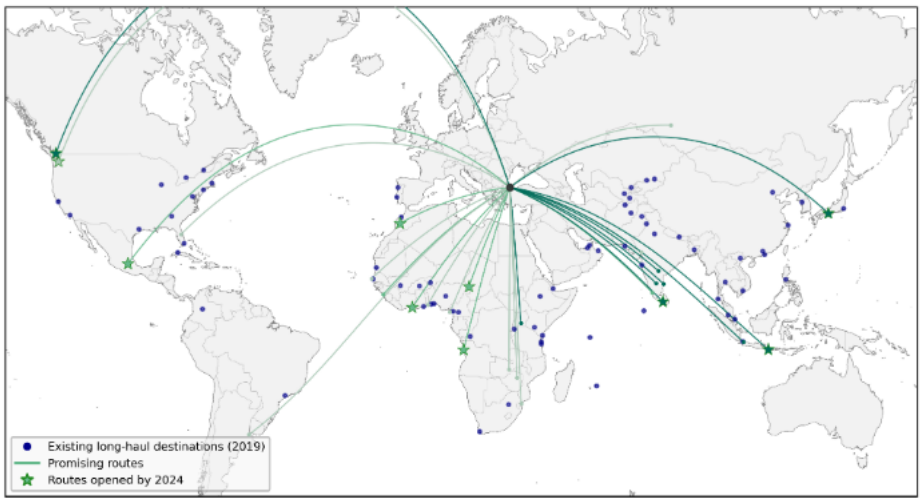
This study set out to empirically investigate the key factors explaining airline decisions to enter new long-haul markets. To this end, we developed a tailored classification model applied to a global dataset spanning six years (2014–2019) aimed at investigating promising routes based on a refined set of features derived from publicly available data. These



(a) British Airways (BA).



(b) Qatar Airways (QR).



(c) Turkish Airlines (TK).

Figure 26: Promising routes for three selected airlines as identified by the classification model applied to 2019 data. Darker lines indicate higher predicted probabilities.

features capture key dimensions such as market potential, competitive intensity, and network connectivity, enabling a structured and data-driven exploration of factors shaping airline route development decisions. Faced with a highly unbalanced dataset, we compare different balancing techniques, demonstrating the better performance of oversampling compared to undersampling.

The results empirically highlight substantial heterogeneity in route selection patterns across airlines. Carrier-specific models clearly outperform the pooled specification. Among the key drivers, connecting demand emerges as a critical factor considered by airlines when assessing the possibility of entering a new route. Some airlines appear more inclined to prioritize markets with strong socio-economic growth potential, while others place greater emphasis on fleet availability and network connectivity. This diversity reinforces the need for customized predictive tools to individual carrier profiles. The practical applicability of the proposed model is demonstrated through case studies involving three major airlines. In each case, the model demonstrates good capabilities in identifying potential routes that were subsequently launched, providing evidence of its applicability to real-world network planning decisions. Overall, the model proves effective in narrowing down the set of promising routes and can serve as a pre-screening tool, especially when integrated with airline proprietary data.

Despite the model's practical applicability and the insights provided, predicting new routes with publicly available data remains an exceptionally challenging task for several reasons. First, route entries are relatively sparse events. Accordingly, also when examining a large period, the actual number of new long-haul routes launched is relatively small, constraining the sample size and, in turn, the model's predictive power. Second, although the model incorporates a comprehensive set of factors, little information in terms of determinants can be easily collected with a global coverage and the appropriate granularity. Examples include data concerning events, public holidays, and tourism intensity- factors that often influence airline decisions but are not consistently available at the global scale. The scarcity of detailed and high-quality data and the relatively

small sample size also reduce the applicability of machine learning techniques. We tested such methods obtaining poorer results compared to the classification approach adopted in this study. Nevertheless, the availability of richer data sources and big data on touristic events, pricing strategies, and other variables affecting airlines' decisions and unlock the full potential of machine learning in this domain.

Chapter 6

Conclusion

Forecasting aviation demand is a significant challenge in the airline industry. The design of commercial aviation networks heavily relies on reliable travel demand predictions. It enables the aviation industry to plan ahead of time, evaluate whether an existing strategy needs to be revised, and prepare for new demands and challenges. Right from designing route networks, arranging for vehicles with proper seat capacity for a network route to pricing the inventory for each vehicle in a route, every operation in every planning horizon depends on estimation of the demand. Passenger transportation occupies a major share in global business and therefore, passenger demand forecasting is a key subject of research. This thesis focuses on the modeling of air passenger demand across different empirical contexts and time horizons. Starting from the longest-term strategic decisions to daily schedules, demand represents the most important input to the design, planning, and allocation of aviation resources. The aim of this thesis is to develop a comprehensive framework that spans multiple time horizons and encompasses a variety of modeling tasks. It investigates how the approach to demand modeling evolves depending on the temporal scope and the specific objectives of the analysis. Throughout Chapters 3-5, we have discussed the different models and highlighted their capability to deliver valuable insights and practically aid decision-making.

This thesis begins with an initial contribution — Chapter 3 — which focuses on short-term demand forecasting, primarily aimed at informing operational decisions within the airline industry. In this study we take into account the intrinsic relationship between demand and supply. While the added value of capturing the two-way relationship between supply and demand is not under debate, the explicit modeling of supply-demand interactions requires addressing a methodological challenge. In this respect, we have in-

vestigated supervised machine learning methodologies that are still less explored in air market demand domain. We have conducted an empirical comparison among different algorithms, discussing the trade off between results' accuracy and their interpretability. A consistent effort have been made in the direction of machine learning results' interpretation, trying to unveil meaningful insights about relevant demand determinants and their influence. The validity of the proposed approach has been evaluated through an experimental test based on real-world scenario. As final contribution, results show how machine learning algorithms can consistently improve the forecasting accuracy in a context of two-way relationship and interdependency, paving the way for a more widespread adoption and investigation of the aforementioned models. From the industrial point of view the main contribution is addressed to air operators. Indeed, these findings offer practical value: more accurate short-term forecasts enable better-informed decisions in areas such as capacity planning, scheduling, pricing strategies, and resource allocation. Additionally, the interpretability of the models provides strategic insights into market behavior, helping operators to anticipate fluctuations in demand and adapt proactively in a competitive environment.

