Journal of the Air Transport Research Society 3 (2024) 100043

Contents lists available at ScienceDirect

Journal of the Air Transport Research Society

journal homepage: www.elsevier.com/locate/jatrs

A hybrid econometric–machine learning framework to support market development in intercity passenger transport: the causal and predictive analytics of economic mobility features

Alessandro V.M. Oliveira [a,*], Luca J. Santos [a], Dante Mendes Aldrighi [b]

[a] Center for Airline Economics, Aeronautics Institute of Technology, Brazil
[b] University of São Paulo, Brazil

ARTICLE INFO

JEL Classification:
D22
L11
L93

Keywords:
Econometrics
Machine learning
Transport equity
Consumer insights
Business analytics

ABSTRACT

Capturing potential travel demand is crucial for carriers to improve their market performance, especially in developing economies with an emerging middle class and increasing socioeconomic inclusion. However, the impact of upward economic mobility on deregulated transport systems and how carriers can capitalize on this trend to increase revenues remain unclear, as this phenomenon is influenced by several confounding factors. This study aims to estimate and decompose the impact of the inclusiveness boom and bust in Brazil on its domestic intercity travel industry. By utilizing Instrumental Variables Least Absolute Shrinkage and Selection Operator (IV-LASSO) and Quantile Regression, our high-dimension sparse approach intends to estimate the effects of a set of economic mobility features on travel markets. We also employ a meta-machine learning approach based on Stacking Regression to assess the contribution of these features to revenue generation. Our findings suggest that airlines are more efficient than bus carriers at implementing market development strategies to achieve effective market inclusion. The customer retention rate for bus carriers is 32% lower, indicating the need to enhance demand management. Moreover, Stacking outperforms base machine learners in predicting revenues for both transport modes. Finally, an event study carried out for the economic downturn period shows a persistent adverse effect on demand and prices and identifies the moments when the machine learning models perform most poorly.

1. Introduction

In developing countries, the recovery of tourism after the pandemic has been hindered by socioeconomic constraints—such as unequal income distribution, financial restrictions, limited access to credit, and economic uncertainty. Promoting social inclusion that drives economic mobility (EM) remains a major challenge for middle and low-income economies, particularly in the current context of economic inequalities exacerbated by recent global crises. In emerging travel markets, upward EM creates opportunities for firms to explore new potential markets, improve segmentation and revenue management, and foster market inclusion and transport equity by incorporating potential customers previously excluded during economic crises.

Several studies have highlighted the impact of socioeconomic factors on travel demand growth in emerging economies. Foremost among them, Brazil experienced moderate economic growth and living conditions' steady improvement for the less-favored classes from the mid-1990s to the mid-2010s. The deregulation of its airline industry in the early 2000s along with socioeconomic changes were decisive to attract many new airline consumers. Despite attempts to revive high-speed train projects, buses have remained as the only alternative for medium to long-distance intercity transportation. The liberalization of the intercity bus transportation industry in the mid-2010s enabled the entry of new technology-driven platforms offering shared and on-demand charter bus services, which have operated as alternatives to traditional scheduled bus lines.

This study explores the impact of EM on the Brazilian domestic intercity travel industry, focusing on demand, price, and revenue. Our approach combines econometric models for causal analysis with machine learning for feature importance and pattern recognition. We extend Huang & Rojas' (2013, 2014) logit model to incorporate EM features, improving its capacity to capture aggregate demand and

* Corresponding author.
E-mail address: alessandro@ita.br (A.V.M. Oliveira).

https://doi.org/10.1016/j.jatrs.2024.100043
Received 12 July 2024; Received in revised form 3 October 2024; Accepted 5 October 2024
Available online 16 October 2024

unobserved options. We also employ Instrumental Variables LASSO (IV-LASSO) and Quantile Regression. The machine learning component includes multiple base learners, such as Gradient Boosting and Neural Networks, with a meta-learner using Stacking Regression. By integrating these methodologies, we provide both explanatory insights and improved predictive power. To our knowledge, this is the first empirical study to use a hybrid econometric-machine learning framework to assess the effect of EM features on carrier revenue performance in emerging travel markets.

The remainder of the paper is structured as follows. Section 2 reviews the literature on emerging economies' travel demand growth, focusing on potential demand drivers related to social, financial, and digital inclusion. Section 3 describes the study object and the methodology used to investigate it. Section 4 presents the results from the IV-LASSO regression analysis, providing insights from a reduced-form pricing model. Section 5 discusses the results of the Quantile Regression approach. Section 6 outlines the machine learning modeling approach and systematically assesses the predictions generated by the stacking regression framework. Section 7 features the event study examining the dynamics of quantities, prices, and revenues during Brazil's economic downturn in the mid-2010s. Finally, Section 8 offers concluding remarks.

2. Literature review

2.1. Transport and economic mobility socioeconomic drivers

The socioeconomic conditions of a country or region are amid the main factors conditioning the chances for an individual or a family to be economically better off. Accordingly, sustainable economic growth plays a critical role in generating better employment opportunities and increased income for all individuals, thus contributing to reducing income inequality. Inclusive economic development, which also prioritize investments in education, housing in urban areas, and transportation infrastructure, fosters equity. Moreover, public policies aimed at improving low-income people's living conditions and promoting social, financial, and digital inclusion can be pivotal in achieving EM. Potential consumer demand in the medium- and long-haul travel markets hinges on growing EM. Transport companies interested in expanding markets, sales, and profits must develop analytical and predictive tools to evaluate demand potential, taking stock of all these factors to define their market development strategies.

The diversification of the consumption basket and the social inclusion of emerging consumers also require financial inclusion and digital inclusion. According to the World Bank, financial inclusion means that "individuals and businesses have access to useful and affordable financial products and services that meet their needs—transactions, payments, savings, credit, and insurance—delivered in a responsible and sustainable way."[1] Garcia-Escribano & Han (2015) note that consumer credit led the credit expansion in Brazil, Indonesia, and Turkey, which grew by more than 10% a year in the 2000s. Several studies, such as Olney (1999), Di Maggio & Kermani (2017), and Bahadir, De & Lastrapes (2020), endorse the view that higher access to credit boosts short-term household consumption. Hanke (2019) documents that airlines' novice passengers in developing countries are usual, with many of them without a bank account or a debit or credit card nor the experience in purchasing airline tickets and other ancillary services.

Digital inclusion is a type of social inclusion that "ensures individuals and disadvantaged groups to have access to, and skills to use, Information and Communication Technologies (ICT) and are therefore able to participate in and benefit from today's growing knowledge and information society."[2] According to NHS Digital (2020), digital inclusion covers three aspects: digital skills, as the ability to use smartphones and the Internet; connectivity, which means access to the Internet through broadband, Wi-Fi, and mobile; and accessibility, which includes services designed to meet users' needs. Studying the US airline market between 1993 and 2007, Dana Jr & Orlov (2014) show that the intensified use of Internet for searching airfares reduces market frictions, allowing airlines to meet the demand with less capacity and higher load factors. Sun et al. (2024) point out that cognitive and affective factors can also interfere in passengers' travel decisions. In this context, it can be argued that digital inclusion may engender social influences that attract potential passengers, particularly in the form of digital word of mouth, or eWOM (electronic Word of Mouth). Through access to information about opinions, recommendations, and feedback on tourist destinations and trips on digital media such as social networks, forums, blogs, and review sites, the desire to travel may be fostered among first-time travelers. Among other related studies are Crespo-Almendros & Del-Barrio García (2016), Seo, Park & Choi (2020), Wattanacharoensil, Schuckert & Graham (2016), Zaki Ahmed & Rodríguez-Díaz (2020), Kim & Hyun (2019).

2.2. Economic mobility, market development, and transport equity

Understanding the relationship between EM and transportation demand, prices, and revenues in medium-to-long-distance travel markets can shed light on its effect on transportation equity. Greater transportation equity is associated with a fairer distribution of resources and benefits from existing transportation infrastructure and operator services across different socioeconomic groups. Di Ciommo & Shiftan (2017) emphasize that transport policies and investments should include equity among their goals. Transportation equity involves the distribution of transportation benefits and costs across population groups of a society based on more egalitarian distributive principles. Thus, accessibility to reliable modes of transportation should be a critical input in the prescription of public policies directed to improve transportation equity.

Although public authorities can enhance transportation equity through supply-side regulation of transportation terminals and carriers or via subsidy schemes to induce demand, transportation equity in medium-to-long-distance travel markets has mostly been driven by economic deregulation. In many countries, increased competition resulting from deregulation in some industries, as airlines, long-distance bus and rail transportation, has generally led to lower prices, benefitting mostly low-income consumers previously excluded from the market. Competition can also expand the availability of transportation services to new locations, thereby improving equity from a spatial perspective. Overall, deregulation has fostered transportation equity in travel industries across the world.

In addition to lower fares and new services, competition can promote market inclusion through companies' market development strategies, contemplating investments in market research, product development, advertising, public relations, and any other initiatives to expand the reach and appeal of a service or a good. Economic mobility creates opportunities for opening new consumer segments underserved or neglected by companies. In sum, market development strengthens transportation operators' market segmentation capabilities through revenue management, which may simultaneously capture demand, increase profits, and stimulate transportation equity.

EM refers to individuals' ability to change their socioeconomic position in a society over time. Works dealing with this topic generally address intergenerational EM, which investigates issues such as the

[1] "Financial Inclusion," available at www.worldbank.org/en/topic/financialinclusion/overview.

[2] "Digital Inclusion definition," available on digitalinclusion.nz/about/digital-inclusion-definition.

probability that a child reaches the top quintile of the national income distribution starting from a family in the bottom quintile, and how parents' income influences their children's income. Chetty et al. (2014), from a data set of more than 40 million children in the United States, show that upward EM is correlated with lower income inequality, better primary schools, greater social capital, and family stability. Bergantino & Madio (2020) investigate intermodal substitution between high-speed rail (HSR) and air transport on the Bari-Rome and Brindisi-Rome routes, highlighting the importance of socio-economic variables in modal decision-making. They find that the probability of opting for HSR increases with age, income, education, and business travel.

Closely related to this paper, Hofer et al. (2018) examine the impact of EM on domestic passenger enplanements in U.S. airports. Using absolute and relative mobility metrics proposed by Chetty et al. (2014), they emphasize the importance of adding EM to the conventional factors employed to analyze the demand for air travel and fares. Their counterintuitive finding is that EM is linked to lower passenger volume: greater EM only increases the numbers of passengers when prices come down, having ceteris paribus a decreasing effect on demand.

We investigate the impact of EM on transportation demand and prices in Brazil, an emerging country characterized by greater inequalities and more limited social inclusion and EM than those studied by Hofer et al. (2018). Furthermore, we propose an alternative approach to their work, assessing the model's predictive performance. Without specifically addressing intergenerational economic mobility, we employ a broad concept of economic mobility that encompasses overall improvements in individuals' socioeconomic position irrespective of their generation. As this study focuses on a developing country whose socioeconomic conditions have rapidly shifted from boom to bust, we deem that many forms of socioeconomic changes, both intergenerational and intragenerational, may have occurred over the period.

Also intimately related to this paper's concerns, Santos et al. (2021) analyze the impact of the coronavirus crisis on air travel demand in Brazil. Their contribution lies in investigating the role of EM by using proxies for income inclusion (HDI), digital inclusion (number of cell phones), and financial inclusion (credit and indebtedness). Building on their framework, our approach goes further, incorporating a broader set of EM metrics, a second mode of transport, and a pricing equation. We innovate by including an inequality-adjusted income indicator as well as proxies for unemployment, customer relationship management efforts by carriers, consumer confidence, service awareness, and customer retention and switching behavior. Moreover, we analyze the asymmetry and heterogeneity of demand and prices between airlines and buses. Lastly, focusing on the recession event's impact during the sample period delves deeper into the dynamics of quantities, prices, and total revenues in the travel markets. We extend the logit model of Huang & Rojas (2013, 2014) to incorporate characteristics specific to emerging economies. With this theoretical framework, our estimation strategy includes several econometric controls not addressed by them, such as transport mode/market, market region/season, and time/transport mode-specific idiosyncrasies that influence market size variations due to the emergence of new consumers.

Four recent articles addressing topics connected to this paper should be mentioned. Dobruszkes & Vandermotten (2022) analyze the factors influencing air traffic in domestic and international markets, comparing national and sub-national levels. They conclude that sub-national units offer more accurate insights, as national-level data can obscure important regional differences, especially in large countries. Chen et al. (2020) present a method to examine the link between air traffic volume and macroeconomic development in Taiwan, distinguishing the most critical among 32 macroeconomic factors. Defining unmet demand as the passengers unable to fly due to supply and demand constraints, Carmona-Benítez & Nieto (2023) estimate unmet demand using machine learning algorithms. Their model incorporates socioeconomic variables at both community zone and airport levels to forecast air travel market size in origin-destination markets. Wang et al. (2019) develop a model to estimate air passenger travel, focusing on both internal driving forces and social influence factors. They simultaneously account for dynamic personal behaviors, the influence of fellow passengers, and the impact of similar passengers.

3. Application

3.1. Economic mobility evolution

According to the World Bank (2012), the share of the middle class in Brazil's population increased from 15% in the 1980s to 33% in the early 2010s, contributing for more than 40% of the overall population expansion in Latin America over that period. Oliven & Pinheiro-Machado (2012) show that the share of Class C households in total national income in Brazil in 2012 was 46%, against 44% of Classes A and B, "which have traditionally prevailed in the Brazilian economy." They claim that the increased social inclusion had a ripple effect during the period, booming consumption and introducing Class C into the air travel market. Similarly, the World Bank (2012) notes that the emergent middle class increasingly used airlines. To reach the emerging middle classes' vast potential demand during the economic boom in the early 2010s, some airlines concentrated their sales efforts in promoting air ticket payment by installments and creating non-digital ticket sales kiosks at popular spots, such as supermarkets, São Paulo subway stations, Rio de Janeiro's Central do Brasil railroad carrier, and even a new travel agency in the Rocinha slum. Fig. 1 shows Brazil's main socioeconomic trends for the years 2010, 2014, and 2018.

Brazil's GDP per capita varied significantly over that period. While it rose 5.7% (from US$ 8,700 to US$ 9,200) between 2010 and 2014, boosting the consumer disposable income, the economic recession in the mid-2010s reduced GDP per capita by 6.5% from 2014 to 2018, when it reached US$ 8,600. The inequality-adjusted income indicator (SEN) provides further insight into the evolution of income inequality in Brazil:[3] after increasing by 7.5% between 2010 and 2014, jumping from US$ 4,000 to US$ 4,300, it regressed in 2018 to the previous decade's values, partially due to the decline in the Gini Index (x 100) from 53.7 to 53.3.

The proportion of income held by the bottom 80% of the population in Brazil increased from 37.5% in 2001 to a peak of 43.3% in 2015, falling thereafter to 42.1% in 2019. Growth was particularly strong for the poorest (bottom 40%) and middle-class (middle 40%) segments, whose shares in total income rose over the period 2001–2015 from 8.5% to 11.3% and from 29% to 31.7%, respectively. Meanwhile, the proportion of income held by the top 20% declined from 62.5% to 56.7%. The mid-2010s recession would resume the worsening path in income inequality.

Financial inclusion is critical to grasp the economic environment of new consumer segments. As Fig. 1 shows, from 2001 to 2019, Brazil's real interest rate fell from 45.6% p.a. to 31.9% p.a., while the availability of credit increased from 29% of GDP to 62.8%, after peaking at 66.8% in 2015. The combination of reduced interest rates and increased credit availability contributed to boosting household consumption. Nonetheless, the financial market conditions, particularly the prevailing high interest rates, continue to impair low- and middle-income consumers. Concerning labor market, the unemployment rate fell from 12.5% in 2001 to a historic low of 7.2% in 2013. With the onset of the recession in 2015, the rate went up again, reaching 12.0% in 2019. Digital inclusion has benefited from significant advances in communication technologies, which have led to relatively high market penetration: the number of cell phones peaked at 125.6 per 100 people in 2015; the number of fixed broadband subscriptions reached 15.5 per 100 people in 2019; and access to the internet improved significantly over

[3] This indicator, proposed by Sen (1976), is calculated by multiplying the per capita income to the complement of the Gini index of income inequality.

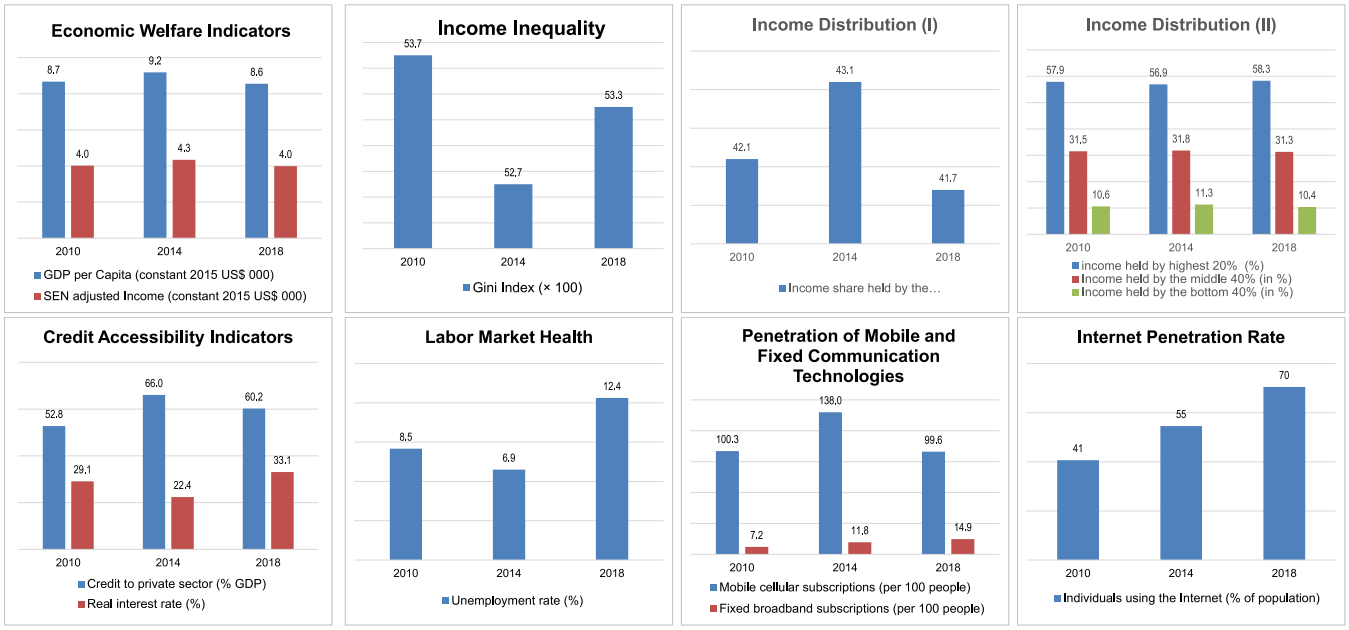

Fig. 1. Socioeconomic trends and temporal patterns–Brazil (2001–2019).
Sources: World Bank, International Monetary Fund (IMF), Brazilian Institute of Geography and Statistics (IBGE), Institute of Applied Economic Research (IPEA), and authors' calculations. "SEN" stands for the welfare indicator developed by Sen (1976). Due to missing data, the 2010 value in the Income Distribution (II) chart was interpolated using the values from the adjacent years, 2009 and 2011.

the years, with 81% of the population having internet access in 2020. These changes have opened up new opportunities for individuals and businesses alike, allowing them to communicate, access information, and conduct transactions more swiftly, effectively, and efficiently.