The second contribution -Chapter 4- has focused on long term demand forecasting. Differently from the previous contribution, this study incorporates the competitive impact of high-speed rail development into the analysis of air travel demand. By doing so, it extends the temporal scope of the analysis and aims to quantify both future air travel demand and the extent of demand substitution driven by high-speed rail. We have developed a demand generation model, providing results that may affect strategic decision within the airline industry. In this paper, we have focused on the Chinese domestic area, given its fast-paced development with HSR that is becoming a more preferred alternative for its accessibility, convenience and location. Using new available data about the HSR development plan for the next decades and the socio-economic projections we have simulated how air demand will be affected. One of the most promising results, compared to the existing literature, has been the identification of air markets that will

be under the most pressure by the advent of new HSR connections. The empirical setting and the consequent discussion have demonstrated how our analysis practically aids decision-making concerning air-HSR competition. Hence, as industry-related contribution, our study highlights the need for airlines to re-adapt their networks in response to HSR development in order to remain efficient in an ever-evolving context. Beyond the implications for airlines, this study also offers valuable insights for policymakers. By forecasting the shifting dynamics between air and rail transport, it provides evidence-based guidance for infrastructure planning, investment prioritization, and intermodal dynamics. It supports the design of balanced transportation policies that promote sustainable mobility and enhance the overall efficiency of national transport networks.

The third contribution -Chapter 5- has focused on middle term allocation perspective. To conclude this thesis, we applied forecasting methodologies within the context of airline network development. Starting from the analysis of air travel demand, we expanded our scope to address a central challenge in the air transport sector: the identification of promising new routes and network development scenarios in a highly competitive and complex global environment. While network development is a core component of airline competitiveness, existing data-driven decision-support tools remain limited in their ability to scale and incorporate the multifaceted factors influencing route entry decisions. Compared to the literature, our study proposes a scalable, data-driven classification model for identifying potential long-haul destinations departing from the main carriers' hubs. By comparing established and newly opened routes with potential alternatives that have not been operated during the time period analyzed, we unveil the key drivers of route development decisions, shedding light on market characteristics, competition, and network structure. This model offers several industry-related benefits, in particular for airlines. First, it enables strategic route planning by highlighting destinations with high potential for successful market entry. Second, it supports cross-carrier benchmarking, allowing airlines to evaluate their decisions relative to competitors. Most importantly, the model demonstrates predictive accuracy by correctly identifying routes that were

launched in subsequent years, thus validating its relevance in real-world applications. As a decision-support tool, it empowers airline planners and strategists to make more informed, data-driven choices about network expansion, reduce risk in route development, and align long-term planning with market dynamics. In doing so, the model contributes directly to enhancing airline agility and profitability in the face of evolving demand and competitive pressures.

Collectively, these three studies contribute to the literature on air passenger demand by developing forecasting tools and methodologies covering different time horizons — short, medium, and long term — and decision-making contexts. Each contribution addresses a key aspect of the air transport sector: operational demand forecasting, strategic response to intermodal competition, and network development planning. Together, they offer a comprehensive and practical framework that enhances decision-making through a more advanced and critical representation of the interdependencies between supply, demand, competition, and network dynamics. Furthermore, the modeling and empirical frameworks proposed in this thesis provide valid foundations for future research toward a more comprehensive and data-driven approach to air demand forecasting, its integration in decision support tools and policy evaluation in an increasingly complex mobility landscape.

Besides the specific points identified in each chapter, three main directions for future research are identified.

First, the integration of demand models into broader airline industry tasks. While air travel demand forecasting has long been a well-established area of research, enriched by increasingly sophisticated methodologies, its investigation and application has traditionally remained somewhat isolated from other decision-making processes within the air transport system. In recent years, however, there has been growing interest in integrating demand models into a wider range of operational and strategic tasks. This shift in the literature reflects a broader industry trend toward greater interconnectivity and system-wide optimization. Future research should explore how demand models can contribute

to enhancing the overall efficiency of the aviation system. This requires the development of holistic and analytical tools capable of capturing the interdependencies across various domains, supporting more integrated and informed decision-making across the entire air transport ecosystem.

Second, embedding route scouting algorithms into network planning processes. Building on the need for more integrated and system-wide approaches, a key future research direction involves the incorporation of route scouting algorithms into airline network planning. While route scouting has proven effective as a stand-alone analytical tool to identify potential market opportunities or evaluate competitive scenarios, its integration into broader network planning frameworks remains unexplored. Incorporating such algorithms into the strategic planning process would represent a novelty in the literature concerning network planning optimization. Ultimately, this approach supports a more adaptive and responsive planning process, where network development is guided not only by historical trends but also by forward-looking insights grounded in validated analytical foundations.

Third, exploring air–HSR integration and policy-driven network adaptation. The second contribution of this thesis highlights the growing influence of high-speed rail on air transport demand and provides an overview of current infrastructure investments and intermodal dynamics in a specific geographical context. Building on these findings, a promising direction for future development lies in the development of policy-oriented research that explore how the air transport sector might adapt to the continued expansion of HSR. Rather than framing air and rail as competing modes, future studies could investigate strategies for fostering complementarity between them —such as coordinated schedules, shared hubs, and integrated ticketing systems. This would involve modeling a joint air-HSR network, examining how airlines might reconfigure their route structures, frequencies, or fleet deployment to remain viable in an evolving intermodal landscape. Such research would provide valuable insights for policymakers and industry stakeholders seeking to design mobility systems that are both competitive and collaborative, promot-

ing connectivity, sustainability, and regional accessibility through coordinated transport planning.

References

Abdelghany, A., Abdelghany, K., and Azadian, F. (2017). Airline flight schedule planning under competition. *Computers & Operations Research*, 87:20–39.

Abdelghany, A. and Guzhva, V. S. (2010). Analyzing airlines market service using panel data. *Journal of Air Transport Management*, 16(1):20–25.

Abdelghany, A. and Guzhva, V. S. (2022). Exploratory analysis of air travel demand stimulation in first-time served markets. *Journal of Air Transport Management*, 98:102162.

Abed, S. Y., Ba-Fail, A. O., and Jasimuddin, S. M. (2001). An econometric analysis of international air travel demand in saudi arabia. *Journal of air transport management*, 7(3):143–148.

Acuna-Agost, R., Thomas, E., and Lhéritier, A. (2023). Price elasticity estimation for deep learning-based choice models: an application to air itinerary choices. In *Artificial Intelligence and Machine Learning in the Travel Industry: Simplifying Complex Decision Making*, pages 3–16. Springer.

Adekunle, I. A., Quadri, K. A., and Maialeh, R. (2025). Social-economic determinant of air travel demands in africa. *Transport Policy*.

Adler, N. and Hashai, N. (2005). Effect of open skies in the middle east region. *Transportation research part a: policy and practice*, 39(10):878–894.

Adler, N., Njoya, E. T., and Volta, N. (2018). The multi-airline p-hub median problem applied to the african aviation market. *Transportation Research Part A: Policy and Practice*, 107:187–202.

Air Transport Action Group (ATAG) (2024). Aviation: Benefits beyond borders 2024. Technical report, Air Transport Action Group.

Albalate, D., Bel, G., and Fageda, X. (2015). Competition and cooperation between high-speed rail and air transportation services in europe. *Journal of transport geography*, 42:166–174.

Alumur, S. A., Campbell, J. F., Contreras, I., Kara, B. Y., Marianov, V., and O'Kelly, M. E.

(2021). Perspectives on modeling hub location problems. *European Journal of Operational Research*, 291(1):1–17.

Amadou, B., Sbihi, M., Gustavo, Z.-C. L., and Mora-Camino, F. (2021). A long term demand forecasting framework for a network of airports. *Brazilian Journal of Development*, 6:42499–42512.

Antunes, A., Martini, G., Porta, F., and Scotti, D. (2020). Air connectivity and spatial effects: regional differences in europe. *Regional Studies*, 54(12):1748–1760.

Avenali, A., D'Alfonso, T., and Reverberi, P. (2025). Airline-high speed rail cooperation, hub congestion, and airport conduct. *Transportation Research Part E: Logistics and Transportation Review*, 194:103818.

Aydemir, R. (2012). Threat of market entry and low cost carrier competition. *Journal of Air Transport Management*, 23:59–62.

Bagwell, C. B. (2005). Hyperlog—a flexible log-like transform for negative, zero, and positive valued data. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 64(1):34–42.

Bannò, M. and Redondi, R. (2014). Air connectivity and foreign direct investments: economic effects of the introduction of new routes. *European Transport Research Review*, 6:355–363.

Baykasoğlu, A., Dudaklı, N., Subulan, K., and Taşan, A. S. (2022). An integrated fleet planning model with empty vehicle repositioning for an intermodal transportation system. *Operational Research*, 22(3):2063–2098.

Behrens, C. and Pels, E. (2012). Intermodal competition in the london–paris passenger market: High-speed rail and air transport. *Journal of Urban Economics*, 71(3):278–288.

Belobaba, P., Odoni, A., and Barnhart, C. (2015). *The global airline industry*. John Wiley & Sons.

Bergantino, A. S. and Madio, L. (2020). Intermodal competition and substitution. hsr versus air transport: Understanding the socio-economic determinants of modal choice. *Research in Transportation Economics*, 79:100823.

Bilotkach, V. and Lakew, P. A. (2014). On sources of market power in the airline industry:

Panel data evidence from the us airports. *Transportation Research Part A: Policy and Practice*, 59:288–305.

Birolini, S., Cattaneo, M., Malighetti, P., and Morlotti, C. (2020). Integrated origin-based demand modeling for air transportation. *Transportation Research Part E: Logistics and Transportation Review*, 142:102050.

Birolini, S. and Jacquillat, A. (2023). Day-ahead aircraft routing with data-driven primary delay predictions. *European Journal of Operational Research*, 310(1):379–396.

Birolini, S., Jacquillat, A., Cattaneo, M., and Antunes, A. P. (2021). Airline network planning: Mixed-integer non-convex optimization with demand–supply interactions. *Transportation Research Part B: Methodological*, 154:100–124.

Birolini, S., Jacquillat, A., Schmedeman, P., and Ribeiro, N. (2023). Passenger-centric slot allocation at schedule-coordinated airports. *Transportation Science*, 57(1):4–26.

Boguslaski, C., Ito, H., and Lee, D. (2004). Entry patterns in the southwest airlines route system. *Review of Industrial Organization*, 25:317–350.

Boonekamp, T., Zuidberg, J., and Burghouwt, G. (2018). Determinants of air travel demand: The role of low-cost carriers, ethnic links and aviation-dependent employment. *Transportation Research Part A: Policy and Practice*, 112:18–28.

Brueckner, J. K., Lee, D., and Singer, E. (2014). City-pairs versus airport-pairs: A market-definition methodology for the airline industry. *Review of Industrial Organization*, 44:1–25.

Button, K. and Drexler, J. (2005). Recovering costs by increasing market share: an empirical critique of the s-curve. *Journal of Transport Economics and Policy (JTEP)*, 39(3):391–410.

Cadarso, L. and Vaze, V. (2023). Passenger-centric integrated airline schedule and aircraft recovery. *Transportation Science*, 57(3):813–837.

Cadarso, L., Vaze, V., Barnhart, C., and Marín, Á. (2017). Integrated airline scheduling: Considering competition effects and the entry of the high speed rail. *Transportation Science*, 51(1):132–154.

Cai, D.-l., Xiao, Y.-b., and Jiang, C. (2021). Competition between high-speed rail and airlines: Considering both passenger and cargo. *Transport Policy*, 110:379–393.

Calzada, J. and Fageda, X. (2019). Route expansion in the european air transport market. *Regional Studies*, 53(8):1149–1160.

Carmona-Benítez, R. B., Nieto, M. R., and Miranda, D. (2017). An econometric dynamic model to estimate passenger demand for air transport industry. *Transportation Research Procedia*, 25:17–29.

Carreira, J. S., Lulli, G., and Antunes, A. P. (2017). The airline long-haul fleet planning problem: The case of tap service to/from brazil. *European Journal of Operational Research*, 263(2):639–651.

Carson, R. T., Cenesizoglu, T., and Parker, R. (2011). Forecasting (aggregate) demand for us commercial air travel. *International journal of Forecasting*, 27(3):923–941.

Cartus, A. R., Bodnar, L. M., and Naimi, A. I. (2020). The impact of undersampling on the predictive performance of logistic regression and machine learning algorithms: a simulation study. *Epidemiology*, 31(5):e42–e44.

Cascetta, E. and Cascetta, E. (2009). Random utility theory. *Transportation systems analysis: models and applications*, pages 89–167.

Casey Jr, H. J. (1955). The law of retail gravitation applied to traffic engineering. *Traffic Quarterly*, 9(3).