3.2. Travel markets dynamics

Table 1 presents the evolution of the domestic travel markets and some indicators of social, financial, and digital inclusion in Brazil. The number of airline passengers increased from 28.0 million in 2000 to 93.6 million in 2018—a growth of more than 234%. By contrast, the size of the intercity bus market shrank from 70.9 million to 42.8 million in the same period. More importantly, the size of the travel market expanded from 98.9 million in 2000 to 135.4 million in 2018, implying a growth of 37%, with the number of trips per capita recording 64.9 per 100 people. A great part of such growth was driven by falling airline prices (-60.3%) and increased overall income, as measured by the GDP per capita (57.0%).

Fig. 2 illustrates the evolution of air and bus transport networks from 2010 to 2018. It is worth noting that airlines have greater network coverage across the country's territory, particularly reaching remote and difficult-to-access areas in the northwest part of Brazil, which covers a significant portion of the Amazon region. As seen in the map, this area shows very limited road transport coverage, evidenced by the sparse green lines, while air transport, represented by red lines, connects these areas in a less sparse manner.

As the dynamics of the travel market is affected by intramodal competition, particularly in air transportation, any discussion about the potential shift between air and bus transport should take into account the impact of low-cost carriers (LCCs).[4] Unfortunately, our dataset does not include observations related to the entry of a rapidly growing emerging LCC that could attract new consumer segments through aggressive pricing strategies. The phase most associated with market penetration pricing strategies for the only two LCCs that operated during our sample period, Gol and Azul, was during their newcomer periods—2001 for Gol and 2008 for Azul. This is well before the start of our sample period, June 2012, when these companies had become larger incumbents. Nevertheless, we believe that, for certain periods of favorable macroeconomic conditions, the pricing power of these LCCs along with the full-service carrier Latam Airlines may have contributed to attracting new passengers.

3.3. Conceptual framework

Fig. 3 represents our conceptual framework for analyzing the effect of economic mobility (EM) on medium- and long-distance travel market between cities. The diagram synthetizes our modeling strategy, facilitating the development of the empirical modeling and the necessary adjustments to the theoretical demand framework. It also shows that an economy's overall socioeconomic conditions are the EM driving forces. Evidently, part of the factors that constitute the socioeconomic conditions is external to the transportation sector, as the overall economic growth, public policies, levels of income inequalities and social inclusion, and the extent of financial and digital inclusion. An environment favorable to medium and low-income consumer segments' upward EM creates a potential demand for travel. Nonetheless, these consumers' effective market inclusion largely depends on carriers' commercial policies and, conversely, the intensity of market development activities increases with carriers' perceived demand potential. The effectiveness of the firms' market development policies promotes inclusion. Market factors related to market development and dependent on the competition status can also attract novice consumers, such as market segmentation, revenue management, travel financing, social media marketing, and customer loyalty, among others.

Market inclusion means that previously underserved individuals (prospects) can access travel services. However, the distinction between a prospect, who shows interest in purchasing but does not yet complete a transaction, and a consumer who is still considering a purchase is blurring. After becoming a member of the market, a customer can be categorized as an "inside good" if s/he has made an effective purchase of a particular travel product, such as air travel or bus travel. If a customer

[4] We thank the anonymous reviewer for bringing this point to our attention.

Table 1
Travel markets indicators–Brazil (2000–2018).

Indicator	(i)	(ii)	(iii)	(iv)	% Change			
	2000	2010	2014	2018	(ii)/(i)	(iii)/(ii)	(iv)/(iii)	(iv)/(i)
Airline passengers (million)	28.0	70.0	95.8	93.6	*149.7%*	*37.0%*	*-2.3%*	*234.1%*
Intercity bus passengers (million)	70.9	46.7	42.8	41.8	*-34.1%*	*-8.4%*	*-2.3%*	*-41.0%*
Total intercity passengers (million)	98.9	116.7	138.6	135.4	*17.9%*	*18.8%*	*-2.3%*	*36.9%*
Passengers per population (× 100)	57.0	59.9	68.7	64.9	*5.0%*	*14.8%*	*-5.5%*	*13.9%*
Bus yield (BRL × 100)	15.5	21.6	20.6	21.0	*38.8%*	*-4.3%*	*1.8%*	*35.3%*
Airline yield (BRL × 100)	88.9	48.9	43.9	35.3	*-45.0%*	*-10.1%*	*-19.7%*	*-60.3%*

Sources: National Civil Aviation Agency (ANAC), National Land Transportation Agency (ANTT), Brazilian Institute of Geography and Statistics (IBGE), and authors' calculations.

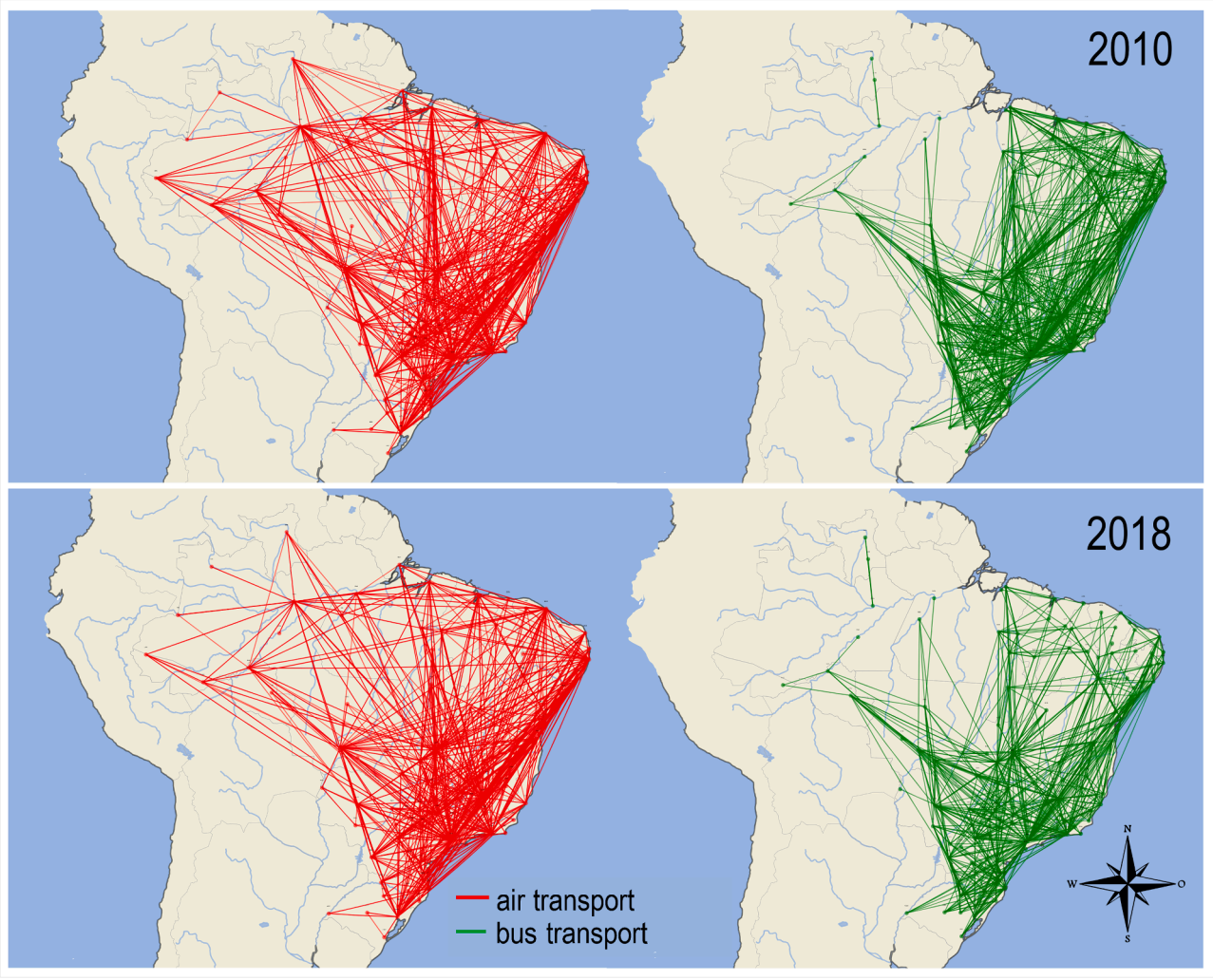

Fig. 2. Air and bus transport network evolution patterns: 2010 vs. 2018.
Sources: Each line corresponds to a pair of mesoregions served by the respective transport mode. The coordinates of the city with the highest population in each mesoregion are used to define the endpoints of the line. National Civil Aviation Agency (ANAC), National Land Transportation Agency (ANTT), Brazilian Institute of Geography and Statistics (IBGE), and authors' calculations.

who has not yet made a purchase is still in the process of searching for the best travel products, s/he is considered an "outside good". A customer who is in the period between two trips is also classified as an "outside good". Finally, once the participation of new consumers from medium- and low-income population segments is materialized, transport equity grows, purely induced by market factors. However, other factors can yield the same or even a stronger effect, such as deregulation, sustainable investment in infrastructure, and transportation policies.

Our primary goal lies in investigating the relationship between upward EM and the size of inside good. Specifically, we examine how improved socioeconomic conditions, by promoting EM, can increase demand for air and bus travel and affect carriers' prices and revenues. Nonetheless, instead of directly tackling the impact of EM on transport equity, we assume its beneficial effects. Thus, we recommend that future studies address this limitation by providing a thorough and comprehensive understanding of the links between socioeconomic conditions and transport equity.

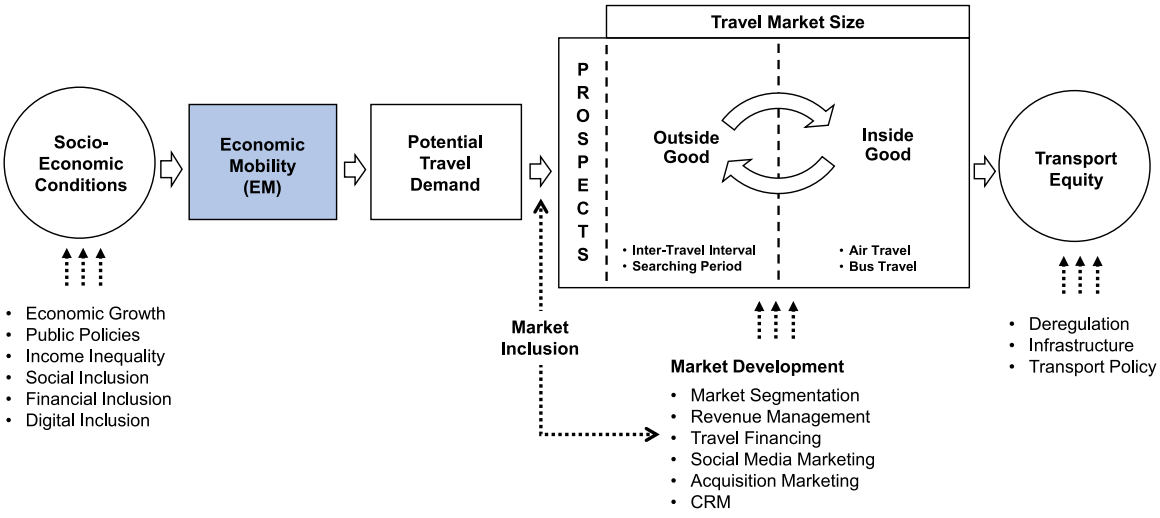

Fig. 3. Conceptual framework.

3.4. Research methodology

As Fig. 4 illustrates, our analysis of the medium and long-distance travel markets in Brazil relies on a logit model with aggregate data. This model incorporates alternatives of both internal and external goods, as described in Berry (1994) and Huang & Rojas (2013, 2014), and is further discussed in Sections 4.1.1 and 4.1.2. Here, consumers' available internal alternatives are air and bus travel modes. Our approach considers the observed characteristics of the transport modes and a modeling structure that incorporates unobserved characteristics as well as total market shifters. To construct a comprehensive dataset on the transportation industry in Brazil, we utilize data from the two travel modes as well as data for socioeconomic, financial, and digital inclusion, including data on travel quantities, prices, consumer loyalty, economic conditions, and EM. Sections 4.1.3 and 4.1.4 thoroughly discuss the dataset's elaboration and contents.

Fig. 4 depicts our two main quantitative approaches, namely an econometric and a machine learning modeling. For the econometric modeling, we employ two configurations. The first one is the Instrumental Variables LASSO (IV-LASSO) model (Section 4), developed by Belloni et al. (2012, 2014a, b), which involves regression procedures with multiple controls and instrumental variables to deactivate irrelevant variables and shrink the model. After estimating the regressions, the best specifications are identified using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Furthermore, we perform an analysis of the high-dimension sparse (HDS) stability of the parameters, examining the behavior of the coefficients when a larger number of high-dimensional controls are utilized.

The second approach, Quantile Regression (Section 5), investigates the parameters' behavior across the dependent variables' quartiles for each transportation mode, providing a distributional curve analysis. This approach helps to identify specific demand and pricing behaviors that may indicate revenue generation opportunities. We develop the machine learning models based on the approach proposed by Wolpert (1992) and Breiman (1996) and the implementation of Ahrens et al. (2023). These models consist of "Level-0" learners, including Neural Networks (Multilayer Perceptron), Gradient Boosting, Random Forest, and Support Vector Regression, and a "Level-1" learner, which is a Stacking Regression. To train the models, we perform hyperparameter tuning based on cross-validation criteria for each Level-0 learner. The Stacking Regression is then used to determine the optimal weights of the predictions generated by each Level-0 learner to estimate the Level-1 learner model.

The econometric models yield insights into the demand for travel and transportation operators' pricing strategies, as discussed in Sections 4.1.6 and 4.2. We conduct a joint interpretation of the estimated quantity and price relationships to provide a comprehensive understanding of the results. The machine learning model, in turn, focuses on revenue prediction analysis, as presented in Section 6. We carry out hyperparameter tuning procedures and employ cross-validation to compare the performance of the final models with alternative testing models. The section's key goal is to explore the chosen set of features to enhance the prediction of the transportation industry revenues over the sample period, with a view to analyzing carriers' prospective sales and profits.

Finally, we combine the results from both modeling approaches to implement an event study (Section 7) of the dynamic behavior of both the predicted and observed total revenues over the economic downturn in Brazil from 2014 onwards. The event study is designed to generate predictive analytics and insights on consumers' behavior that may provide policy suggestions for businesses operating in emerging economies in periods of crises. Specifically, the study takes stock of how the travel markets functioned during the economic slowdown to propose improvements in the transportation operators' commercial management in such challenging periods.

4. The IV-LASSO regression

4.1. Demand framework

4.1.1. Logit with aggregate data

We employ the baseline logit with aggregate data framework proposed by Berry (1994) and discussed in the works of Huang & Rojas (2013, 2014). We first consider a two-way panel dataset with data from a single market of products over time and a simplified consumers' utility function. The next section extends the analysis to a multi-way panel dataset that includes information from multiple markets, products, and time periods, using a more realistic, yet still simple, utility function. The developed demand modeling aims to be aligned with the conceptual model discussed in Section 3.3 and illustrated in the diagram of Fig. 3. Table 2 provides a list of the symbols and notations utilized in the framework.

Consider a market with N differentiated products wherein the consumer utility function incorporates a single observed characteristic of the product (x), its price (p), and a single unobserved characteristic (ξ). In each period t, an individual i chooses product j that maximizes her utility u_{ijt} among the N alternatives, plus the outside good, denoted by $j = 0$. We therefore have the following utility maximization problem:

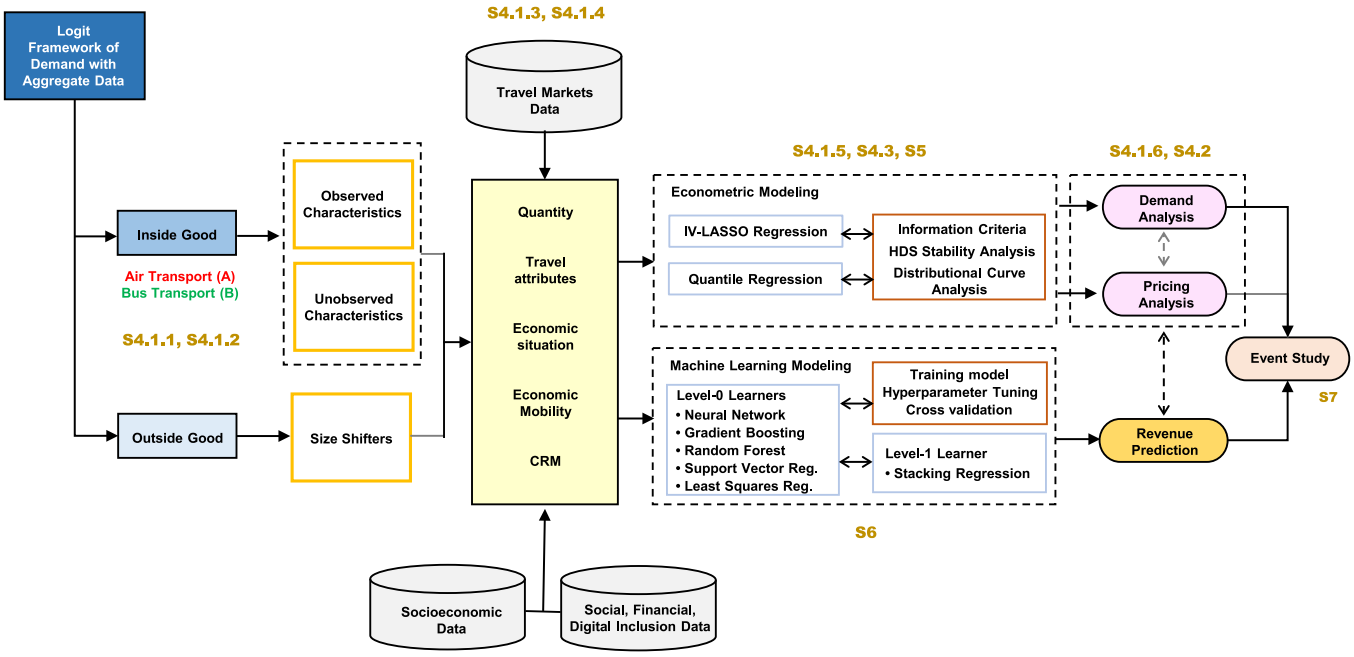

Fig. 4. Methodology schematic.

Table 2
Notations.

Symbol	Description	Symbol	Description
A	subscript for air transport	t	subscript for time period
B	subscript for bus transport	u	Utility
H	outside good shifters vector	w	subscript for season
i	subscript for individual	x	single observable characteristic
j	subscript for product	X	vector of observable characteristics
k	subscript for market	α	price parameter
K	total number of markets	β	mean utility function parameter/ vector
M	total market size	γ	parameter vector
N	number of inside goods	δ	mean utility level/vector/ function
p	price	δ_0	outside good's mean utility level
q	quantity	ϵ	component error
q_0	outside good quantity	ζ	unobserved idiosyncrasies
r	subscript for market region	ν	unobserved individual idiosyncrasies
s	market share	ξ	unobserved characteristic(s)
s_0	outside good market share	π	choice probability/function

$$\max_j u_{ijt} = \delta_{jt} + \nu_{ijt}; j = 0, 1, \ldots, N, \tag{1}$$

where δ_{jt} is the mean utility level of j, and ν_{ijt} is the logit error term. Assume δ_{jt} as below:

$$\delta_{jt} = \beta_0 + \beta_1 x_{jt} - \alpha p_{jt} + \xi_{jt}. \tag{2}$$

Consumer i's probability of choosing alternative j is given by

$$\pi_{jt} = \pi(\delta_t) = \frac{e^{\delta_{jt}}}{\sum_{l=0}^{N} e^{\delta_{lt}}}; j = 0, 1, \ldots, N. \tag{3}$$

where δ_t is a vector of the N mean utilities. In the typical model setup, researchers with aggregate data employ the following procedures:

(i) assume that ξ_{jt} is correlated with p_{jt};

(ii) replace π_{jt} by $s_{jt} = q_{jt}/M_t$, where s_{jt} and q_{jt} are, respectively, the market share and the quantity of product j, $j = 0, 1, \ldots, N$, and $M_t = \sum_{l=0}^{N} q_{lt}$ is the total market size; and

(iii) normalize the utility of the outside good to zero ($\delta_{0t} = 0$).

Our framework employs only the first two procedures, adding to the literature initiated by Berry (1994) by providing a less restrictive approach, albeit with its own limitations, which are discussed later. We therefore have:

$$\frac{e^{\delta_{jt}}}{\sum_{l=0}^{N} e^{\delta_{lt}}} = s_{jt}. \tag{4}$$

Dividing each side by the share of the outside good (s_{0t}) and after some manipulations, we reach:

$$\ln\left(q_{jt}\right) - \ln(q_{0t}) = \beta_0 + \beta_1 x_{jt} - \alpha p_{jt} - \delta_{0t} + \xi_{jt}, \tag{5}$$

where q_{0t} is the size of the outside good, which is an unobserved quantity. Huang & Rojas (2014) point out that researchers commonly set q_{0t} by a "reasonable guess," which is indirectly obtained by assuming a guess for the market size M_t. We interpret M_t as firms' total market potential to gain market share. Building upon the authors, we employ an alternative approach to account for the unobserved factors associated with market potential. More specifically, we move $\ln(q_{0t})$ to the right-hand side of Eq. (5) and, unlike those authors, introduce the following subtle yet crucial modification (5):

$$\begin{aligned} &\ln\left(q_{jt}\right) = \beta_0 + \beta_1 x_{jt} - \alpha p_{jt} + \epsilon_{jt}, \\ &\text{where } \epsilon_{jt} = \ln(q_{0t}) - \delta_{0t} + \xi_{jt}. \end{aligned} \tag{6}$$

Our approach innovates by incorporating $\ln(q_{0t})$ into ϵ_{jt}, thus simplifying the specification of the systematic component of Eq. (6), while posing a challenge in modeling the unobserved portion of ϵ_{jt}. The following subsection presents our proposed estimation strategy, which addresses this difficulty and accounts for the potentially non-zero δ_{0t} term.

4.1.2. Outside good and economic mobility

To account for the possible non-random variations of ϵ_{jt} in Eq. (6), we

follow Huang & Rojas (2014) and introduce a third dimension into the model, the index for market k, along with indexes j for product and t for time period. Considering the multi-way panel data framework and incorporating alternative-specific parameters and multiple product characteristics into the utility function, Eq. (6) is expressed as follows:

$$\ln\left(q_{jkt}\right) = \beta_j^{'}\mathrm{X}_{jkt} - \alpha_j p_{jkt} + \epsilon_{jkt}, \tag{7}$$

where $\epsilon_{jkt} = \ln(q_{0kt}) - \delta_{0kt} + \xi_{jkt}$, X_{jkt} is a vector of observed characteristics of product j.

Concerning δ_{0kt}, we cannot take for granted a zero-mean utility for the outside good because consumers may value the no-purchase option differently in different markets within the panel data. However, if the panel data structure has not a very long time series, it may be reasonable to assume the utility constant over time, even though being market-related. Our contribution to the literature is to set a model that assumes a market-specific, time-constant mean utility of the outside good δ_{0k} in substitution for δ_{0kt} in (7). Note that this structure is flexible for allowing local nullity possibilities in the data ($\delta_{0k} = 0$, for a subset of $k = 1, 2,\ldots,K$) and, in addition, also leaves the global nullity possibility ($\delta_{0k} = 0, \forall k$) as a special case. Finally, this approach has the benefit of accounting for such an unobserved term (δ_{0k}) if we use market-fixed effects.

We propose the following empirical counterpart for the error term in (7):

$$\epsilon_{jkt} = \gamma_j^{'}\mathrm{H}_{kt} + \gamma_{0k} + \xi\left(\zeta_{jk}, \zeta_{rw}, \zeta_{jt}\right) + \zeta_{jkt}, \tag{8}$$

where H_{kt} is a vector of outside good shifters and $\xi\left(\zeta_{jk}, \zeta_{rw}, \zeta_{jt}\right)$ is an additive function targeted at modeling ξ_{jkt} in a more detailed way; ζ_{jk}, ζ_{rw}, and ζ_{jt} are controls for, respectively, product/market, market region/season, and product/time-specific idiosyncrasies that may affect the utility portion associated with the unobserved characteristics of j; ζ_{jkt} is a random term; $\gamma_j^{'}$ is a parameters vector; and γ_{0k} is a parameter associated with the outside good mean utility term δ_{0k}, discussed above.

To model H_{kt}, we use time-invariant and time-varying covariates and controls associated with firms' existing market potentials. As discussed in 3.3, the outside good size has two driving forces. One is the *existing customers'* unrealized demand, i.e., those clients who do not consume any alternative at time t but have done so before. These customers may be in an inter-travel interval period or in a process of searching for fare or travel conditions. The other is *new consumers*' unrealized demand, i.e. those novice clients who have never purchased any of the N inside products before and are now considering them. These consumers may be attracted to the market due to improved EM conditions, representing a segment of consumers that can be seen as the demand stemming from market inclusion, which should be the focus of firms' sales prospecting efforts to boost market penetration and ultimately generate new revenue.

While the available dataset does not allow us to distinguish between unrealized demand from existing consumers and new consumers or prospects, instead of explicitly separating the unobserved components associated with each group, we estimate the impact of a set of EM indicators that affect the overall size of the travel market by either expanding or contracting the outside good. This approach is in line with Hofer et al. (2018), as it makes it possible to investigate the hypothesis that demand is also driven by additional socioeconomic factors, besides the conventional indicators related to income and employment. Moreover, we examine how these factors can contribute to predicting the total revenue generated.

We use subscript j to index vector $\gamma_j^{'}$ in Eq. (8), wherein lies a limitation of our study. Some proxies we use to control for the impact of EM on demand may also reflect the utility associated with unobservable characteristics of the product, ξ_{jkt}. This is because new customers' tastes and preferences may be different from those of existing customers, and each customer segment's preferences may differ for each product, adding another source of endogeneity in the model, which we try to mitigate with control variables associated with ζ_{jk}, ζ_{ws}, and ζ_{jt}. Furthermore, we assume that the demand for each product is solely influenced by changes in the size of the outside good. For instance, airlines may employ more effective acquisition strategies than bus carriers, leading to a higher likelihood that newly acquired customers choose air travel. We therefore replace H_{kt} with H_{jkt} in (8). However, this approach prevents us from isolating the factors affecting the outside good size from those associated with the unobservable utility of j. Plugging (8) into (7), we reach the following final model:

$$\ln\left(q_{jkt}\right) = \beta_j^{'}\mathrm{X}_{jkt} - \alpha_j p_{jkt} + \gamma_j^{'}\mathrm{H}_{jkt} + \gamma_{0k} + \xi\left(\zeta_{jk}, \zeta_{rw}, \zeta_{jt}\right) + \zeta_{jkt}, \tag{9}$$

which allows us to investigate EM impacts on customers demand in a market comprising both observable (inside) goods and an unobservable (outside) good, whose size is influenced by the inclusion of new consumers.

4.1.3. Data and sources

We construct a comprehensive dataset to explore the influence of EM on prospective sales and revenues of carriers operating in Brazil's medium- and long-distance travel markets. It comprises a panel of directional domestic city-pair markets, with monthly observations spanning from June 2012 to December 2018.[5] We group multiple airports belonging to the same catchment area and only consider routes with more than 6 observations. The sample comprises 587 markets served by airlines and 339 markets served by buses, with a total sample of 34,589 observations for air travels and 18,209 for bus travels. We collect most of the data from publicly available sources, including the National Civil Aviation Agency (ANAC), the National Secretariat of Civil Aviation (SAC), the National Land Transport Agency (ANTT), the Brazilian Institute of Geography and Statistics (IBGE), the National Agency for Petroleum, Natural Gas and Biofuels (ANP), and the Ministry of Labor and Employment. To estimate the EM indicators, we rely on data from the IBGE, while for the financial inclusion indicators we use the Central Bank of Brazil's ESTBAN and Credit Information System (SCR) databases. For the digital inclusion indicators, we employ data from the National Telecommunication Agency (ANATEL) and Google Trends. Below, we explain how our database is integrated into the demand framework of subsections 4.1.1 and 4.1.2.

4.1.4. Model specification

To present the empirical specification of Eq. (9), adjusted to fit our travel market dataset, we begin by denoting air and bus transportation by A and B, respectively, the transport mode by subscript j, the directional city-pair market by subscript k, and the time period by subscript t. All variables are expressed in logarithmic form. Concerning the regressand term in (9), q_{jkt}, we use the empirical counterpart PAX_{jkt}:[6]

$$q_{jkt} = \left\{\mathrm{PAX}_{\mathrm{jkt}}\right\},\ j = \{A, B\}, \tag{10}$$

where PAX_{jkt} is the total number of passengers for each transport mode.[7] As for the vector of observable characteristics X, we rely on the following specification:

$$\mathrm{X}_{jkt} = \left\{\mathrm{FREQ}_{jkt}, \mathrm{TIME}_{jkt}, \mathrm{P}_{jkt}\right\},\ j = \{A, B\}, \tag{11}$$

[5] Observations for June and July 2014 are missing.

[6] We use capitalized letters for empirical model variables or features and their associated vectors to distinguish them from theoretical and conceptual variables.

[7] Sources: ANAC's Airfares Microdata (number of total tickets sold); ANTT's (number of revenue passengers).

where $FREQ_{jkt}$ denotes the total number of frequencies,[8] $TIME_{jkt}$ is the mean travel time in minutes,[9] and P_{jkt}, the mean price.[10] To address multicollinearity issues in the demand equation of bus transportation, we define these variables as the difference between the bus and the air transport service attributes. Our specification for the vector of outside good shifters is as follows:

$$H_{jkt} = \{ECONOMY_{kt}, EM_{kt}, CRM_{jkt}\},\ j = \{A, B\}, \tag{12}$$

where $ECONOMY_{kt}$ represents the national and local economic factors that impact on consumers' propensity to travel, which in turn determine the effectiveness of firms' market penetration strategies based on economic conditions; EM_{kt} includes EM factors that promote market inclusion through sales prospecting; and CRM_{jkt} stands for carriers' customer relationship management efforts, which can have an impact on customer experience and satisfaction, potentially attracting customers away from the rival transport mode.

Using (12), we define the external good shifter vectors that may also be associated with the unobservable characteristics of the transport modes. The proxies used in each vector are:

$$ECONOMY_{kt} = \{INCOME_{kt}, VACAT_{kt}, UNEMPL_{kt}, CONFID_{kt}\},\ j = \{A, B\}, \tag{13}$$

where $INCOME_{kt}$ is a proxy for consumers' mean income, which is equal to the geometric mean of the origin and destination cities' per capita gross value added;[11] $VACAT_{kt}$ is a proxy for the tourism intensity of the traffic on the directional city-pair, being equal to the proportion of the amount of wages from tourism-related jobs to the respective local economy's gross domestic product;[12] $UNEMPL_{kt}$ is an index of unemployment;[13] and $CONFID_{kt}$ is a composite index of confidence in national and local economic conditions, resulting from multiplying an index of industrial manufacturers' expectations on the overall economy and the year-over-year growth in local GDP.[14]

$$EM_{kt} = \{SEN_{kt}, CREDIT_{kt}, DEFAULT_{kt}, SEARCH_{kt}\},\ j = \{A, B\}, \tag{14}$$

where SEN is Sen's inequality-adjusted income indicator (Sen, 1976; Foster & Sen, 1997), which we calculate as the complement of the local Gini index times the mean wage in the local economy (rather than the GDP per capita, as originally conceived by Sen, to avoid multicollinearity with INCOME);[15] CREDIT is a proxy for financial inclusion, being the average amount of bank credit available per operation for clients with a total income of up to 10 minimum wages;[16] DEFAULT is an indicator of the debt burden among low and middle income households, defined as the percentage of both outstanding credit operations with overdue installments of more than 90 days and credits from high-risk, restructured, or defaulted credit operations in total bank credit received by customers with total income of up to 10 minimum wages;[17] and SEARCH is a composite measure designed to capture two important dimensions of EM in the context of travel: digital inclusion and consumer service awareness. SEARCH is computed by combining three different indicators, all measured at the origin mesoregion of each travel route:[18] the number of cell phones in the region, the proportion of low and middle-income families, and a web-based travel search interest metric, the Google Trends' Search Index for the topic "viagem" ("travel" in Portuguese).[19] The first two variables reflect the extent of digital inclusion in the region, while the third variable measures the level of consumer awareness related to travel services and attributes. To obtain the final figure, we extract the six-month moving average of the resulting variable. This accounts for the typical time window when consumers search for fares before making actual travel purchases, and consumer service awareness may emerge and intensify.

$$CRM_{jkt} = \{RETENT_{jkt}, SWITCH_{jkt}\},\ j = \{A, B\}, \tag{15}$$

where $RETENT_{jkt}$ is a proxy for customers' potential retention, which refers to the likelihood that a customer will continue to use the same travel services for future trips. This is a crucial metric for travel companies, as retaining customers can translate into increased revenue, higher growth rates, and long-term profitability. The potential factors that drive customer retention in the travel industry encompass a range of elements, such as the quality of the travel experience, the level of customer service, availability and appeal of loyalty programs. We set consumer retention as the maximum number of passengers who traveled via a given mode of transport in the past 24 months, as Eq. (16) shows.[20] If positive, its coefficient indicates how much past consumption has the power to generate future trips: the higher it is, the greater the firms' average retention capacity.

[8] Sources: ANAC and ANTT.

[9] Sources: ANAC, ANTT, and National Civil Aviation Secretary (SAC).

[10] Measured in inflation-adjusted local currency values. Source: ANAC's Airfares Microdata and IBGE. The mean price of buses is a proxy extracted from the IBGE's city-level consumer price index. For the medium and small cities, we utilize the countrywide bus price inflation. For each route, we calculate the geometric mean price index between origin and destination cities. We allow for inter-route variation by multiplying the time-varying index by the ANTT's published price reference of 2019 for each route.

[11] Measured in inflation-adjusted local currency values. The gross value added is equal to GDP plus subsidies on products, minus taxes on products. We compute a monthly interpolation of the original yearly data. We also aggregate the data at the mesoregion (grouping nearby cities) level. Source: IBGE's Gross Domestic Product of Municipalities.

[12] Multiplied by one thousand. We aggregate this figure to the mesoregion (grouping nearby cities) level. Source: Annual Relation of Social Information (RAIS) of the Ministry of Labor and Employment and IBGE's GDP of Municipalities.