Çetin, T. and Eryigit, K. Y. (2018). Estimating the economic effects of airline deregulation. *Journal of Transport Economics and Policy (JTEP)*, 52(4):404–426.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, F.-Y., Chang, Y.-H., and Lin, Y.-H. (2012). Customer perceptions of airline social responsibility and its effect on loyalty. *Journal of Air Transport Management*, 20:49–51.

Chen, J., Yan, N., Lin, S., and Chen, S. (2023). Comparative analysis of the influence of transport modes on tourism: High-speed rail or air? city-level evidence from china. *Transportation Research Record*, 2677(2):1592–1604.

Chen, Y., Guo, F., Wang, J., Cai, W., Wang, C., and Wang, K. (2020). Provincial and gridded population projection for china under shared socioeconomic pathways from 2010 to 2100. *Scientific Data*, 7(1):83.

Chen, Z. (2017). Impacts of high-speed rail on domestic air transportation in china. *Journal of Transport Geography*, 62:184–196.

Chen, Z., Ni, H., and Wang, Z. (2025). Analyzing the propagation effect of high-speed rail from air passenger to air cargo traffic in china. *Journal of Transport Geography*, 128:104335.

Cheung, T. K., Wong, C. W., and Lei, Z. (2022). Assessment of hub airports' connectivity and self-connection potentials. *Transport Policy*, 127:250–259.

Clewlow, R. R., Sussman, J. M., and Balakrishnan, H. (2014). The impact of high-speed rail and low-cost carriers on european air passenger traffic. *Transport Policy*, 33:136–143.

Cook, A., Kluge, U., Paul, A., and Cristóbal, S. (2017). Factors influencing european passenger demand for air transport. In *Air Transport Research Society World Conference*. Air Transport Research Society.

Dahiya, N., Gupta, S., and Singh, S. (2022). A review paper on machine learning applications, advantages, and techniques. *ECS Transactions*, 107(1):6137.

Dayioglu, T. and Alnipak, S. (2023). Dynamic effecting factors of air travel demand: an econometric analysis. *Quality & Quantity*, 57(4):3713–3727.

De Rus, G. and Nash, C. (2007). In what circumstances is investment in hr worthwhile?

Deese, W., y US International Trade Commission, et al. (2013). *Determinants of inbound travel to the United States*. US Internat. Trade Commission, Office of Economics.

Delahaye, T., Acuna-Agost, R., Bondoux, N., Nguyen, A.-Q., and Boudia, M. (2017). Data-driven models for itinerary preferences of air travelers and application for dynamic pricing optimization. *Journal of Revenue and Pricing Management*, 16(6):621–639.

Dobruszkes, F. (2025). Is air/high-speed rail integration the panacea to curb the impact of aviation on climate change? the case of frankfurt airport. *European Journal of Transport and Infrastructure Research*, 25(1).

Dobruszkes, F., Dehon, C., and Givoni, M. (2014). Does european high-speed rail affect the current level of air services? an eu-wide analysis. *Transportation Research Part A: Policy and Practice*, 69:461–475.

Dresner, M., Eroglu, C., Hofer, C., Mendez, F., and Tan, K. (2015). The impact of gulf carrier competition on us airlines. *Transportation Research Part A: Policy and Practice*, 79:31–41.

D'Alfonso, T., Jiang, C., and Bracaglia, V. (2016). Air transport and high-speed rail competition: Environmental implications and mitigation strategies. *Transportation Research Part A: Policy and Practice*, 92:261–276.

Fang, H., Wang, L., and Yang, Y. (2024). Competition and quality: evidence from high-speed railways and airlines. *Review of Economics and Statistics*, pages 1–16.

Firat, M., Yiltas-Kaplan, D., and Samli, R. (2021). Forecasting air travel demand for selected destinations using machine learning methods. *Journal of Universal Computer Science (JUCS)*, 27(6).

Fu, X., Lei, Z., Wang, K., and Yan, J. (2015). Low cost carrier competition and route entry in an emerging but regulated aviation market–the case of china. *Transportation Research Part A: Policy and Practice*, 79:3–16.

Fu, X., Oum, T. H., and Zhang, A. (2010). Air transport liberalization and its impacts on airline competition and air passenger traffic. *Transportation journal*, 49(4):24–41.

Fu, X., Zhang, A., and Lei, Z. (2012). Will china's airline industry survive the entry of high-speed rail? *Research in Transportation Economics*, 35(1):13–25.

Gaggero, A. A. and Piazza, G. (2021). Multilayer networks and route entry into the airline industry: Evidence from the us domestic market. *Research in Transportation Economics*, 90:101044.

Garrow, L. A. (2016). *Discrete choice modelling and air travel demand: theory and applications*. Routledge.

Geursen, I. L., Santos, B. F., and Yorke-Smith, N. (2023). Fleet planning under demand and fuel price uncertainty using actor–critic reinforcement learning. *Journal of Air Transport Management*, 109:102397.

Gillen, D. and Lall, A. (2004). Competitive advantage of low-cost carriers: some implications for airports. *Journal of Air Transport Management*, 10(1):41–50.

Givoni, M. and Banister, D. (2007). Role of the railways in the future of air transport. *Transportation planning and technology*, 30(1):95–112.

Givoni, M. and Dobruszkes, F. (2013). A review of ex-post evidence for mode substitution

and induced demand following the introduction of high-speed rail. *Transport reviews*, 33(6):720–742.

Gosling, G. D. and Ballard, D. (2019). Addressing household income distribution in air travel demand models: case study of the baltimore–washington region. *Transportation Research Record*, 2673(1):491–502.

Grosche, T., Rothlauf, F., and Heinzl, A. (2007). Gravity models for airline passenger volume estimation. *Journal of Air Transport Management*, 13(4):175–183.

Gu, H. and Wan, Y. (2020). Can entry of high-speed rail increase air traffic? price competition, travel time difference and catchment expansion. *Transport Policy*, 97:55–72.

Gu, H. and Wan, Y. (2022). Airline reactions to high-speed rail entry: Rail quality and market structure. *Transportation Research Part A: Policy and Practice*, 165:511–532.

Hahn, J. and Hausman, J. (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review*, 93(2):118–125.

Hakim, M. M. and Merkert, R. (2019). Econometric evidence on the determinants of air transport in south asian countries. *Transport Policy*, 83:120–126.

Halpern, N. and Graham, A. (2015). Airport route development: A survey of current practice. *Tourism Management*, 46:213–221.

Hanson, D., Delibasi, T. T., Gatti, M., and Cohen, S. (2022). How do changes in economic activity affect air passenger traffic? the use of state-dependent income elasticities to improve aviation forecasts. *Journal of Air Transport Management*, 98:102147.

Hausladen, I. and Schosser, M. (2020). Towards a maturity model for big data analytics in airline network planning. *Journal of Air Transport Management*, 82:101721.

Hazledine, T. (2017). An augmented gravity model for forecasting passenger air traffic on city-pair routes. *Journal of Transport Economics and Policy (JTEP)*, 51(3):208–224.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Hofer, C., Kali, R., and Mendez, F. (2018). Socio-economic mobility and air passenger demand in the us. *Transportation Research Part A: Policy and Practice*, 112:85–94.

Hsiao, C.-Y. (2008). *Passenger demand for air transportation in a hub-and-spoke network*. University of California, Berkeley.

Hsiao, C.-Y. and Hansen, M. (2011). A passenger demand model for air transportation in a hub-and-spoke network. *Transportation Research Part E: Logistics and Transportation Review*, 47(6):1112–1125.

Hsu, C.-I., Li, H.-C., and Yang, L.-H. (2013). Intertemporal demand for international tourist air travel. *Transportmetrica A: Transport Science*, 9(5):385–407.

Hsu, C.-I. and Wen, Y.-H. (2003). Determining flight frequencies on an airline network with demand–supply interactions. *Transportation Research Part E: Logistics and Transportation Review*, 39(6):417–441.

International Air Transport Association (2020). Safety report 2019. Technical report, International Air Transport Association (IATA).

International Air Transport Association (IATA) (2021). Our commitment to fly net zero by 2050. Technical report, International Air Transport Association.

International Air Transport Association (IATA) (2024). Global outlook for air transport 2024. Technical report, International Air Transport Association.

Jaillet, P., Song, G., and Yu, G. (1996). Airline network design and hub location problems. *Location science*, 4(3):195–212.

Javadinasr, M., Maggasy, T., Mohammadi, M., Mohammadain, K., Rahimi, E., Salon, D., Conway, M. W., Pendyala, R., and Derrible, S. (2022). The long-term effects of covid-19 on travel behavior in the united states: A panel study on work from home, mode choice, online shopping, and air travel. *Transportation Research Part F: Traffic Psychology and Behaviour*, 90:466–484.

Jiang, C., D'Alfonso, T., and Wan, Y. (2017). Air-rail cooperation: Partnership level, market structure and welfare implications. *Transportation Research Part B: Methodological*, 104:461–482.

Jiang, C., Wang, K., Wang, Q., and Yang, H. (2022). The impact of high-speed rail competition on airline on-time performance. *Transportation Research Part B: Methodological*, 161:109–127.

Jiang, C. and Zhang, A. (2014). Effects of high-speed rail and airline cooperation under hub airport capacity constraint. *Transportation Research Part B: Methodological*, 60:33–49.

Jiang, C. and Zhang, A. (2016). Airline network choice and market coverage under high-speed rail competition. *Transportation Research Part A: Policy and Practice*, 92:248–260.

Jiang, M., Jiang, C., Xiao, Y.-b., and Wang, C. (2021). Air-hsr cooperation: Impacts on service frequency and environment. *Transportation Research Part E: Logistics and Transportation Review*, 150:102336.

Jiménez, J. L. and Betancor, O. (2012). When trains go faster than planes: the strategic reaction of airlines in spain. *Transport Policy*, 23:34–41.

Jorge-Calderón, J. (1997). A demand model for scheduled airline services on international european routes. *Journal of Air Transport Management*, 3(1):23–35.

Jungmittag, A. (2016). Combination of forecasts across estimation windows: An application to air travel demand. *Journal of Forecasting*, 35(4):373–380.

Kampp, M., Sedelmeier, J., Schüth, J., Thust, M., Kaiser, D., Scherr, W., Schlaich, J., and 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.

Keane, M. P. and Neal, T. (2024). A practical guide to weak instruments. *Annual Review of Economics*, 16.

Khanzode, K. C. A. and Sarode, R. D. (2020). Advantages and disadvantages of artificial intelligence and machine learning: A literature review. *International Journal of Library & Information Science (IJLIS)*, 9(1):3.

Kim, M.-S., Kim, K.-W., and Park, S.-S. (2012). A study on the air travel demand forecasting using time series arima-intervention model. *Journal of the korean Society for Aviation and Aeronautics*, 20(1):66–75.

Koç, İ. and Arslan, E. (2018). Demand forecasting for domestic air transportation in turkey using artificial neural networks. In *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, pages 1–6. IEEE.

Kölker, K. and Lütjens, K. (2015). Using genetic algorithms to solve large-scale airline network planning problems. *Transportation research procedia*, 10:900–909.

Koppelman, F. S., Coldren, G. M., and Parker, R. A. (2008). Schedule delay impacts on air-travel itinerary demand. *Transportation Research Part B: Methodological*, 42(3):263–273.

Kristoffersson, I. and 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(4):62–88.

Kroes, E. and Savelberg, F. (2019). Substitution from air to high-speed rail: the case of amsterdam airport. *Transportation Research Record*, 2673(5):166–174.

Kubus, M. (2020). Evaluation of resampling methods in the class unbalance problem. *Econometrics. Ekonometria. Advances in Applied Data Analytics*, 24(1):39–50.

Lai, S. B. S., Shahri, N., Mohamad, M. B., Rahman, H., and Rambli, A. B. (2021). Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data. *Mathematics and Statistics*, 9(3):379–385.

Lee, E., Kawakita, T., Huai, Y., Lo, H. K., and Zhang, A. (2024). Airline and high-speed rail collaboration and competition under travel time variability. *Transportation Research Part A: Policy and Practice*, 185:104104.

Lee, S. Y., Yoo, K. E., and Park, Y. (2014). A continuous connectivity model for evaluation of hub-and-spoke operations. *Transportmetrica A: Transport Science*, 10(10):894–916.

Lhéritier, A., Bocamazo, M., Delahaye, T., and Acuna-Agost, R. (2019). Airline itinerary choice modeling using machine learning. *Journal of choice modelling*, 31:198–209.

Li, H., Strauss, J., and Lu, L. (2019). The impact of high-speed rail on civil aviation in china. *Transport Policy*, 74:187–200.