[13] It equals the inverse of the Ministry of Labor and Employment's formal employment index (multiplied by 100).

[14] We extract the geometric mean of the indexes for the endpoint cities and aggregate this figure to the mesoregion (grouping nearby cities) level. Source: IBGE's GDP of Municipalities and National Confederation of Industry.

[15] It equals the geometric mean between the origin and destination cities (inflation-adjusted local currency values, in logarithm). We compute a monthly interpolation of the original yearly data. For the Gini index figures, we use interpolation and extrapolation. We also aggregate it to the mesoregion (grouping nearby cities) level. Sources: Annual Relation of Social Information (RAIS) of the Ministry of Labor and Employment, and IBGE's GDP of Municipalities and IBGE/Demographic Censuses 2000, 2010 for the Gini coefficient of per capita household income by Municipality, Brazil.

[16] Inflation-adjusted local currency values. The data source is the Central Bank of Brazil, with state/monthly disaggregated data. To calculate this variable, we use the mean values from the origin and destination states for each route.

[17] The data source is the Central Bank of Brazil, with state/monthly disaggregated data. To calculate this variable, we average the values from both the origin and destination states for each route.

[18] Sources: ANATEL, IBGE's Household Budget Survey, and Google Trends, respectively.

[19] Measured at the state level. For a visual representation of the data source, please refer to the trends.google.com.br/trends/explore?date=2012-01-01%202021-12-31&q=%2Fm%2F014dsx&geo=BR-SP link.

[20] The sources for these variables are the same as those for the models' dependent variables: ANAC's Airfares Microdata (number of total tickets sold) and ANTT's (number of revenue passengers).

$$\text{RETENT}_{jkt} = \max\left(q_{jk,t-h}, q_{jk,t-h+1}, \ldots, q_{jk,t}\right),\ j = \{A, B\} \text{ and } h = 1, 2, \ldots, 24 \tag{16}$$

SWITCH_{jkt} is a measure of the firms' ability to attract passengers from the competing transport mode. Peelen & Beltman (2013) argue that striking a balance between acquiring new customers and nurturing existing relationships is crucial for any corporation. This variable therefore captures the potential of acquiring customers who are likely to switch modes. The ability to attract new customers and retain existing ones can result in increased revenue and market share for firms. Our proxy is calculated by considering the maximum number of passengers who used the rival mode of transportation in the past 24 months and may potentially switch to the current mode.

$$\text{SWITCH}_{jkt} = \max\left(q_{-jk,t-h}, q_{-jk,t-h+1}, \ldots, q_{-jk,t}\right),\ -j = \{A, B\} \text{ and } h = 1, 2, \ldots, 24. \tag{17}$$

where $-j$ refers to mode j's rival transport mode. Table 3 provides a concise overview of the variables used in our empirical framework. The Appendix presents the descriptive statistics and a correlation analysis.

4.1.5. Estimation strategy

I. Component error specification

Eq. (18) presents the empirical demand specification, which incorporates the settings proposed in (9) to (17):

$$\begin{aligned}\text{PAX}_{jkt} = {} & \beta_{1j}\text{FREQ}_{jkt} + \beta_{2j}\text{TIME}_{jkt} + \beta_{3j}\text{P}_{jkt} + \beta_{4j}\text{INCOME}_{kt} + \beta_{5j}\text{VACAT}_{kt} \\ & + \beta_{6j}\text{UNEMPL}_{kt} + \beta_{7j}\text{CONFID}_{kt} + \beta_{8j}\text{SEN}_{kt} + \beta_{9j}\text{CREDIT}_{kt} \\ & + \beta_{10j}\text{DEFAULT}_{kt} + \beta_{11j}\text{SEARCH}_{kt} + \beta_{12j}\text{RETENT}_{jkt} \\ & + \beta_{13j}\text{SWITCH}_{jkt} + \gamma_{0k} + \xi_{jkt},\ j = \{A, B\},\end{aligned} \tag{18}$$

where $\xi_{jkt} = \xi(\zeta_{jk}, \zeta_{rw}, \zeta_{jt}) + \zeta_{jkt}$, as discussed before. The component error in this model is represented by $\gamma_{0k} + \xi_{jkt}$, which accounts for various factors that may influence demand but are not observed by the econometrician. These factors are associated with the transport mode, city-pair, endpoint regions, seasons, and period. We define the empirical counterpart of ξ_{jkt} as:

$$\xi_{jkt} = \text{E}\left(\text{CPAIR}_{jk},\ \text{REG}^{o}_{jkm}, \text{REG}^{d}_{jkm}, \text{PER}_{jy}\right) + \omega_{jkt} \tag{19}$$

Table 3
Brief description of the model variables.

Vector	Variable	Description
$\mathbf{q_{jkt}}$	PAX_{jkt}	Total passengers (mode-specific) (ln)
$\mathbf{X_{jkt}}$	FREQ_{jkt}	Total frequency (mode-specific) (ln)
	TIME_{jkt}	Mean travel time (mode-specific) (ln)
	P_{jkt}	Mean fare (mode-specific) (ln)
$\mathbf{ECONOMY_{kt}}$	INCOME_{kt}	Per capita gross domestic product (O, D geom. mean) (ln)
	VACAT_{kt}	Tourism intensity (O, D geom. mean) (ln)
	UNEMPL_{kt}	Unemployment index (O, D geom. mean) (ln)
	CONFID_{kt}	Economic confidence (O, D geom. mean) (% var. × ln)
$\mathbf{EM_{kt}}$	SEN_{kt}	Amartya Sen's adjusted income (O, D geom. mean) (ln)
	CREDIT_{kt}	Available credit in the economy (O, D geom. mean) (ln)
	DEFAULT_{kt}	Household indebtedness (O, D geom. mean) (ln)
	SEARCH_{kt}	Internet search intensity (O, D geom. mean) (ln)
$\mathbf{CRM_{jkt}}$	RETENT_{jkt}	Retention potential of current customers (ln)
	SWITCH_{jkt}	Acquisition potential of customers who are likely to switch modes (ln)

where E(.) is an additive function of control variables; CPAIR_{jk} is the fixed effect of route k designed to account for transport mode/market-specific and time invariant idiosyncrasies, which controls for both γ_{0k} (market-specific mean utility of the outside good) and ζ_{jk} (the effects of route distance on demand, among other possible factors); REG^{o}_{jkm} and REG^{d}_{jkm} are dummy variables to account for the unobserved effects associated with the region of the endpoint city (origin o or destination d) and the month of the year (m) on route k;[21] PER_{jy} are mode-specific year dummies to account for time-varying changes; and ω_{jkt} accounts for the mean zero random disturbances.

Our approach to modeling the component error has the limitation of imposing additivity in E(.), implying a strong linear restriction on the disturbances of Eq. (18). Moreover, the use of the annual dummies PER_{jy} to model the unobserved time-varying error factors only captures medium- to long-term effects, while short-term factors require continuous variables. Although necessary to mitigate multicollinearity in the model and thus to improve model identification, these restrictions may introduce omitted variable biases into the results, which therefore should be interpreted with caution.

The sample period, 2012 to 2018, includes Brazil's economic downturn, which had a substantial impact on economic activity level and very probably on the travel markets. Neglecting the huge effects associated with this period would lead these factors to be subsumed in the error term, thereby further confounding estimation. We address this fact by including annual dummies, as discussed above, and by adding an interaction term of $\text{P}_{jk,t}$ with a dummy variable indicative of this period, EDT_t, which assumes the value of 1 from the third quarter of 2014, when the recession began.[22] The interaction term $\text{P}_{jk,t} \times \text{EDT}_t$ captures changes in the price elasticity of demand during the economic downturn. We emphasize that the methodology presented in Section 3.3 allows us to investigate the impact of this period on the studied markets, which we thoroughly present in Section 7.

Eq. (20) shows the full empirical demand specification:

$$\begin{aligned}\text{PAX}_{jkt} = {} & \beta_{1j}\text{FREQ}_{jkt} + \beta_{2j}\text{TIME}_{jkt} + \beta_{3j}\text{P}_{jkt} + \widetilde{\beta}_{3j}\text{P}_{jkt} \times \text{EDT}_t \\ & + \beta_{4j}\text{INCOME}_{kt} + \beta_{5j}\text{VACAT}_{kt} + \beta_{6j}\text{UNEMPL}_{kt} + \beta_{7j}\text{CONFID}_{kt} \\ & + \beta_{8j}\text{SEN}_{kt} + \beta_{9j}\text{CREDIT}_{kt} + \beta_{10j}\text{DEFAULT}_{kt} + \beta_{11j}\text{SEARCH}_{kt} \\ & + \beta_{12j}\text{RETENT}_{jkt} + \beta_{13j}\text{SWITCH}_{jkt} + + \varphi_{1j}\text{CPAIR}_{jk} + \varphi_{2j}\text{REG}^{o}_{jkm} \\ & + \varphi_{3j}\text{REG}^{d}_{jkm} + \varphi_{4j}\text{PER}_{jy} + \omega_{jkt},\ j \\ = {} & \{\text{A}, \text{B}\},\end{aligned} \tag{20}$$

where $\widetilde{\beta}_{3j}$ is an interaction coefficient that indicates the possible moderation effect of EDT on price, and φ_{1j}, φ_{2j}, φ_{3j}, φ_{4j} are unknown parameters.

II. Endogeneity, heteroscedasticity, and autocorrelation

Our framework considers the possibility of endogeneity in prices, which means that the covariance between the price of a product (P_{jkt}) and the error term (ξ_{jkt}) may be different from zero. In addition, since the price of a product may be correlated with the variables in the E(.) function, particularly with CPAIR_{jk}, we use a fixed instead of a random effect approach. Moreover, the price of a product may be correlated not only with unobservable product characteristics but also with the unobservable determinants of the potential demand that may be associated with ω_{jkt}. Thus, there is likely to be a non-null correlation between P_{jkt} and ω_{jkt}. To account for these issues, we employ the instrumental vari-

[21] Brazil is geographically divided into five regions: North, South, North-East, South-East, and Middle-West. By incorporating these alternative-specific controls, we consider the seasonality impact on demand, which may vary by region and month.

[22] See Section 7 for further details.

ables (IV) regression method, which absorbs multiple levels of fixed effects, as described in Correia (2017).

More specifically, we rely on a set of exogenous costs and operational shifters for both transport modes as IVs. These IVs include variables such as fuel costs and taxes, fuel efficiency, maintenance, and congestion. Additionally, we use lagged observations of price, frequency, and time (up to six lags), as well as Hausman-type instruments. For instance, we instrument P_{jkt} with the price and operational characteristics of other routes. To create further instruments, we interact all the proposed instruments with EDT_t.

We conduct a battery of diagnostic tests for the residuals to pinpoint possible estimation problems: for heteroscedasticity, the Pagan-Hall, White/Koenker, and Breusch-Pagan/Godfrey/Cook-Weisberg tests; and for autocorrelation, the Cumby-Huizinga test. As these tests reveal both heteroscedasticity and autocorrelation in ω_{jkt}, we address them with the Newey-West procedure, which allows for consistent estimation of the standard errors by adjusting the variance-covariance matrix of the parameter estimates using a kernel function that captures the dependence structure of the residuals over time. We utilize the Two-Step Generalized Method of Moments (2SGMM) estimation, as it can handle both the endogeneity and the heteroscedasticity of the error terms. Moreover, GMM can provide consistent estimates even when the distribution of the error terms is unknown or the errors are correlated over time.

III. High-dimension sparse (HDS) approach

Grasping how economic mobility affects travel markets within a country is thorny because of the various unobserved factors—such as changing consumer preferences, evolving tourism and migration patterns, and varying conditions of business travel, among others. Hence, we opt for a high-dimensional approach whereby we rely on a broad subset of control variables to account for such multiplicity of confounders. Apart from the fixed effects related to the 587 air and 339 bus routes, as expressed by (20), our approach also estimates 252 parameters of control variables for the equations of both the transport modes, which we refer to as "nuisance parameters," since they are model parameters that hold no direct interest for the researcher, but nonetheless are still necessary in the model to account for possible sources of unobserved variability. By including them, we attempt to ensure proper statistical inference of the parameters of interest.

Besides the fixed effects and controls, our model contains up to 200 instrumental variables. To guarantee that we select only the most appropriate control and instrumental variables for each estimation and to reduce the dimensionality of the model, we employ the high dimensional sparse (HDS) regression models developed by Belloni et al. (2012), Belloni et al. (2014a, b), and Chernozhukov et al. (2015). These models build upon the Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996). We utilize the estimation procedure known as IV-LASSO (Instrumental Variables LASSO), employing the theory-driven penalization criteria for rigorous LASSO (Belloni et al., 2012), as described in Ahrens et al. (2020). The final step of our approach uses the Two-Step Feasible GMM procedure as the post-LASSO estimator (2SGMM-IVLASSO).[23]

4.1.6. Estimation results

Table 4 presents the estimation results of the empirical model of demand (PAX), with air transport (A) in Columns (1)–(3) and bus transport (B) in Columns (4)–(6). The specifications between columns differ only in the number of variables considered. Columns (1) and (4) show the results of the models that include only variables from the X and ECONOMY vectors. Columns (2) and (5) include the variables of EM: SEN, CREDIT, DEFAULT, and SEARCH. Finally, Columns (3) and (6) add the CRM variables, i.e. RETENT and SWITCH. We focus on the more comprehensive model specifications of Columns (3) and (6) because they present the lowest AIC and BIC statistics. To simplify the exposition, we henceforth remove the subscripts from the variables.

The results are consistent with our expectations. The variable FREQ has a positive estimated coefficient, while the coefficients for P, P × EDT, and TIME are negative across the columns. However, TIME in most cases is not statistically significant for bus transport. Overall, the price estimates for both modes of transport are reasonably close. Nonetheless, the absolute coefficients for buses are from 7.2% to 12.8% higher than those for air transport. The estimated negative coefficient for P × EDT suggests that the demand became more elastic to price changes during the economic downturn period, especially for bus transport.

The negative coefficient for INCOME might indicate that bus transport is an inferior good, but these estimates fail to be statistically significant, implying that it may be an income-inelastic or income-neutral good. By contrast, INCOME estimates for air transport are consistently positive and statistically significant. While VACAT is a statistically significant demand driver only for airlines, the negative coefficient of this variable in Column (6) suggests that buses may face a competitive disadvantage in tourism markets. Moreover, UNEMPL is generally not statistically significant, whereas CONFID is statistically significant only for airlines.

The coefficients for most of the EM variables are statistically significant, which are in line with our argument about the implications of the changing socioeconomic conditions over the sample period. However, while these indicators always present statistical significance for air transport, some of them are not for bus operators' demand, suggesting that there is room for demand management improvements in the bus industry. For example, our results indicate that SEN has a significant positive effect on airline demand, which holds even after controlling for the effect of INCOME, while it is positive but non-significant for buses, further supporting our claim that bus transportation is an income-inelastic good, even when using inequality-adjusted income measures.

The findings for financial inclusion measures are insightful. Coefficients for the variable CREDIT are consistently statistically significant and positive for airlines and negative for buses, indicating that greater credit availability for low and middle-income consumers drives intermodal substitution, as bus customers are more likely to switch to airlines. DEFAULT is consistently negative and statistically significant for both modes of transport, suggesting that indebtedness of economically disadvantaged households constrains consumption and contributes to leaving potential demand in the travel market in Brazil. Furthermore, it is worth remarking that the magnitude of the estimated coefficients for bus transport is more than twice that for air transport.

The variable SEARCH is statistically significant only for airlines, indicating a positive effect in capturing online customers' attention, particularly neophyte consumers who have recently purchased cell phones. This finding points to airlines' probable competitive advantage vis-à-vis bus carriers in certain online marketing strategies, such as Search Engine Optimization (SEO) and social media marketing. Nonetheless, the recent emergence of online bus ticket startups, such as Buser and FlixBus, may come to mitigate bus carriers' disadvantage.

Comparing the estimated parameters in the model specifications of Columns (3) and (6), air transport tends to benefit more from upward EM than bus transport in several aspects. The results indicate greater effects for SEN and SEARCH on the demand for air transport. In addition, CREDIT has a transport mode switching effect favoring air transport. Finally, the DEFAULT coefficient for air transport is two thirds lower, suggesting that higher indebtedness in low- and middle-income consumer segments has less impact on the demand for airlines. The analysis seems to indicate that airlines are proficient in generating demand when socioeconomic conditions are favorable. Furthermore, the evidence suggests that airlines are better at market segmentation to attract non-

[23] We conducted an HD stability analysis to assess the robustness of our empirical results. It involved adding a substantial set of high-dimensional controls to the model and comparing the estimated coefficients with the original specifications. The results demonstrated that most of the estimates remained stable, confirming the robustness of the findings.

Table 4
Estimation results–IV-LASSO regression: travel demand (PAX).

Variable	{Y, j} = {PAX, A}			{Y, j} = {PAX, B}		
	(1)	(2)	(3)	(4)	(5)	(6)
FREQ	0.1706***	0.1598***	0.1679***	0.1690***	0.1706***	0.1803***
TIME	-0.1603***	-0.1705***	-0.1578***	-0.0711	-0.0123	-0.0771
P	-1.1201***	-1.1505***	-0.9765***	-1.2002***	-1.2982**	-1.0860***
P × EDT	-0.0048***	-0.0076***	-0.0063***	-0.1923***	-0.2303***	-0.1900***
INCOME	0.7561***	0.4004***	0.1795**	-0.0722	-0.1739	-0.3077
VACAT	0.6803***	0.4839***	0.3945***	-0.2099	-0.2687	-0.4134***
UNEMPL	-0.4283**	-0.1105	-0.0950	-0.2873	-0.0409	0.6932
CONFID	0.0516*	0.1632***	0.1519***	0.2358*	-0.3071	-0.1056
SEN		0.4957***	0.7044***		0.0013	0.2975
CREDIT		0.2007***	0.1666***		-0.6873***	-0.5642***
DEFAULT		-0.2474***	-0.2238***		-0.7507***	-0.5875***
SEARCH		0.2467***	0.2173***		0.0304	0.0524
RETENT			0.2950***			0.2017***
SWITCH			<lasso>			0.2384***
Estimator	IV-LASSO	IV-LASSO	IV-LASSO	IV-LASSO	IV-LASSO	IV-LASSO
Post-Lasso	2SGMM	2SGMM	2SGMM	2SGMM	2SGMM	2SGMM
Fixed Effects	yes	yes	yes	Yes	yes	yes
AIC Stat.	-7,204	-7,584	-12,201	25,491	26,482	23,843
BIC Stat.	-7,086	-7,432	-12,040	25,600	26,622	23,999
Adj R2 Stat.	0.9750	0.9753	0.9784	0.9256	0.9215	0.9321
RMSE Stat.	0.2202	0.2189	0.2048	0.4929	0.5065	0.4710
Underid. Stat	2107.6149	1946.6069	1731.9285	43.0993	19.3855	127.8954
Weak Id. Stat	124.3536	128.5539	106.9260	21.4736	9.6441	42.6700
No. Controls	99/120	99/120	99/120	92/120	91/120	91/120
No. IVs	18/176	16/176	17/176	2/32	2/32	3/32
No. Obs.	34,589	34,589	34,589	18,209	18,209	18,209

Notes: Estimation results produced by the instrumental variables, post-double-selection LASSO-based methodology of Belloni et al. (2012, 2014a, b), with fixed effects (IV-LASSO). Post-LASSO estimation is performed with a Two-Step Feasible Generalized Method of Moments (2SGMM), fixed-effects, procedure with standard errors robust to heteroskedasticity and autocorrelation. LASSO penalty loadings account for the clustering of city-pairs. Region-specific seasonality control variables estimates omitted. FREQ, TIME, P, P × EDT, INCOME, VACAT, and UNEMPL not penalized by LASSO. Endogenous variables: P and P × ECDOWN. "<lasso>" denotes that the LASSO procedure discarded the variable. Blank cells indicate that the respective variable is not used. Adjusted R2 and RMSE statistics extracted from an equivalent Least Squares Dummy Variables (LSDV) model. P-value representations: *** p<0.01, ** p<0.05, * p<0.10. "A" denotes air transport, and "B" denotes bus transport.

traditional travel market clients, such as younger consumers from medium- or low-income families, owing to the convenience, speed, and perceived status associated to air travel. By contrast, bus companies appear to be failing in attracting this key market segment to demand growth during economic prosperity. Yet, we also find evidence that an economic crisis that lowers INCOME and SEN and restricts credit availability is likely to affect more airlines than bus operators.

Regarding the proxy for potential consumer retention, RETENT, both air transport and bus companies show some ability to retain customers, as its coefficients are always positive and statistically significant. However, bus companies exhibit significantly lower retention capabilities, with an estimated coefficient (0.2017) 32% lower than those for airlines (0.2950). Interestingly, the IV-LASSO analysis shows that the proxy for attracting consumers, SWITCH, is statistically significant and positive for bus companies whereas it is discarded for airlines. The findings have significant implications for carrier market positioning: while bus companies should focus on enhancing their ability to retain customers by establishing stronger relationships and offering loyalty incentives, airlines can benefit from conquest strategies—such as market research, social media monitoring, and targeted marketing campaigns to attract customers from rivals.

4.2. Insights from a pricing model

Using the methodology outlined in subsection 3.3, we develop a complementary empirical price model to provide insights into carriers' competitive behavior, without the pretension of being exhaustive.[24] The empirical pricing modeling of air and bus transport relies on the same IV-LASSO estimator and empirical configuration used in the demand modeling. However, since price is now the dependent variable, we remove it and its interaction with EDT from the right-hand side of the equation and add two other covariates: FUELP, which is a proxy for the price of jet A1 fuel in the case of airlines and diesel in the case of buses,[25] and HHI, the Herfindahl-Hirschman index of city-pair concentration (multiplied by 100), considering the revenue passengers of both the air and bus carriers.[26] We also interact HHI with EDT to examine the moderating impacts of economic downturns on the relationship between market concentration and pricing. Eq. (21) below shows the full specification of the pricing model:[27]

$$\begin{aligned} P_{jk,t} = {} & \gamma_{1j}\mathrm{FUELP}_{jk,t} + \beta^{p}_{1j}\mathrm{FREQ}_{jkt} + \beta^{p}_{2j}\mathrm{TIME}_{jkt} + \beta^{p}_{3j}\mathrm{HHI}_{jkt} + \widetilde{\beta}^{p}_{3j}\mathrm{HHI}_{jkt} \\ & \times \mathrm{EDT}_{t} + \beta^{p}_{4j}\mathrm{INCOME}_{kt} + \beta^{p}_{5j}\mathrm{VACAT}_{kt} + \beta^{p}_{6j}\mathrm{UNEMPL}_{kt} \\ & + \beta^{p}_{7j}\mathrm{CONFID}_{kt} + \beta^{p}_{8j}\mathrm{SEN}_{kt} + \beta^{p}_{9j}\mathrm{CREDIT}_{kt} + \beta^{p}_{10j}\mathrm{DEFAULT}_{kt} \\ & + \beta^{p}_{11j}\mathrm{SEARCH}_{kt} + \beta^{p}_{12j}\mathrm{RETENT}_{jkt} + \beta^{p}_{13j}\mathrm{SWITCH}_{jkt} + \varphi^{p}_{1j}\mathrm{CPAIR}_{jk} \\ & + \varphi^{p}_{2j}\mathrm{REG}^{o}_{jkm} + \varphi^{p}_{3j}\mathrm{REG}^{d}_{jkm} + \varphi^{p}_{4j}\mathrm{PER}_{jy} + \nu_{jk,t},\ j \\ = {} & \{A, B\}, \end{aligned} \tag{21}$$

where the symbols γ, β^{p}, and φ^{p} denote unknown parameters, with $\widetilde{\beta}^{p}_{3j}$

[24] The reduced-form of our price model is a limitation.

[25] These variables are available at the region/daily level in ANP database. We then compute a mean region/monthly basis and extract the geometric mean of the origin and destination cities' regions. Their values in local currency are deflated.

[26] We calculate the index with data from ANAC's Airfares Microdata (number of total tickets sold) and ANTT's (number of revenue passengers).

[27] For the sake of simplicity, we exclude controls for variables with indexes other than j, k, m, and y from the equation because, although they are part of the model, their effect is null in this case.

being the interaction coefficient and $v_{jk,t}$ the random term. We treat HHI_{jkt} and $HHI_{jkt} \times EDT_t$ as endogenous variables and instrument them with a set of exogenous demand shifters. Additionally, to instrument the market concentration of a particular route, we augment the IV set with Hausman-type instruments that rely on other routes' market concentration levels (all instruments are in logs). We have 148 instrumental variables to be penalized by the IV-LASSO method.

Table 5 presents the estimated results for air transport in Columns (1)-(3) and bus transport in Columns (5)-(8). Once again, completeness and superior information criterion performance, as indicated by the AIC and BIC statistics at the bottom of the table, lead us to choose the full specifications, shown in Columns (3) and (6).

As expected, FUELP is positively associated with prices. It is unclear whether the magnitude of the fuel-to-price transmission is higher for any of the considered transport modes. FREQ has a negative estimated coefficient, which is probably due to the presence of traffic density economies, but the effect is significant only for air transport. The positive coefficient of TIME for airlines, unlike buses, may be associated to the higher cost caused by flight delays and inefficient trajectories stemming from air traffic management problems.

Regarding market concentration, the positive estimated coefficients for HHI suggest the presence of pricing power affecting ticket prices in air travel, while evidence of such possibility is limited for buses, as not all coefficients are statistically significant at the 5% level. Moreover, the negative and statistically significant interaction of HHI with EDT is only observed for air transport, possibly indicating airlines' declining pricing power during the recession period. It is worth reminding that our analysis is not based on a structural model of competition, thus implying that this finding is not enough to confirm market power and should be interpreted as evidence of pricing power instead.

The economy-related variables—INCOME, VACAT, and UNEMPL—are statistically significant only for airlines. Interestingly, the results for CONFID suggest that greater confidence in the economy's performance is associated with mean prices higher for airlines and lower for bus, pointing to airlines' demand rationing behavior and bus companies' attempt to avoid losing customers to their rivals. Apparently, airlines do not fear price-sensitive customers to switch from buses, so their revenue management systems appear to restrict additional demand.

Amid the EM indicators, the results for SEN, with statistically significant estimates only for airlines, are consistent with those for INCOME and also with the evidence discussed in the demand model suggesting that the demand elasticity for bus transport may be income-neutral. For CREDIT and DEFAULT, the findings confirm financial inclusion as a pivotal market indicator for all carriers, providing evidence that higher indebtedness of low and middle-income families leads to price reduction, probably to mitigate demand loss. Also consistent with those for CONFID, higher credit availability tends to increase air transport prices and decrease bus prices, indicating opposing price pressures that may result from potential transportation mode switching. SEARCH does not affect pricing for any transportation modes, meaning that, ceteris paribus, pricing does not change according to the average search intensity, implying that pricing can improve, given the increasing digital inclusion and consumer awareness.

Finally, results are intriguing for the CRM variables. In the equation for air transport SWITCH becomes inactive and RETENT is not discarded by IV-LASSO, while in the bus transport equation the opposite happens—RETENT is dropped and SWITCH remains. In both cases, the variables kept in the equation are negatively correlated and highly statistically significant, suggesting that carriers may use the additional

Table 5
Estimation results–IV-LASSO regression: mean travel price (P).

Variable	{Y, j} = {P, A}			{Y, j} = {P, B}		
	(1)	(2)	(3)	(4)	(5)	(6)
FUELP	0.1314***	0.0831***	0.0573**	0.0425***	0.1000***	0.0999***
FREQ	-0.0984***	-0.0972***	-0.0861***	-0.0009**	-0.0002	-0.0004
TIME	0.0906**	0.0913**	0.0677*	-0.0069*	-0.0044	-0.0035
HHI	0.6248***	0.6529***	0.6120***	0.0533***	0.0245*	0.0248*
HHI × EDT	-0.0129***	-0.0174***	-0.0181***	-0.0010**	0.0000	0.0000
INCOME	0.2205**	0.1681*	0.2659***	0.0150	0.0175	0.0235*
VACAT	0.2312***	0.2282***	0.2920***	-0.0027	-0.0107	-0.0061
UNEMPL	-0.4028**	-0.4816**	-0.4174**	-0.0032	0.0018	0.0166
CONFID	0.1163***	0.2460***	0.2318***	0.0314***	-0.0220***	-0.0234***
SEN		-0.3698**	-0.4158**		0.0352	0.0269
CREDIT		0.2403***	0.2434***		-0.1006***	-0.1010***
DEFAULT		-0.1578***	-0.1615***		-0.0240***	-0.0263***
SEARCH		-0.0050	-0.0020		-0.0055	-0.0056
RETENT			-0.1637***			<lasso>
SWITCH			<lasso>			-0.0069***
Estimator	IV-LASSO	IV-LASSO	IV-LASSO	IV-LASSO	IV-LASSO	IV-LASSO
Post-Lasso	2SGMM	2SGMM	2SGMM	2SGMM	2SGMM	2SGMM
Fixed Effects	yes	Yes	yes	yes	yes	yes
AIC Stat.	-678	-254	-2,007	-79,802	-82,664	-82,704
BIC Stat.	-552	-94	-1,838	-79,685	-82,515	-82,548
Adj R2 Stat.	0.6273	0.6207	0.6299	0.9969	0.9973	0.9973
RMSE Stat.	0.2412	0.2434	0.2404	0.0274	0.0253	0.0253
Underid. Stat	572.4706	438.6469	436.8177	522.9307	205.1625	206.6121
Weak Id. Stat	25.2157	19.2419	19.1600	21.3874	8.5819	8.6428
No. Controls	99/120	99/120	99/120	95/120	95/120	95/120
No. IVs	23/74	23/74	23/74	25/74	24/74	24/74
No. Obs.	34,589	34,589	34,589	18,209	18,209	18,209

Notes: Estimation results produced by the instrumental variables, post-double-selection LASSO-based methodology of Belloni et al. (2012, 2014a, b), with fixed effects (IV-LASSO). Post-LASSO estimation is performed with a Two-Step Feasible Generalized Method of Moments (2SGMM), fixed-effects, procedure with standard errors robust to heteroskedasticity and autocorrelation. LASSO penalty loadings account for the clustering of city-pairs. Region-specific seasonality control variables estimates omitted. FUELP, FREQ, TIME, P, P × EDT, INCOME, VACAT, and UNEMPL not penalized by LASSO. Endogenous variables: P and P × ECDOWN. "<lasso>" denotes that the LASSO procedure discarded the variable. Blank cells indicate that the respective variable is not used. Adjusted R2 and RMSE statistics extracted from an equivalent Least Squares Dummy Variables (LSDV) model. P-value representations: *** p<0.01, ** p<0.05, * p<0.10. "A" denotes air transport, and "B" denotes bus transport.

demand, generated by retention by airlines and switching by buses, to increase price competitiveness.

4.3. HD stability analysis

To assess the robustness of the empirical results, we conducted an HD stability analysis, employing 252 control parameters in the above estimations for both modes of transportation and adding an extra set of 1532 control variables as nuisance parameters in the models. The additional control variables consist of various local economy indicators, such as cash and bank deposit values, the number of bank branches, and the amount of real estate credit available. These variables are measured both in absolute terms and normalized by population or GDP. We include in addition a nominal retail index and the human development index, a composite measure of long and healthy life, education, and standard of living.[28] Once more, we rely on the IV-LASSO approach to identify and select the most significant control variables.

Most of the estimates remain relatively stable after adding the high-dimensional controls, with a few exceptions, such as INCOME, UNEMPL, and SEN, which overall show different estimates from the main models. Concerning bus transport pricing, for example, DEFAULT, SEN, INCOME, and UNEMPL exhibit coefficients at odds with their previous estimates. The results suggest that these variables may be influenced by a range of unobserved idiosyncrasies that vary over time in a complex way that is not captured by the REG and PER in Eqs. (20) and (21). All other estimated coefficients remain unchanged, indicating that a significant portion of our estimation results is robust to the inclusion of additional high-dimensional controls.

5. Quantile regression

The IV-LASSO procedure provides flexible estimations of the causal effects of the proposed regressors on travel demand and price while taking control variables into account. However, this approach assumes symmetric and uniform conditional distribution across all quantiles, estimating only the mean of the response variable given the covariates. To address potential asymmetry and heterogeneity in the conditional distribution, we employ a quantile regression modeling approach to investigate the effects of EM and market inclusion on travel markets in Brazil. This extension of the IV-LASSO procedure allows us to explore how the effects of the regressors may vary across different quantiles, clarifying the relationship between EM and travel demand and price.

Our approach relies on quantile regressions with instrumental variables and fixed effects, covering percentiles in increments of 10%, from 10% to 90%. We use the full demand and price models of Section 4. The quantile regression procedure estimates the variance-covariance matrix of the estimators using bootstrap resampling and, to ensure the reliability of the results, calculating bootstrap standard errors with 250 replications. Fig. 5 displays the coefficient estimates for PAX and P across different percentiles. As a reference, the graphs also present the values of the corresponding ordinary least squares (OLS) estimates as a solid line, along with their confidence interval represented by dashed lines. These figures provide a clearer interpretation of the effects of our variables across different points of the distribution, suggesting valuable insights for corporate policy and carriers' decision-making concerning a more effective market penetration.

With respect to travel demand (PAX), only TIME, SEN, DEFAULT, and CONFID remain relatively stable across percentiles on the airline side. By contrast, FREQ, P, P × EDT, INCOME, VACAT, CREDIT, SEARCH, UNEMPL, and notably RETENT vary significantly. The results indicate airlines' higher consumer retention potential and higher price demand elasticity in denser transportation markets. Conversely, the estimated effect of the number of frequencies (FREQ) tends to decrease in these markets, suggesting diminishing returns in demand attraction. The EM vector variables show that higher demand reduces the effect of SEN and reinforces the effects of CREDIT and SEARCH, meaning that, as demand grows, the demand enhancing impact of higher income distribution tends to be substituted by other aspects of EM, such as financial and digital inclusion.

For bus transportation, the number of covariates that exhibit variability across the percentiles is higher than for airlines, with only P, UNEMPL, and CONFID being relatively stable. Again, FREQ, P, P × EDT, INCOME, VACAT, CREDIT, SEARCH, and RETENT vary considerably. SWITCH appears in this specification as not being inactivated by IV-LASSO, with its impact being stronger on bus demand in markets with medium to high demand and weaker in markets with low demand. Regarding EM variables, the variability pattern in SEN, CREDIT, DEFAULT, and SEARCH for buses is the opposite of that estimated for airlines. In bus markets with high demand, income distribution has a stronger impact, while the effects of CREDIT and SEARCH are weaker.

Regarding travel prices, the effect of fuel prices (FUELP) for both modes of transport tends to be more pronounced in high-demand markets. Although this result should be further investigated in future studies, it may be due to carriers' greater market power in high-demand markets, facilitating to pass a greater proportion of their costs to prices.[29]

6. Machine learning modeling

In the previous two sections, we examined the causal relationship between EM, market inclusion, transport demand, and pricing in Brazil's air and bus transport systems. This section employs predictive analytics to explore how EM factors that promote inclusion may assist carriers in their revenue prospecting activities. We undertake a meta-learning approach to predict carriers' total revenue (TREV).[30] The set of variables is the same used in the IV-LASSO and the quantile regression models, albeit in levels instead of logarithmic transformations. As a final step, we evaluate the effectiveness of the EM vector features in improving predictions.

Ahrens et al. (2023) claim that researchers seldom have *a priori* a clear view of the best suited machine learning models to carry out a new prediction or a classification task. They note that a common approach is to evaluate the performance of a set of machine learners using a hold-out partition of the data or cross-validation and subsequently select the machine learner that minimizes a chosen loss metric. Wolpert (1992) and Breiman (1996) propose an alternative approach, called "stacked generalization" (or just "stacking"), which relies on the idea that combining multiple learners into one final prediction may lead to superior performance than that for each individual learner. The theoretical approach of Van der Laan et al. (2007) shows that a super learner is a flexible prediction procedure that can have high performance on different data-generating distributions. They argue that researchers, instead of restricting this technique to a limited set of learners, should include any sensible model available to achieve the best performance, allowing for a diversity of base learners that ultimately can lead to that goal when combined into a meta-learner.