Li, J., Sun, X., Cong, W., Miyoshi, C., Ying, L. C., and Wandelt, S. (2024). On the air-hsr mode substitution in china: From the carbon intensity reduction perspective. *Transportation Research Part A: Policy and Practice*, 180:103977.

Li, Z.-C. and Sheng, D. (2016). Forecasting passenger travel demand for air and high-speed rail integration service: A case study of beijing-guangzhou corridor, china. *Transportation Research Part A: Policy and Practice*, 94:397–410.

Lieshout, R., Malighetti, P., Redondi, R., and Burghouwt, G. (2016). The competitive landscape of air transport in europe. *Journal of Transport Geography*, 50:68–82.

Lin, Y., Zhang, J.-w., and Liu, H. (2019). Deep learning based short-term air traffic flow prediction considering temporal–spatial correlation. *Aerospace Science and Technology*, 93:105113.

Liu, S., Wan, Y., Ha, H.-K., Yoshida, Y., and Zhang, A. (2019). Impact of high-speed rail network development on airport traffic and traffic distribution: Evidence from china and japan. *Transportation Research Part A: Policy and Practice*, 127:115–135.

Lundaeva, K. A., Saranin, Z. A., Pospelov, K. N., and Gintciak, A. M. (2024). Demand forecasting model for airline flights based on historical passenger flow data. *Applied Sciences*, 14(23):11413.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lurkin, V., Garrow, L. A., Higgins, M. J., Newman, J. P., and Schyns, M. (2017). Accounting for price endogeneity in airline itinerary choice models: An application to continental us markets. *Transportation Research Part A: Policy and Practice*, 100:228–246.

Lurkin, V., Garrow, L. A., Higgins, M. J., Newman, J. P., and Schyns, M. (2018). Modeling competition among airline itineraries. *Transportation Research Part A: Policy and Practice*, 113:157–172.

Ma, X., Jiang, C., and Zhang, Z. (2024). Spatio-temporal pattern of china's urban network under the substitution effect of high-speed rail on aviation. *Transportation Research Record*, 2678(8):74–92.

Maheshwari, A., Davendralingam, N., and DeLaurentis, D. A. (2018). A comparative study of machine learning techniques for aviation applications. In *2018 Aviation Technology, Integration, and Operations Conference*, page 3980.

Martín, J. C. and Nombela, G. (2007). Microeconomic impacts of investments in high speed trains in spain. *The Annals of Regional Science*, 41(3):715–733.

Megahed, F. M., Chen, Y.-J., Megahed, A., Ong, Y., Altman, N., and Krzywinski, M. (2021). The class imbalance problem.

Milan, J. (1993). A model of competition between high speed rail and air transport. *Transportation planning and technology*, 17(1):1–23.

Mizutani, J. and Sakai, H. (2021). Which is a stronger competitor, high speed rail, or low cost carrier, to full service carrier?–effects of hsr network extension and lcc entry on fsc's airfare in japan. *Journal of Air Transport Management*, 90:101965.

Mohammadi, M., Jula, P., and Tavakkoli-Moghaddam, R. (2019). Reliable single-allocation hub location problem with disruptions. *Transportation Research Part E: Logistics and Transportation Review*, 123:90–120.

Mohri, S. S., Nasrollahi, M., Pirayesh, A., and Mohammadi, M. (2022). An integrated global airline hub network design with fleet planning. *Computers & Industrial Engineering*, 164:107883.

Mokhtarian, P. L. (2016). Presenting the independence of irrelevant alternatives property in a first course on logit modeling. *Journal of choice modelling*, 21:25–29.

Morrison, S. A. and Winston, C. (1990). The dynamics of airline pricing and competition. *The American Economic Review*, 80(2):389–393.

Moyano, A., Martínez, H. S., and Coronado, J. M. (2018). From network to services: A comparative accessibility analysis of the spanish high-speed rail system. *Transport Policy*, 63:51–60.

Mueller, F. (2015). Estimating own-price elasticities of air travel demand: The case of norway. Master's thesis, Høgskolen i Molde-Vitenskapelig høgskole i logistikk.

Müller, M. E. (2004). Can user models be learned at all? inherent problems in machine learning for user modelling. *The Knowledge Engineering Review*, 19(1):61–88.

Nicholas, A. (2021). Forecasting us overseas travelling with univariate and multivariate models. *Journal of Forecasting*, 40(6):963–976.

Nurhidayat, A. Y., Widyastuti, H., Sutikno, and Upahita, D. P. (2023). Research on passengers' preferences and impact of high-speed rail on air transport demand. *Sustainability*, 15(4):3060.

Oliveira, A. V. (2008). An empirical model of low-cost carrier entry. *Transportation Research Part A: Policy and Practice*, 42(4):673–695.

Olmedo, E. (2016). Comparison of near neighbour and neural network in travel forecasting. *Journal of Forecasting*, 35(3):217–223.

Oommen, T., Baise, L. G., and Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, 43:99–120.

Ozmec-Ban, M. and Babić, R. Š. (2023). A literature review of recent causality analyses between air transport demand and socio-economic factors. *Transportation Research Procedia*, 73:85–93.

Park, Y. and Ha, H.-K. (2006). Analysis of the impact of high-speed railroad service on air transport demand. *Transportation Research Part E: Logistics and Transportation Review*, 42(2):95–104.

Pita, J. P., Adler, N., and Antunes, A. P. (2014). Socially-oriented flight scheduling and fleet assignment model with an application to norway. *Transportation Research Part B: Methodological*, 61:17–32.

Pita, J. P., Antunes, A. P., Barnhart, C., and de Menezes, A. G. (2013). Setting public service obligations in low-demand air transportation networks: Application to the azores. *Transportation Research Part A: Policy and Practice*, 54:35–48.

Rajendran, S., Srinivas, S., and Grimshaw, T. (2021). Predicting demand for air taxi urban aviation services using machine learning algorithms. *Journal of Air Transport Management*, 92:102043.

Redyuk, S., Schelter, S., Rukat, T., Markl, V., and Biessmann, F. (2019). Learning to validate the predictions of black box machine learning models on unseen data. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–4.

Ren, T., Huang, H.-J., Luo, S.-d., and Nie, Y. M. (2023). High-speed rail in china: Implications for intercity commuting and urban spatial structure. *Sustainable Cities and Society*, 97:104719.

Román, C., Espino, R., and Martín, J. C. (2007). Competition of high-speed train with air transport: The case of madrid–barcelona. *Journal of Air Transport Management*, 13(5):277–284.