We rely on Stacking Regression, henceforth ST, to obtain TREV predictions. We consider the following machine learning models as base learners (Level-0 learners): Random Forest (RF), Gradient Boosting (GB), Neural Networks/Multilayer Perceptron (NN), Support Vector

[28] All these variables are calculated for the origin and destination of each route, with minimum, arithmetic mean, geometric mean, and maximum values extracted between the endpoints. The squared values are also computed as well as the first lag. Finally, the values are normalized between zero and 100, and computed in logarithm.

[29] Gayle & Lin (2021) focus on a similar topic relying on a structural model rather than a quantile regressions.

[30] We define total revenue as total passengers × mean price.

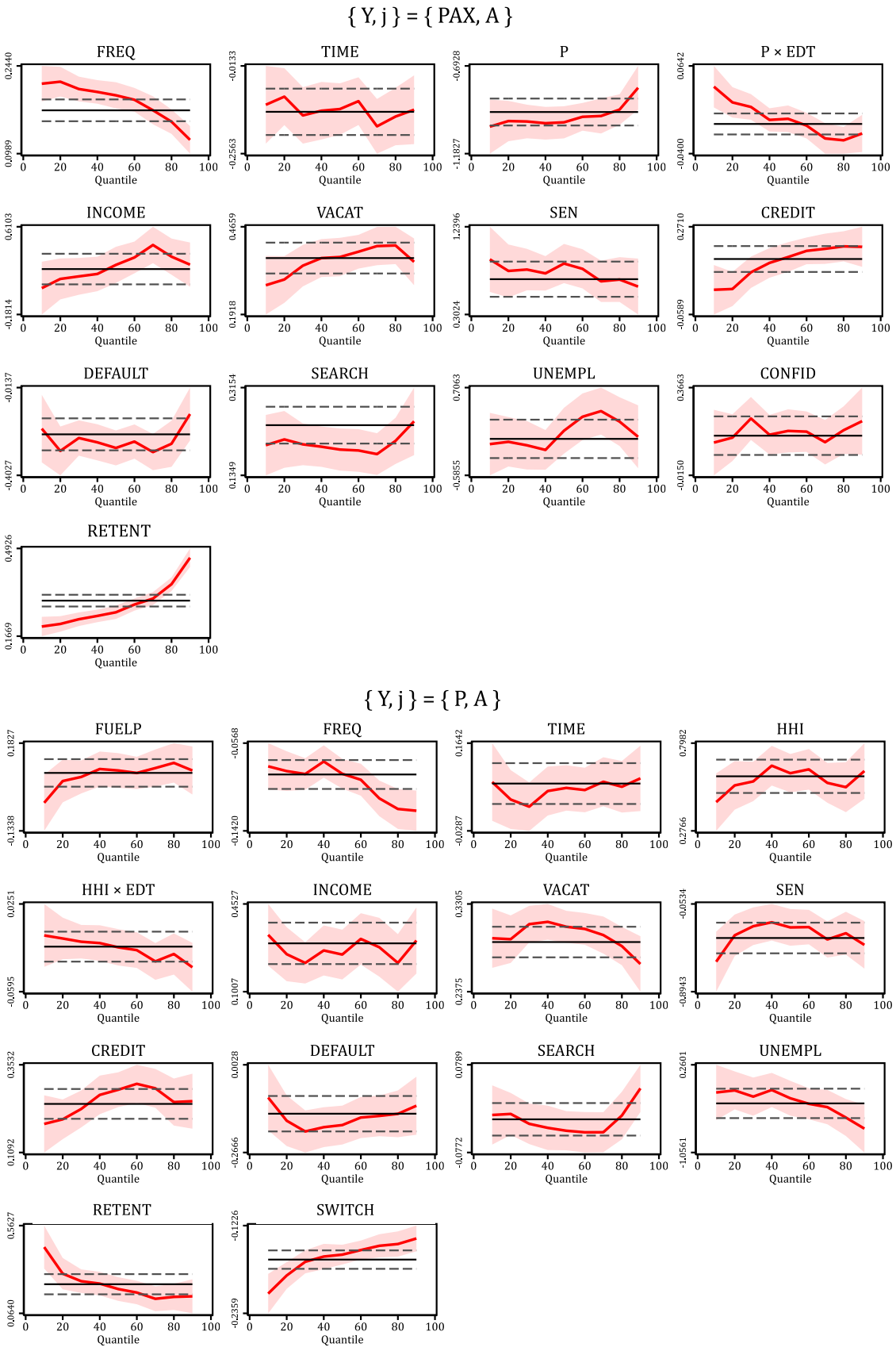

Fig. 5. (part I)-Estimation results-Quantile Regression: travel demand (PAX) and price (P): airlines (A).
Notes: Estimation results produced by an instrumental variable, fixed-effects, Quantile Regression, with same model specification as in Section 4. The red lines display the Quantile Regression results, with the red areas indicative of their confidence interval. The black lines display the Ordinary Least Squares (OLS) results, with the respective confidence interval demarcated with dotted lines. Endogenous variables: P and P × EDT (upper charts), and HHI and HHI × EDT (lower charts).

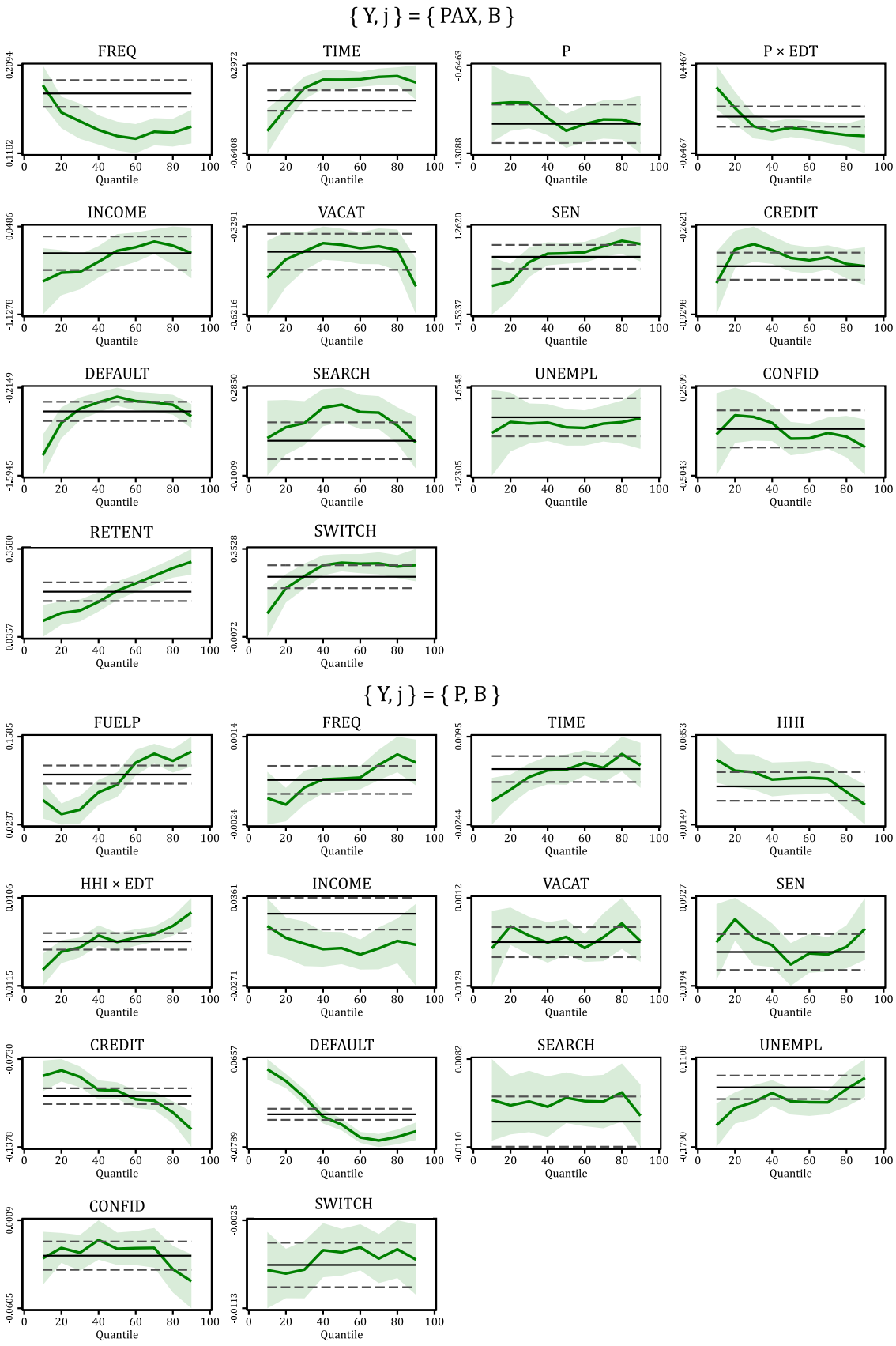

Fig. 5. (part II)–Estimation results–Quantile Regression: travel demand (PAX) and price (P): buses (B).
Notes: Estimation results produced by an instrumental variable, fixed-effects, Quantile Regression, with same model specification as in Section 4. The green lines display the Quantile Regression results, with the green areas indicative of their confidence interval. The black lines display the Ordinary Least Squares (OLS) results, with the respective confidence interval demarcated with dotted lines. Endogenous variables: P and P × EDT (upper charts), and HHI and HHI × EDT (lower charts).

Regression (SV), and Least Squares (LS).[31] ST serves as the meta-learner (Level-1 learner), which is constructed by taking a weighted average of the individual predictions, using a constrained least squares approach. More specifically, the model uses a non-negative linear squares estimator (NNLS) with the additional constraint of the sum of the coefficients being one.[32]

To train the Level-0 models, we conduct hyperparameter tuning by performing a random search of 100 trials for each model. As we use panel data, cross-validation is appropriate to determine the best performing learners because it can yield the loss of the temporal structure of each route's series. Instead, we employ a hold-out sample procedure for both model validation and testing. For validation, we train the models on the initial period of the series (July 2012 to December 2015). To prevent correlation due to temporal proximity from contaminating the hold-out predictions, we introduce a 6-month gap between the training and validation sets, which spans from July 2016 to June 2017. For model testing, we also use a 6-month gap and perform hold-out predictions for the year 2018. Throughout all procedures, we measure prediction performance using the Coefficient of Variation of the Root Mean Square Error (CVRMSE). In addition, we report the Mean Absolute Percentage Error (MAPE) in the analysis of the model's test performance.[33] As we study a variety of heterogeneous travel markets, we use a weighted version of these indicators, where the weights are based on the respective number of passengers. This allows us to assign higher weights to markets that contribute more to the total revenue of transportation operators from a system-wide standpoint.

To evaluate the robustness of the ST predictions, we consider two versions. The first version, ST1, uses weights estimated from the model with the lowest CVRMSE during the validation phase for each of the five Level-0 learners. As a robustness check for ST1, the second version, ST2, employs weights estimated from the second lowest CVRMSE model.

Table 6 exhibits the results of the CVRMSE and MAPE metrics for each machine learning model evaluated in the testing procedure. The weights of each ST model relative to the Level-0 learners are displayed on the left-hand side of the table. To provide a more detailed analysis, we present the statistics for the full sample, as well as for the "high season" and "low season" periods. In Brazil, these periods correspond to December through February (Summer) and July (mid-year school break), respectively. Additionally, the table presents the ranking of each model, highlighting the top performers for each sample. The results presented in Table 6 indicate that the stacking procedure performs exceptionally well for most of the specifications and transportation modes. Specifically, ST1 and ST2 consistently outperform all Level-0 learners for both CVRMSE and MAPE statistics, with ST1 being the highest-performing model in most rankings. For instance, regarding air transport and the full sample, ST1 and ST2 present the lowest estimated CVRMSE values of 20.0% and 20.5%, respectively, relative to the sample mean of TREV; for bus transport, the corresponding values are 20.8% and 20.7%. It is worth noting that these figures are the smallest among all models. However, it should observe that the stacking models do not consistently achieve the best holdout prediction performance. For example, ST1 and ST2 rank third and fourth in the smallest CVRMSE ranking for bus transport during the high season. These results suggest that although the stacking procedure overall performs well, analysts should explore other models to address specific market conditions affected by seasonality.

Furthermore, Table 6 indicates that, among the Level-0 learners, Random Forest (RF) and Gradient Boosting (GB) provide the best predictive performances for air transport and bus transport, respectively, often ranking third in the overall ranking. Conversely, Support Vector Regression (SV) and Least Squares (LS) generally present inferior performance for both CVRMSE and MAPE metrics.

Lastly, Table 6 reveals that, irrespective of the method and metric used, the predictive performance of the models for both transport models is relatively weaker for the high season. There is a notable increase in both the CVRMSE and MAPE statistics in high seasons vis-à-vis the low seasons. This could imply that a more diverse range of consumers with varying preferences may be present in the market, making predictions more challenging. To address this issue, carriers need to enhance their predictive analytic capabilities mainly for these high-demand, high-profit periods.

Fig. 6 focuses on the correlation between the selected EM variables (SEN, CREDIT, DEFAULT, and SEARCH) and the predicted revenues, which may provide insights into the relevance of each feature in the accuracy of predictions. We utilize the Pearson correlation metric to evaluate the potential for a linear relationship, but also compute the Spearman and Kendall metrics to account for possible non-linear relationships. We present measures for both the training and holdout (testing) sets. Furthermore, we examine the impact of omitting the ECONOMY features, including INCOME, VACAT, UNEMPL, and CONFID, which are commonly used in explanatory and predictive analyses. By comparing the results of these variables with those of the EM features, further insights into the relevance of market inclusion may be found.

Among the EM variables, SEARCH most frequently has the highest correlation with the predicted revenues, especially for air transport, indicating that consumers are increasingly sensitive to online promotions and offers and that revenue management strategies of firms should consider individuals' online behavior, both on their own websites and smartphone apps as well as on travel consolidators, metasearch sites, and social media. SEN has the second place for air transport, while for bus transport it shows a behavior indicative of overfitting, given that the correlation of this feature with the revenue predictions drops considerably when comparing the results applied to the training sample and the testing sample. Finally, the ECONOMY vector variables (INCOME, VACAT, UNEMPL, and CONFID) generally have similar relevance for both training and testing predictions of revenues compared to the EM variables. The results demonstrate that the assumption that the variables typically used in demand and pricing models would outperform the EM variables in revenue prediction is not always true, providing evidence that combining EM variables with the traditional variables in economic models of transportation improves revenue prediction. In sum, our results suggest that EM variables associated with market inclusion are relevant for revenue prediction models, but nonetheless feature-specific and transport-specific analyses cannot be neglected.

7. Event study

In this section, we carry out an event study to complement the explanatory and predictive analytics developed in Sections 4, 5, and 6 by investigating the dynamic behavior of quantities, price, and total revenues in the transport markets over a period of economic decline. We analyze the estimation results within an annual time window, starting from one year before the onset of the 2014 technical recession.[34] The technical recession ended in late 2016, resulting in an impressive cumulative GDP decline of 8.6%.[35] Barbosa, Souza, & Soares, 2020 point

[31] We adjust the hyperparameters of each of these models using a random search procedure.

[32] See Ahrens, Hansen & Schaffer (2022) for details on the Stacking Regression procedure and routine.

[33] See Adedeji et al. (2021) for a discussion of the CVRMSE and MAPE.

[34] Source: Getulio Vargas Foundation (FGV). See "*Brazil entered into a recession starting from the 2nd quarter of 2014, according to FGV*" (in Portuguese), available at g1.globo.com, August 4th, 2015.

[35] Source: Getulio Vargas Foundation (FGV). See "*The Brazilian recession ended in late 2016, says committee of FGV that studies economic cycles*" (in Portuguese), available at g1.globo.com, October 30, 2017.

Table 6
Testing performance of the machine learning models: total revenue (TREV).