Rosenfield, A., Attanucci, J. P., and Zhao, J. (2020). A randomized controlled trial in travel demand management. *Transportation*, 47(4):1907–1932.

Santana, M., De La Vega, J., Morabito, R., and Pureza, V. (2023). The aircraft recovery

problem: A systematic literature review. *EURO Journal on Transportation and Logistics*, page 100117.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

Schosser, M. and Schosser, M. (2020). Status quo of strategic network planning in airlines. *Big Data to Improve Strategic Network Planning in Airlines*, pages 161–194.

Scotti, D. and Dresner, M. (2015). The impact of baggage fees on passenger demand on us air routes. *Transport Policy*, 43:4–10.

Sharma, A., Jakhar, S. K., Vlachos, I., and Kumar, S. (2024). Advances in hub location problems: a literature review and research agenda. *International Journal of Productivity and Performance Management*.

Silva, J. T. M., Santos, L. H., Dias, A. T., and Tadeu, H. F. B. (2019). Intermittent demand forecasting for aircraft inventories: a study of brazilian's boeing 737ng aircraft´ s spare parts management. *TRANSPORTES*, 27(2):102–116.

Sismanidou, A. and Tarradellas, J. (2017). Traffic demand forecasting and flexible planning in airport capacity expansions: Lessons from the madrid-barajas new terminal area master plan. *Case Studies on Transport Policy*, 5(2):188–199.

Soylu, B. and Katip, H. (2019). A multiobjective hub-airport location problem for an airline network design. *European Journal of Operational Research*, 277(2):412–425.

Srinivas, S. and Ravindran, A. R. (2018). Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. *Expert Systems with Applications*, 102:245–261.

Starita, S., Strauss, A. K., Fei, X., Jovanović, R., Ivanov, N., Pavlović, G., and Fichert, F. (2020). Air traffic control capacity planning under demand and capacity provision uncertainty. *Transportation Science*, 54(4):882–896.

Strauss, J., Li, H., and Cui, J. (2021). High-speed rail's impact on airline demand and air carbon emissions in china. *Transport Policy*, 109:85–97.

Su, M., Luan, W., Fu, X., Yang, Z., and Zhang, R. (2020). The competition effects of low-cost carriers and high-speed rail on the chinese aviation market. *Transport Policy*, 95:37–46.

Su, M., Luan, W., and Sun, T. (2019). Effect of high-speed rail competition on airlines' intertemporal price strategies. *Journal of Air Transport Management*, 80:101694.

Suh, D. Y. and Ryerson, M. S. (2019). Forecast to grow: aviation demand forecasting in an era of demand uncertainty and optimism bias. *Transportation Research Part E: Logistics and Transportation Review*, 128:400–416.

Sun, X., Wandelt, S., and Zhang, A. (2021a). Comparative accessibility of chinese airports and high-speed railway stations: A high-resolution, yet scalable framework based on open data. *Journal of Air Transport Management*, 92:102014.

Sun, X., Wandelt, S., and Zhang, A. (2022). Covid-19 pandemic and air transportation: Summary of recent research, policy consideration and future research directions. *Transportation research interdisciplinary perspectives*, 16:100718.

Sun, X., Wandelt, S., Zheng, C., and Zhang, A. (2021b). Covid-19 pandemic and air transportation: Successfully navigating the paper hurricane. *Journal of Air Transport Management*, 94:102062.

Sun, X., Zhang, Y., and Wandelt, S. (2017). Air transport versus high-speed rail: An overview and research agenda. *Journal of Advanced Transportation*, 2017(1):8426926.

Sun, X., Zheng, C., Li, J., Jiang, C., Zhang, A., and Wandelt, S. (2024). A review on research regarding hsr interactions with air transport and outlook for future research challenges. *Transport Policy*.

Teodorović, D., Kalić, M., and Pavković, G. (1994). The potential for using fuzzy set theory in airline network design. *Transportation Research Part B: Methodological*, 28(2):103–121.

Teoh, L. E. and Khoo, H. L. (2016). Fleet planning decision-making: Two-stage optimization with slot purchase. *Journal of Optimization*, 2016(1):8089794.

Tirtha, S. D., Bhowmik, T., and Eluru, N. (2023). Understanding the factors affecting airport level demand (arrivals and departures) using a novel modeling approach. *Journal of Air Transport Management*, 106:102320.

Trapero, J. R., Fildes, R., and Davydenko, A. (2011). Nonlinear identification of judgmental forecasts effects at sku level. *Journal of Forecasting*, 30(5):490–508.

Truong, D. (2021). Using causal machine learning for predicting the risk of flight delays in air transportation. *Journal of Air Transport Management*, 91:101993.

Tsunoda, Y. (2018). Transportation policy for high-speed rail competing with airlines. *Transportation Research Part A: Policy and Practice*, 116:350–360.

Ustyannie, W. and Suprapto, S. (2020). Oversampling method to handling imbalanced datasets problem in binary logistic regression algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 14(1):1–10.

Valdes, V. (2015). Determinants of air travel demand in middle income countries. *Journal of Air Transport Management*, 42:75–84.

Vowles, T. M. (2006). Airfare pricing determinants in hub-to-hub markets. *Journal of Transport Geography*, 14(1):15–22.

Wan, Y., Ha, H.-K., Yoshida, Y., and Zhang, A. (2016). Airlines' reaction to high-speed rail entries: Empirical study of the northeast asian market. *Transportation Research Part A: Policy and Practice*, 94:532–557.

Wandelt, S., Signori, A., Chang, S., Wang, S., Du, Z., and Sun, X. (2025). Unleashing the potential of operations research in air transport: A review of applications, methods, and challenges. *Journal of Air Transport Management*, 124:102747.

Wang, H., Zhu, R., and Ma, P. (2018a). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844.

Wang, J., Jiao, J., Du, C., and Hu, H. (2015). Competition of spatial service hinterlands between high-speed rail and air transport in china: Present and future trends. *Journal of Geographical Sciences*, 25:1137–1152.

Wang, K., Tsui, K. W. H., Liang, L., and Fu, X. (2017). Entry patterns of low-cost carriers in hong kong and implications to the regional market. *Journal of Air Transport Management*, 64:101–112.

Wang, K., Xia, W., Zhang, A., and Zhang, Q. (2018b). Effects of train speed on airline demand and price: Theory and empirical evidence from a natural experiment. *Transportation Research Part B: Methodological*, 114:99–130.

Wang, L., Liu, Y., Mao, L., and Sun, C. (2018c). Potential impacts of china 2030 high-speed rail network on ground transportation accessibility. *Sustainability*, 10(4):1270.

Wang, S. and Gao, Y. (2021). A literature review and citation analyses of air travel demand studies published between 2010 and 2020. *Journal of Air Transport Management*, 97:102135.

Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., and Chen, S.-C. (2019). A framework for airfare price prediction: a machine learning approach. In *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)*, pages 200–207. IEEE.

Wang, T. and Sun, F. (2022). Global gridded gdp data set consistent with the shared socioeconomic pathways. *Scientific data*, 9(1):221.

Wang, X., Meng, X., and Long, Y. (2022a). Projecting 1 km-grid population distributions from 2020 to 2100 globally under shared socioeconomic pathways. *Scientific Data*, 9(1):563.

Wang, X., Peng, J., Tang, J., Lu, Q., and Li, X. (2022b). Investigating the impact of adding new airline routes on air transportation resilience in china. *Transport Policy*, 125:79–95.

Wang, Y., Jiang, X., and Ma, J. (2024). Effects of price discrimination based on the heterogeneity of passenger travel purpose on air-hsr competition: Implications for traffic, welfare and government regulation. *Travel Behaviour and Society*, 36:100758.

Wang, Y., Lu, Q., Cao, X., Zhou, X., Latora, V., Tong, L. C., and Du, W. (2020a). Travel time analysis in the chinese coupled aviation and high-speed rail network. *Chaos, Solitons & Fractals*, 139:109973.

Wang, Y., Sun, L., Teunter, R. H., Wu, J., and Hua, G. (2020b). Effects of introducing low-cost high-speed rail on air-rail competition: Modelling and numerical analysis for paris-marseille. *Transport Policy*, 99:145–162.

Wanner, J., Herm, L.-V., and Janiesch, C. (2020). How much is the black box? the value of explainability in machine learning models.

Webb, E. K. (1970). Profile relationships: The log-linear range, and extension to strong stability. *Quarterly Journal of the Royal Meteorological Society*, 96(407):67–90.

Wei, W. and Hansen, M. (2005). Impact of aircraft size and seat availability on airlines' demand

and market share in duopoly markets. *Transportation Research Part E: Logistics and Transportation Review*, 41(4):315–327.

Wen, X., Chung, S.-H., Ma, H.-L., and Khan, W. A. (2024). Airline crew scheduling with sustainability enhancement by data analytics under circular economy. *Annals of Operations Research*, 342(1):959–985.

Wong, C. W., Cheung, T. K. Y., and Zhang, A. (2023). A connectivity-based methodology for new air route identification. *Transportation Research Part A: Policy and Practice*, 173:103715.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wu, J., Zhang, P.-w., Wang, Y., and Shi, J. J. (2022). Integrated aviation model and metaheuristic algorithm for hub-and-spoke network design and airline fleet planning. *Transportation Research Part E: Logistics and Transportation Review*, 164:102755.

Wu, S. and Han, D. (2022). Accessibility of high-speed rail (hsr) stations and hsr–air competition: Evidence from china. *Transportation Research Part A: Policy and Practice*, 166:262–284.

Xia, W. and Zhang, A. (2016). High-speed rail and air transport competition and cooperation: A vertical differentiation approach. *Transportation Research Part B: Methodological*, 94:456–481.

Xu, H., Shi, J., and Wang, T. (2022). Departure flight delay prediction model based on deep fully connected neural network. *J. Comput. Appl*, 42(10):3283.

Xu, M., Shuai, B., Wang, X., Liu, H., and Zhou, H. (2023). Analysis of the accessibility of connecting transport at high-speed rail stations from the perspective of departing passengers. *Transportation Research Part A: Policy and Practice*, 173:103714.

Xu, Y., Adler, N., Wandelt, S., and Sun, X. (2024). Competitive integrated airline schedule design and fleet assignment. *European Journal of Operational Research*, 314(1):32–50.

Yang, H., Du, D., Wang, J., Wang, X., and Zhang, F. (2023). Reshaping china's urban networks and their determinants: High-speed rail vs. air networks. *Transport Policy*, 143:83–92.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., and Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling

imbalanced datasets. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)*, pages 13–22. Springer.

Yu, B., Guo, Z., Asian, S., Wang, H., and Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, 125:203–221.

Yu, K., Strauss, J., Liu, S., Li, H., Kuang, X., and Wu, J. (2021). Effects of railway speed on aviation demand and co2 emissions in china. *Transportation research Part D: Transport and environment*, 94:102772.

Zachariah, R. A., Sharma, S., and Kumar, V. (2023). Systematic review of passenger demand forecasting in aviation industry. *Multimedia tools and applications*, 82(30):46483–46519.

Zhang, Q., Yang, H., and Wang, Q. (2017a). Impact of high-speed rail on china's big three airlines. *Transportation Research Part A: Policy and Practice*, 98:77–85.

Zhang, R., Johnson, D., Zhao, W., and Nash, C. (2019). Competition of airline and high-speed rail in terms of price and frequency: Empirical study from china. *Transport Policy*, 78:8–18.

Zhang, Y., Hu, R., Chen, R., Cai, D.-l., and Jiang, C. (2024). Competition in cargo and passenger between high-speed rail and airlines—considering the vertical structure of transportation. *Transport Policy*, 151:120–133.

Zhang, Y., Wang, K., and Fu, X. (2017b). Air transport services in regional australia: Demand pattern, frequency choice and airport entry. *Transportation Research Part A: Policy and Practice*, 103:472–489.

Zhou, J., Qiu, G., et al. (2018a). China's high-speed rail network construction and planning over time: a network analysis. *Journal of Transport Geography*, 70:40–54.

Zhou, J., Yang, L., Li, L., et al. (2018b). The implications of high-speed rail for chinese cities: Connectivity and accessibility. *Transportation Research Part A: Policy and Practice*, 116:308–326.

Zhu, F., Wu, X., and Cao, C. (2021). High-speed rail and air transport competition under high flight delay conditions in china: A case study of the beijing-shanghai corridor. *Utilities Policy*, 71:101233.

Zou, L. and Yu, C. (2020). The evolving market entry strategy: A comparative study of southwest and jetblue. *Transportation Research Part A: Policy and Practice*, 132:682–695.