Machine Learner				{Y, j} = {TREV, A}													
				Full Sample				High Season				Low Season					
				CVRMSE	Rank		MAPE	Rank		CVRMSE	Rank		MAPE	Rank		CVRMSE	Rank		MAPE	Rank	
Level 1	ST1			*20.0%*	1	◄	*15.5%*	1	◄	*22.8%*	1	◄	*18.2%*	1	◄	*18.5%*	1	◄	*14.2%*	1	◄
	ST2			*20.5%*	2	◄	*16.1%*	2	◄	*24.0%*	3		*18.7%*	2	◄	*18.7%*	2	◄	*14.9%*	2	◄
Level 0	*weights*	ST1	ST2																		
	NN	*9.2%*	*9.6%*	*22.5%*	4		*18.5%*	4		*23.3%*	2	◄	*20.9%*	4		*22.1%*	4		*17.4%*	4	
	SV	*7.2%*	*0.2%*	*44.4%*	7		*18.8%*	5		*29.5%*	6		*21.0%*	5		*49.8%*	7		*17.8%*	5	
	GB	*0.0%*	*0.0%*	*25.2%*	5		*27.9%*	6		*27.5%*	5		*31.7%*	6		*24.1%*	5		*26.2%*	6	
	RF	*83.6%*	*90.2%*	*22.0%*	3		*16.8%*	3		*25.4%*	4		*19.3%*	3		*20.1%*	3		*15.6%*	3	
	LS	*0.0%*	*0.0%*	*29.8%*	6		*30.3%*	7		*34.5%*	7		*36.3%*	7		*27.3%*	6		*27.5%*	7	
Mean	*(a) Level 1 mean*			*20.2%*			*15.8%*			*23.4%*			*18.5%*			*18.6%*			*14.5%*		
	(b) Level 0 mean			*28.7%*			*22.5%*			*28.0%*			*25.8%*			*28.7%*			*20.9%*		
	(a) - (b)			*-8.5%*			*-6.7%*			*-4.6%*			*-7.4%*			*-10.1%*			*-6.4%*		
Machine Learner				{Y, j} = {TREV, B}													
				Full Sample				High Season				Low Season					
				CVRMSE	Rank		MAPE	Rank		CVRMSE	Rank		MAPE	Rank		CVRMSE	Rank		MAPE	Rank	
Level 1	ST1			*20.8%*	2	◄	*16.1%*	1	◄	*26.4%*	3		*17.8%*	1	◄	*15.2%*	3		*14.9%*	1	◄
	ST2			*20.7%*	1	◄	*16.3%*	2	◄	*27.5%*	4		*18.0%*	2	◄	*13.4%*	2	◄	*14.9%*	2	◄
Level 0	*weights*	ST1	ST2																		
	NN	*1.8%*	*0.0%*	*25.8%*	5		*32.3%*	6		*25.8%*	1	◄	*30.7%*	6		*25.8%*	5		*33.5%*	6	
	SV	*0.0%*	*0.0%*	*41.9%*	6		*28.8%*	5		*41.1%*	7		*30.3%*	5		*42.6%*	6		*27.8%*	5	
	GB	*67.2%*	*41.0%*	*21.0%*	3		*16.7%*	3		*25.8%*	2	◄	*18.3%*	3		*16.5%*	4		*15.6%*	3	
	RF	*31.0%*	*59.0%*	*21.3%*	4		*16.9%*	4		*28.9%*	5		*18.8%*	4		*12.8%*	1		*15.6%*	4	
	LS	*0.0%*	*0.0%*	*44.7%*	7		*33.4%*	7		*40.2%*	6		*30.9%*	7		*47.9%*	7		*35.3%*	7	
Mean	*(a) Level 1 mean*			*20.7%*			*16.2%*			*26.9%*			*17.9%*			*14.3%*			*14.9%*		
	(b) Level 0 mean			*31.0%*			*25.6%*			*32.3%*			*25.8%*			*29.1%*			*25.5%*		
	(a) - (b)			*-10.2%*			*-9.4%*			*-5.4%*			*-7.9%*			*-14.8%*			*-10.6%*		

Notes: Results produced by two versions of Stacking Regression (abbreviated as ST1 and ST2) to obtain TREV predictions for the testing (holdout) sample, which constitute the Level-1 learners. TREV is total revenue in levels. "A" denotes air transport, and "B" denotes bus transport. We consider the following Level-0: Random Forest (RF), Gradient Boosting (GB), Neural Networks (NN), Support Vector Machine (SV), and Least Squares (LS). Hyperparameters adjusted utilizing a random search procedure with 100 trials. CVRMSE and MAPE mean Coefficient of Variation of the Root Mean Squared Error, and Mean Absolute Percentage Error, respectively. CVRMSE extracted by dividing the RMSE by the respective sample mean TREV. Rank means the position of the prediction model in a ranking of models according to its predictive performance. ◄ and ◄ indicate the best performing models in each prediction level, for air transport and bus transport, respectively.

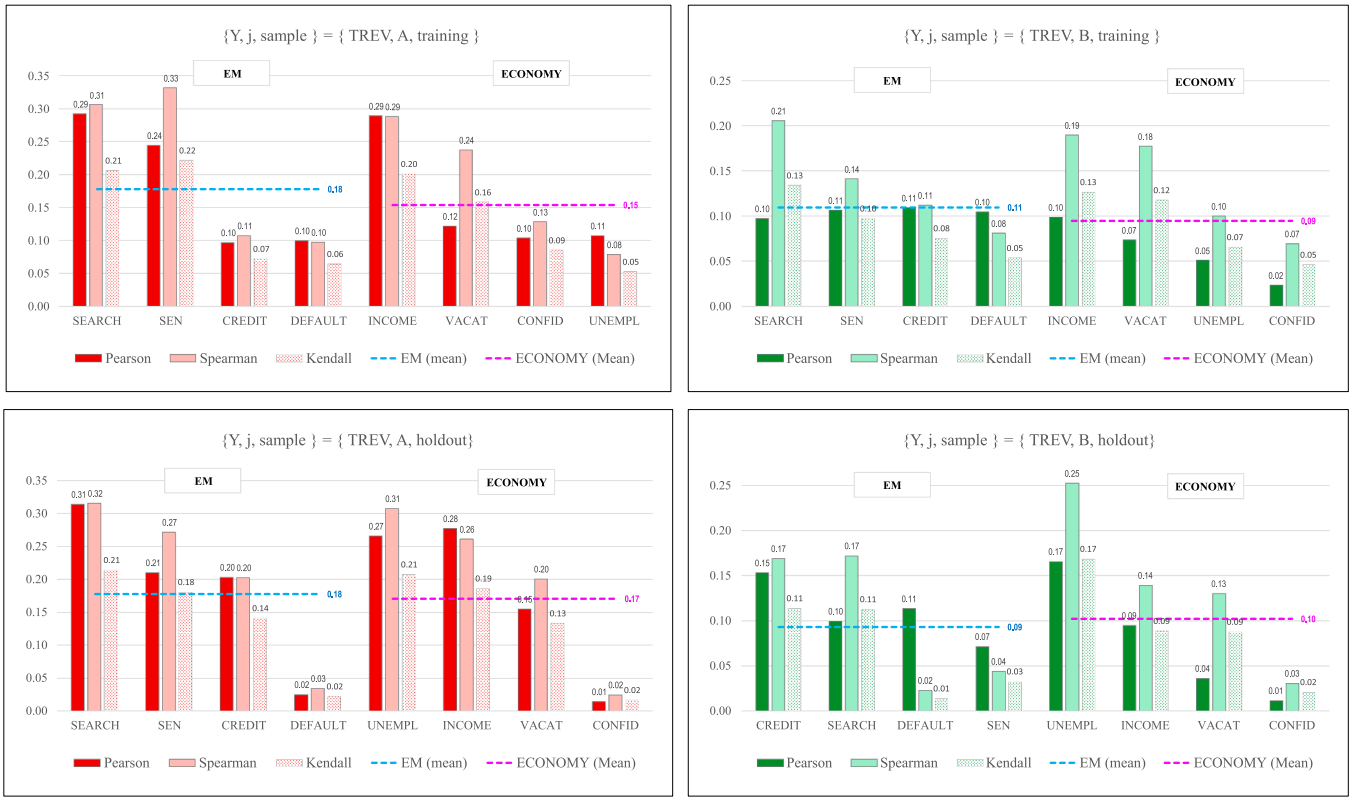

Fig. 6. Impact of Economic Mobility Feature Omission on Predictive Performance.
Notes: Results obtained using the Stacking Regression model (version ST1) to predict TREV, which stands for total revenue in levels. "A" and "B" denote air transport and bus transport, respectively. The horizontal dashed lines in blue and purple represent the average correlation indices for the variables that compose EM and ECONOMY, respectively. Training and holdout refer to the training sample and the holdout sample, respectively.

out that many indicators of social inclusion and EM worsened during and after the recession. Considering the relatively sluggish economic growth recovery after the end of the technical recession, we focus on the period from 2014 to 2018, defined as the "economic downturn" (EDT).

We use a hybrid approach for the event study, combining the findings from both the econometric and machine learning models. In the first step, we use the econometric setup of mode-specific year dummies PER_{jy}, whose coefficients are estimated from Eqs. (20) e (21). The base for these coefficients is 2012, the first year of the data set. To better align to the previous exposition, we rename these coefficients as EDT_τ, where τ denotes the years within the time window under analysis, where $\tau = \{\tau^* - 1, \tau^*, \tau^* + 1, \tau^* + 2, \tau^* + 3, \tau^* + 4\}$, and τ^* is the year when the event was triggered, that is $\tau^* = 2014$.

We calculate the percentage change in total revenue (TREV) for each EDT period using our preferred specifications of Tables 4 and 5.[36] Since TREV is the product of the mean price and the total quantity and the econometric models are in logarithmic form, we sum the coefficients of demand and price for each time dummy, then apply the exponential function to the result minus one, and multiply by 100. The percentages are displayed in the top graphs of Fig. 7, with red and green lines representing the evolution of the air and bus transport estimates, respectively. We present the evolution of PAX, P, and TREV, respectively. For comparison, we include a background chart consisting of blue vertical bar graphs that display the percentage changes in average GDP, which are specifically calculated for the markets served by each transport mode (air on the left-hand side and bus on the right-hand side). The light blue vertical bars represent the periods of technical recession.

Fig. 7 upper graphs reveal a relevant ceteris paribus decline in total revenues for both transport modes during the recessive period. The estimated effects indicate that airlines experienced a substantial drop, ranging from 23.2% in the first year of recession, τ^*, to 26.4% in the second year, $\tau^* + 1$, while buses decreased between 19.7% in the first year and 32.5% in the second year. It is worth contrasting the years of the recession when decline in revenues was steeper: the first year for airlines and the second year for buses, discrepancy which may be assigned to the revenue decline in air transport before the beginning of the recession, i.e. $\tau^* - 1$.

Also intriguing is that the estimated effects of the recession suggest a steady declining trajectory for air transport, especially in $\tau^* + 3$ and $\tau^* + 4$, while for bus transport a modest reversal trend follows the initial decline. This finding may be linked to the fact that in $\tau^* + 4$ the GDP growth rate of bus routes' endpoint cities is the highest of the entire period and nearly the double of the markets served by airlines. Nonetheless, ceteris paribus both transport modes' revenue effects remain negative, indicating falling revenues by the end of the sample period, suggesting that the recession had lingering effects persisting for two years after its onset.

Consistent with the meaning of a recession, both graphs show a sharp decline in demand due to exogenous factors outside the transport sector. Prices exhibit a long-term upward trend during the economic recession. However, in the short term, average airfare prices increased in the years following the recession. The increase, which ceteris paribus is due to fuel prices and HHI, may reflect an attempt to avoid further revenue losses by charging higher prices to price-insensitive passengers. By contrast, bus transport prices declined throughout the time window. None of the transport modes managed to mitigate revenue losses, however.

The bottom graphs of Fig. 7 present the results of a portfolio of alternative specifications based on the best-performing machine learning model, ST1. These alternative models incorporate the EDT

[36] In these experiments, we employ 2SLS in the post-estimation LASSO.

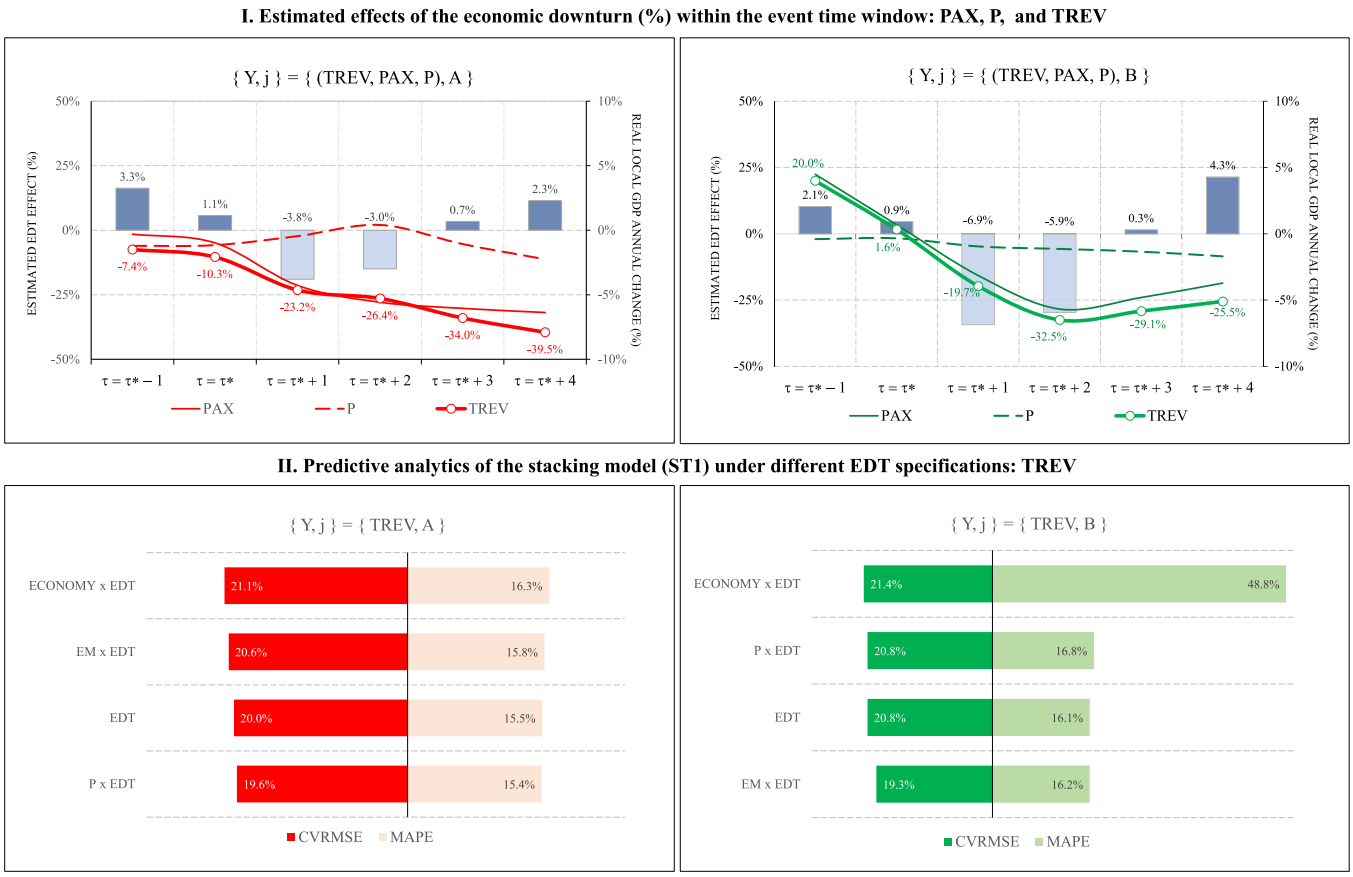

Fig. 7. Combined results of the econometric and machine learning models for the event study period.
Notes: The upper graphics display the estimated percentual effects of the event time-window dummies on the dependent variables PAX and P, obtained using the procedure described in Section 4.1.5. The coefficients of TREV are the sum of the coefficients of the two other equations. Final figures are expressed as equivalent percent changes for each time-window dummy, calculated by taking the exponential of the coefficient of the dummy variable minus one, and multiplying by 100. The bottom graphics show the results of the Stacking Regression, version 1 (abbreviated as ST1), used to predict the total revenue in levels, as described in Section 6. CVRMSE and MAPE mean Coefficient of Variation of the Root Mean Squared Error and Mean Absolute Percentage Error, respectively. CVRMSE is the RMSE over the respective sample mean TREV. The letter "A" denotes air transport, while "B" denotes bus transport.

phenomenon from different perspectives, considering distinct possible configurations for the insertion of this variable in the model. Contrary to the top graphs, which use a set of event window variables, the bottom graphs of Fig. 7 use the EDT dummy to simplify the model, avoiding the overfitting possibly generated by the temporal dummies, but still seeking to predict revenues satisfactorily. The following specifications are attempted: EDT only (EDT), EDT interacted with prices (P × EDT), EDT interacted with each of the variables contained in EM (EM × EDT), and EDT interacted with the variables contained in ECONOMY (ECONOMY × EDT). We report both the CVRMSE and MAPE statistics in these experiments.

The results in bottom graphs of Fig. 7 suggest that none of the specifications significantly affect revenue predictions for both modes of transportation. The only exception is ECONOMY × EDT for bus transportation, which, when measured by MAPE, increases considerably in this specification and approaches 50%. With respect to prediction accuracy, the results indicate that the use of P × EDT for air transportation and EM × EDT for bus transportation minimizes test errors for the models. It is worth noting that these findings allow us once again to conclude that incorporating EM features yield results not significantly different from those using the more traditional variables of ECONOMY. In the case of Fig. 7, incorporating EM x EDT as a feature is more likely to enhance the predictive power of revenue during the economic downturn than using ECONOMY × EDT.

Finally, Fig. 8 shows a set of surface graphs with the estimated coefficients of the time window τ using the Quantile Regression model approach of Section 5. The left charts depict the results for air transport and the right charts, for bus transport. The charts present three types of estimation results: coefficients from Eq. (20) of PAX (upper charts), coefficients from Eq. (21) of P (middle charts), and coefficients of TREV resulting from the combination of both (bottom charts). The graphs confirm a systematic fall in airline revenues throughout the period of economic downturn, in contrast to an initial but not sufficient movement of loss ceasing for bus companies at the beginning of the economic recovery. Notably, the decline in airline revenue due to passenger loss is lower for higher demand quantiles, whereas the opposite occurs for buses, as their loss is greater precisely in these markets. The results suggest that the Quantile Regression's more granular analysis of the results provides valuable implications for setting commercial policies targeting markets where revenue loss is greater.

8. Conclusion

We investigate the effects of economic mobility (EM) on the demand, prices, and revenues of domestic travel transportation' carriers in an emerging country. For this, we utilize a high-dimensional sparse (HDS) method, IV-LASSO, which tackles the complexity of the object as well as of the dataset arising from the use of multiple control variables and instrumental variables. Furthermore, we use quantile regressions for a more granular approach and machine learning models, such as Stacking Regression, to predict total revenues and evaluate the contribution of the proposed features. The analysis relies on intercity air and bus travel

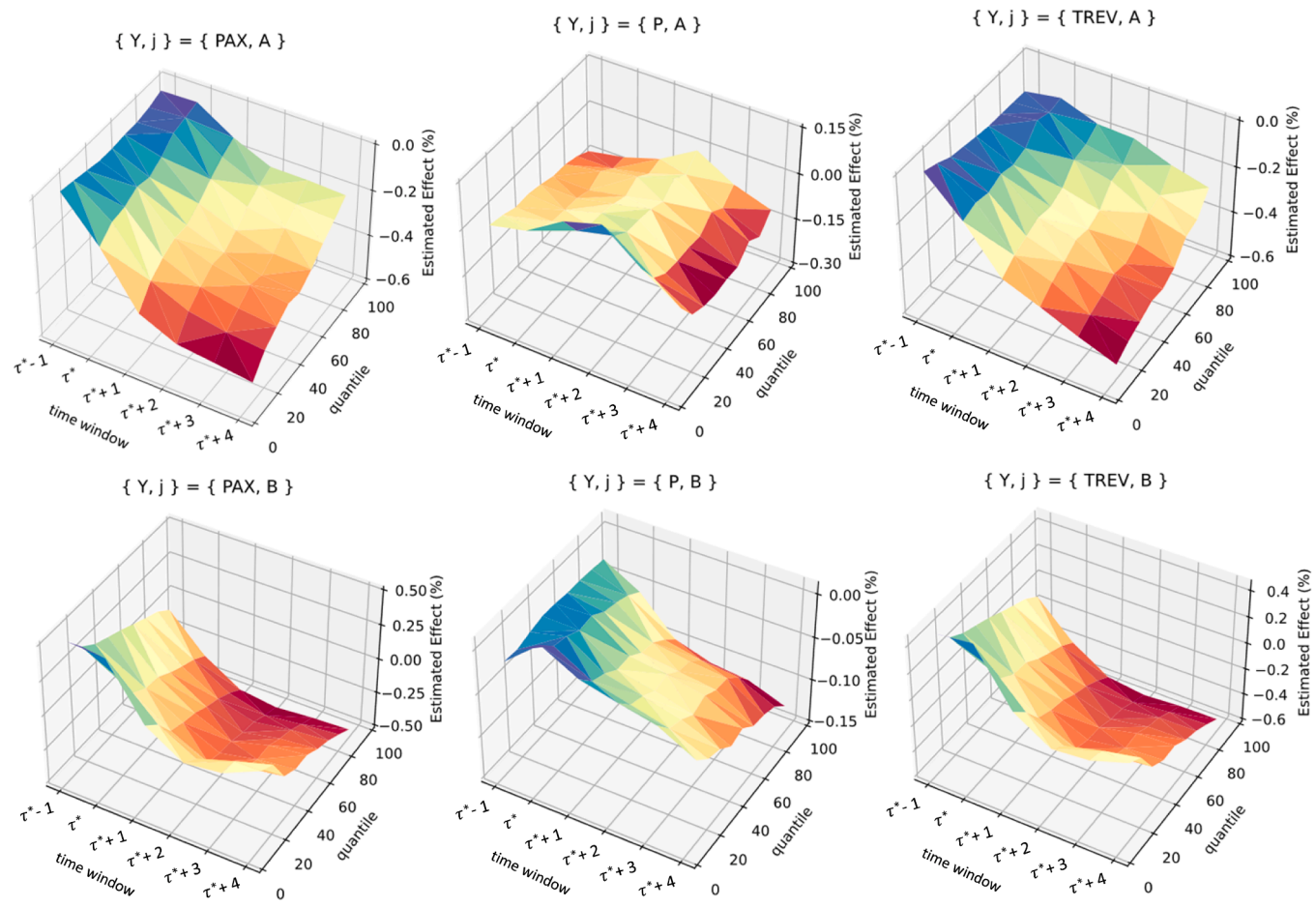

Fig. 8. Estimation results–Quantile Regression: mean travel price (P).
Notes: Estimation results produced by an instrumental variable, fixed-effects, Quantile Regression. The procedure utilizes the same variables selected by the previously estimated IV-LASSO model in Section 5.

data for Brazil from 2012 to 2018. In the first half of the sample period, rapid market growth was accompanied by a surge in the participation of new middle-class travelers, driven by the country's greater inclusiveness. Over the second half of the sample period, a severe economic downturn had an adverse impact on the formerly achieved transport equity gains.

Our causal analysis indicates that bus transportation is an income-neutral good, with lower consumer retention capabilities than airlines'. We also observe that EM proxies are more relevant to the demand for airlines, highlighting the need for the bus industry to improve demand management strategies and market positioning. Furthermore, airlines are generally more effective at capturing potential demand when socioeconomic conditions are favorable, probably due to their more sophisticated revenue management systems. This underscores the efficacy of their market development strategies in attracting consumers from medium and low-income families, particularly when social, financial, and digital inclusion is greater.

We conduct an event study and identify a deflationary recession phenomenon in the transport sector, in which both fares and demand decreased as the economy contracted. Moreover, we show that stacking outperforms the base machine learners used in predicting revenue generation. We also pinpoint the post-recession periods during which the prediction models are more likely to produce higher inaccuracy, finding that carriers should invest in improving their ability to predict revenue during the high season, in order to capitalize on market opportunities.

The initial assumption that variables such as income and unemployment, commonly used in the literature, would outperform the EM variables in revenue prediction does not always hold true, indicating potential benefits of integrating EM variables into economic models of transportation for more accurate revenue prediction. The evidence also highlights the importance of contemplating EM factors when designing marketing policies to target specific passenger segments as well as of paying special attention to early signals of economic downturns and recoveries to design effective market segmentation strategies. By enhancing the predictive analytics of carriers and combining it with transportation operators' effective tactical commercial management, revenue losses can be alleviated or even avoided in those periods. This emphasizes the need for a proactive revenue management approach in the transportation industry, which can be achieved through the development of more adaptable, localized, and scenario-sensitive forecasting models.

Emerging countries often combine consumer diversity and vulnerable economic mobility, challenging carriers to develop effective business intelligence frameworks to capture demand potential. We conduct a quantitative evaluation from a consumer standpoint of the impact of market inclusion in the transportation industry and suggest corporate policy measures to improve market penetration in these travel markets. Nonetheless, given the complexity of the subject and our approach limitations, further research should be carried out to address this topic in other contexts.

CRediT authorship contribution statement

Alessandro V.M. Oliveira: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software,

Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Luca J. Santos:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dante Mendes Aldrighi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author wishes to thank the São Paulo Research Foundation (FAPESP) – grants n. 2015/19444-1 and 2020/01616-0; National Council for Scientific and Technological Development (CNPq) – grants n. 301654/2013-1, n. 301344/2017-5. The second author thanks the Coordination for the Improvement of Higher Education Personnel (CAPES) – Finance Code 001. The authors are grateful to Sérgio Mikio Koyama, Alexandre Castelli, Márcio Nakane, Bruno Alvarenga, Antônio Marcos Ferreira, Ricardo Rocha, Frederico Turolla, Daniel Komesu, Mayara Murça, Rodrigo Scarpel, Paulo Ivo Queiroz. Special thanks to Achim Ahrens. All mistakes are ours.

Appendix

Table 7
Descriptive statistics of the models' variables.

Variable	{j} = {A}					{j} = {B}				
	No Obs	Mean	Std Dev	Min	Max	No Obs	Mean	Std Dev	Min	Max
CONFID	34,589	90.4	17.3	64.9	122.1	18,209	91.4	17.8	64.9	122.1
CREDIT	34,589	4,393.6	942.2	1,905.6	7,109.6	18,209	4,496.1	930.7	2,020.4	6,933.6
DEFAULT	34,589	4.1	0.6	2.2	6.4	18,209	4.1	0.6	2.4	6.4
EDT	34,589	0.7	0.4	0.0	1.0	18,209	0.7	0.5	0.0	1.0
FREQ	34,589	150.2	308.0	1.0	4,112.0	18,209	1,809.5	10,823.9	1.0	313,517.0
FUELP	34,589	2.5	0.5	1.5	3.5	18,209	2.3	0.2	1.8	2.9
HHI	34,589	48.7	23.4	10.9	100.0	18,209	39.2	16.6	10.9	100.0
INCOME	34,589	37,886	11,992	14,522	84,045	18,209	40,376	11,829	15,282	84,045
P	34,589	415.1	168.6	46.0	1,727.1	18,209	179.8	80.4	28.6	517.3
PAX	34,589	7,041	13,754	100	201,479	18,209	34,179	206,371	100	6,000,214
RETENT	34,589	31,850	242,275	0	6,000,214	18,209	14,501	23,698	29	203,831
SEARCH	34,589	5,174,866	8,733,872	43,537	32,995,726	18,209	6,608,033	9,934,004	60,487	32,995,726
SEN	34,589	1,335.9	210.1	614.2	2,127.8	18,209	1,375.0	192.9	820.7	1,972.3
SWITCH	34,589	10,524	18,337	111	203,831	18,209	59,164	330,362	105	6,000,214
TIME	34,589	107.6	46.5	28.0	280.0	18,209	921.7	573.5	117.5	2,895.3
UNEMPL	34,589	60.7	4.3	45.5	69.8	18,209	61.5	3.7	45.5	69.8
VACAT	34,589	6.4	0.4	4.4	7.6	18,209	6.5	0.3	5.2	7.6

Note: All variables and figures have been computed using the authors' own calculations and are presented in their original scale.

Table 8
Matrix of the Pearson correlation coefficients for the models' variables.

{j} = {A, B}		PAX	P	FREQ	TIME	INCOME	VACAT	SEN	CREDIT	DEFAULT	SEARCH	UNEMPL	CONFID	RETENT	ACQUIS	FUELP	HHI	EDT
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
PAX	(1)		-0.41	0.82	-0.53	0.15	0.11	0.11	-0.09	0.08	0.18	0.08	0.07	0.94	0.19	0.00	-0.27	-0.07
P	(2)	-0.21		-0.25	0.88	0.09	0.10	0.11	-0.06	0.10	0.12	0.15	0.03	-0.33	0.30	0.04	0.06	-0.02
FREQ	(3)	0.81	-0.23		-0.34	0.14	0.14	0.10	-0.11	0.20	0.17	0.11	0.18	0.80	0.26	-0.08	-0.23	-0.18
TIME	(4)	0.21	0.60	-0.05		-0.04	0.08	0.01	-0.08	0.13	0.07	0.07	0.01	-0.45	0.26	0.02	0.09	-0.01
INCOME	(5)	0.31	0.04	0.33	0.16		-0.23	0.78	0.47	-0.39	0.31	0.48	0.04	0.18	0.31	0.22	-0.29	-0.04
VACAT	(6)	0.31	-0.08	0.18	0.09	0.10		-0.07	-0.31	0.01	0.02	0.06	-0.12	0.09	0.24	-0.04	-0.03	0.14
SEN	(7)	0.34	-0.04	0.34	0.15	0.82	0.24		0.45	-0.50	0.18	0.43	-0.13	0.12	0.38	0.31	-0.16	0.15
CREDIT	(8)	-0.02	-0.01	0.08	-0.05	0.52	-0.10	0.49		-0.45	-0.03	-0.02	0.04	-0.10	-0.10	0.25	-0.16	-0.12
DEFAULT	(9)	0.04	0.11	-0.02	0.06	-0.43	-0.19	-0.56	-0.47		-0.05	-0.34	0.31	0.09	0.06	-0.43	0.14	-0.42
SEARCH	(10)	0.29	0.02	0.24	0.14	0.33	0.18	0.23	0.01	-0.08		0.28	-0.01	0.21	0.29	0.09	-0.22	0.00
UNEMPL	(11)	0.10	-0.01	0.10	0.07	0.53	0.33	0.51	0.09	-0.44	0.33		-0.09	0.08	0.12	0.11	-0.11	0.12
CONFID	(12)	0.08	0.09	0.07	-0.06	0.05	-0.10	-0.11	0.06	0.24	0.00	-0.09		0.01	0.03	-0.35	0.00	-0.82
RETENT	(13)	0.96	-0.15	0.81	0.20	0.31	0.34	0.35	-0.02	0.03	0.28	0.11	0.02		0.25	0.02	-0.27	-0.03
ACQUIS	(14)	0.30	-0.15	0.29	-0.09	0.27	0.33	0.27	0.08	-0.02	0.24	0.25	0.02	0.32		-0.02	-0.28	-0.04
FUELP	(15)	0.11	0.08	0.11	-0.05	0.09	-0.10	-0.09	0.21	0.08	0.08	-0.18	0.78	0.05	0.01		-0.06	0.37
HHI	(16)	-0.39	0.17	-0.47	0.12	-0.28	-0.13	-0.16	-0.16	0.03	-0.20	-0.03	-0.06	-0.38	-0.41	-0.10		0.00
EDT	(17)	-0.09	-0.08	-0.09	0.07	-0.06	0.13	0.13	-0.16	-0.36	-0.01	0.12	-0.82	-0.05	-0.04	-0.66	0.09	

Note: All variables and figures presented in this analysis have been computed using the authors' own calculations. The color-coding used in the figures represents the absolute values of the variables, with darker shades indicating higher values.

References

Adedeji, P., Akinlabi, S., Madushele, N., & Olatunji, O. (2021). Beyond site suitability: Investigating temporal variability for utility-scale solar-PV using soft computing techniques. *Ren Energy Focus*, 39.

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal, 20*(1), 176–235.

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2023). pystacked: Stacking generalization and machine learning in Stata. *The Stata Journal, 23*(4), 909–931.

Bahadir, B., De, K., & Lastrapes, W. D. (2020). Household debt, consumption and inequality. *Journal of International Money and Finance, 109*, Article 102240.

Barbosa, R. J., Souza, P. H. F., & Soares, S. (2020). Distribuição de renda nos anos 2010: uma década perdida para desigualdade e pobreza (No. 2610). Working Paper. Instituto de Pesquisa Econômica Aplicada, Ipea. Rio de Janeiro. https://doi.org/10.38116/td2610.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica, 80*(6), 2369–2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies, 81*(2), 608–650.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives, 28*(2), 29–50.

Bergantino, A. S., & Madio, L. (2020). Intermodal competition and substitution. HSR versus air transport: understanding the socio-economic determinants of modal choice. *Res. Transp. Econ., 79*, Article 100823 (2020).

Breiman, L. (1996). Stacked regressions. *Machine Learning, 24*, 49–64.

Carmona-Benítez, R. B., & Nieto, M. R. (2023). A methodology for calculating the unmet passenger demand in the air transportation industry. *Research in Transportation Business & Management, 50*, Article 101039.

Chen, J. H., Wei, H. H., Chen, C. L., Wei, H. Y., Chen, Y. P., & Ye, Z. (2020). A practical approach to determining critical macroeconomic factors in air-traffic volume based on K-means clustering and decision-tree classification. *Journal of Air Transport Management, 82*, Article 101743.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review, 105*(5), 486–490.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics, 129*(4), 1553–1623.

Crespo-Almendros, E., & Del-Barrio García, S. (2016). Online airline ticket purchasing: Influence of online sales promotion type and internet experience. *Journal of Air Transport Management, 53*(2016), 23–34.

Correia, S. (2017). reghdfe: Stata module for linear and instrumental-variable/GMM regression absorbing multiple levels of fixed effects. Statistical Software Components s457874. *Boston Col. Dep. Economics.*

Dana, J. D., Jr, & Orlov, E (2014). Internet penetration and capacity utilization in the US airline industry. *American Economic Journal: Microeconomics, 6*(4), 106–137.

Di Maggio, M., & Kermani, A. (2017). Credit-induced boom and bust. *The Review of Financial Studies, 30*(11), 3711–3758.

Dobruszkes, F., & Vandermotten, C. (2022). Do scale and the type of markets matter? Revisiting the determinants of passenger air services worldwide. *Journal of Air Transport Management, 99*, Article 102178.

Foster, J., & Sen, A. (1997). *On Income Inequality*. Oxford University Press.

Garcia-Escribano, M. M., & Han, M. F. (2015). *Credit expansion in emerging markets: propeller of growth? (No. 15-212)*. International Monetary Fund.

Gayle, P. G., & Lin, Y. (2021). Cost Pass Through In Commercial Aviation: Theory And Evidence. *Economic Inquiry, 59*(2), 803–828.

Hanke, M. (2019). Distribution Trends. In Anne Graham, & Frederic Dobruszkes (Eds.), *Air Transport–A Tourism Perspective* (pp. 105–124). Amsterdam: Elsevier.

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

Huang, D., & Rojas, C. (2013). The Outside Good Bias in Logit Models of Demand with Aggregate Data. *Economics Bulletin, 33*(1), 198–206.

Huang, D., & Rojas, C. (2014). Eliminating the outside good bias in logit models of demand with aggregate data. *Review of Marketing Science, 12*(1), 1–36.

Kim, H. L., & Hyun, S. S. (2019). The Relationships among Perceived Value, Intention to Use Hashtags, eWOM, and Brand Loyalty of Air Travelers. *Sustainability, 11*(22), 6523.

Oliven, R., & Pinheiro-Machado, R. (2012). From "country of the future" to emergent country: Popular consumption in Brazil. *In Consumer Culture in Latin America*. New York: Palgrave Macmillan.

Olney, M. L. (1999). Avoiding default: The role of credit in the consumption collapse of 1930. *The Quarterly Journal of Economics, 114*(1), 319–335.

Peelen, E., & Beltman, R. (2013). *Customer Relationship Management* (2nd ed.). Harlow: Pearson Education.

Santos, L. J., Oliveira, A. V. M., & Aldrighi, D. M. (2021). Testing the differentiated impact of the COVID-19 pandemic on air travel demand considering social inclusion. *J. Air Transport Manag., 94*, Article 102082.

Sen, A. (1976). Real national income. *The Review of Economic Studies, 43*(1), 19–39.

Seo, E. J., Park, J. W., & Choi, Y. J. (2020). The Effect of Social Media Usage Characteristics on e-WOM, Trust, and Brand Equity: Focusing on Users of Airline Social Media. *Sustainability, 12*(4), 1691.

Sun, X., Zheng, C., Wandelt, S., & Zhang, A. (2024). Airline competition: A comprehensive review of recent research. *Journal of the Air Transport Research Society*, Article 100013.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1).

Wang, J., Liu, X., & Ding, J. (2019). Air passenger travel forecasting model based on both dynamical individual behavior and social influence force. *Journal of Algorithms & Computational Technology, 13*, Article 1748302619881392.

Wattanacharoensil, W., Schuckert, M., & Graham, A. (2016). An airport experience framework from a tourism perspective. *Transport Reviews, 36*(3), 318–340.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259.

Zaki Ahmed, A., & Rodríguez-Díaz, M (2020). Analyzing the Online Reputation and Positioning of Airlines. *Sustainability, 12*(3), 1184.