# Forecasting the price-elasticity of airline passenger-demand for dynamic price optimization

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

eingereicht von
Jan Felix Meyer
29/07/2023

1. Gutachter: Prof. Dr. Göran Kauermann

2. Gutachter: Prof. Dr. Reinhold Decker

Tag der mündliche Prüfung: 23/10/2023

*Für meine Söhne Julius, Jon und Thilo*

# Acknowledgements

I want to thank everyone who has supported me during my PhD journey.

First and foremost, I am grateful to have stumbled upon the field of statistics, which has become my passion and career. I was a typical student who started studying management and economics without a clear direction. However, thankfully, the bachelor program offered statistics as a specialization, and I was lucky enough to have Prof. Dr Göran Kauermann as my teacher. Prof. Dr Göran Kauermann also helped create the master's program in statistical science that allowed me to evolve my skills and deepen my knowledge of statistics. I am grateful to Prof. Dr Göran Kauermann for allowing me to work as a tutor at the Institution of Statistics, which provided me with financial stability. Prof. Dr Göran Kauermann's enthusiasm for statistics made me attend his lectures in person. He has always been my go-to person for all statistics-related questions.

I also thank him for his patience and support throughout my PhD journey, which took me ten years to complete. There are always reasons why something takes longer than expected, but it is always about priorities. I admire everyone working in the scientific field for their dedication and perseverance, and I am grateful for the opportunities it has given me.

I would also like to acknowledge the support of Dr Christopher Alder, who has helped me understand Revenue Management and kept me away from the politics that often come with working for a large corporation.

I am deeply grateful to my wife, Lina, and my children, Julius, Jon, and Thilo, for their unwavering love and support. They give me the energy and motivation to keep pushing forward and show me daily what is truly important. I would also like to thank my parents and grandparents for their endless love and support.

In conclusion, I am thankful for all the people who have supported me in becoming the statistician I am today, and I look forward to using my knowledge and skills to make a positive impact on the world.

# Danksagung

Ich möchte allen, die mich auf meinem Weg zur Promotion unterstützt haben, von ganzem Herzen danken.

In erster Linie bin ich dankbar, dass ich auf das Gebiet der Statistik gestoßen bin, dass zu meiner Leidenschaft und Karriere geworden ist. Ich war ein typischer Student, der ein Studium der Betriebs- und Volkswirtschaftslehre begann, ohne eine klare Richtung zu haben. Zum Glück bot der Bachelor-Studiengang Statistik als Spezialisierung an, und ich hatte das Glück, Prof. Dr. Göran Kauermann als Lehrer zu haben. Prof. Dr. Göran Kauermann war es auch, der massgeblich dazu beitrug, den Masterstudiengang Statistische Wissenschaft ins Leben zu rufen, der es mir ermöglichte, meine Fähigkeiten weiterzuentwickeln und mein Wissen auf dem Gebiet der Statistik zu vertiefen. Ich bin Prof. Dr. Göran Kauermann dankbar, dass er mir die Möglichkeit gegeben hat, als Tutor am Institut für Statistik zu arbeiten, was mir finanzielle Stabilität verschaffte. Die Begeisterung von Prof. Dr. Göran Kauermann für die Statistik hat mich dazu angetrieben mich für das Fachgebiet der Statistik zu interessieren, weshalb ich jede seine Vorlesungen besuchte. Während der gesamten Zeit meines Studiums und darüber hinaus war Prof. Dr. Göran Kauermann immer der Experte, welcher sich die Zeit nahm, meine Fragen zur Statistik zu beantworten.

Ich danke ihm auch für seine Geduld und Unterstützung während meiner Doktorarbeit, die zehn Jahre andauerte. Es gibt immer Gründe, warum etwas länger dauert als erwartet, aber es geht immer um Prioritäten und ich danke Prof. Dr. Göran Kauermann der mir das Abschliessen der Doktorarbeit als Priorität verdeutlichte. Ich bewundere alle, die in der Wissenschaft arbeiten, für ihr Engagement und ihre Ausdauer, und ich bin dankbar für die Möglichkeiten, die sich mir dadurch eröffnet haben.

Ich möchte mich auch für die Unterstützung von Dr. Christopher Alder bedanken, der mir geholfen hat, den Fachbereich Revenue Management zu verstehen und mich von den politischen Aktivitäten im Unternehmen geschützt hat was mir verhalf mich auf die Doktorarbeit zu fokussieren.

Ich bin meiner Frau Lina und meinen Kindern Julius, Jon und Thilo sehr dankbar für ihre unerschütterliche Liebe und Unterstützung. Sie geben mir die Energie und die Motivation, weiterzumachen und zeigen mir täglich, was im Leben wirklich wichtig ist. Ich möchte mich auch bei meinen Eltern und Großeltern für ihre unendliche Liebe und Unterstützung bedanken.

Abschließend möchte ich mich bei allen Menschen bedanken, die mich dabei unterstützt haben, der Statistiker zu werden, der ich heute bin. Ich freue mich darauf, mein Wissen und meine Fähigkeiten einzusetzen, um die Welt positiv zu beeinflussen.

# Zusammenfassung

In Dienstleistungsbranchen wie dem Luftverkehr bestimmt der Preis die Nachfrage. Um den Umsatz zu maximieren, müssen Fluggesellschaften die Nachfrageelastizität des Preises auf der Mikroebene (täglich für jeden Flug ) vorhersagen, um ihre Preisstrategie so zu optimieren, dass der richtige Kunde das richtige Produktbündel zur richtigen Zeit kauft.

Diese Arbeit leistet einen Beitrag zur Literatur des Operations Research und der Statistik, indem zwei Modelle zur Schätzung der Preissensitivität von Fluggästen bei Flugreisen mit variablen Preisableitungen vorgestellt werden.

Neben der Schätzung der Preissensitivität wird die Anwendungen der Modelle auf das Ertragsmanagement von Fluggesellschaften, insbesondere auf die kontinuierliche Preisgestaltung und die Kundensegmentierung, diskutiert. Da die Ertragsmanagementsysteme der Fluggesellschaften die Nachfrage über den Preis steuern, wird zudem die Preisendogenität berücksichtigt.

Das erste Modell, ein erweitertes verallgemeinertes additives Modell, geht von einer Buchungsintensität als Funktion Kovariaten auf Buchungs- und Flugebenen aus, einschließlich nichtlinearer Effekte, die semiparametrisch mit Hilfe von penalisierten Splines modelliert werden. Durch die Anwendung monotoner ANOVA-artiger glatter Interaktionen bis hin zur bivariaten Ebene können erhebliche Variationen in der Preissensitivität identifiziert und die Prognosegenauigkeit zu gängigen Alternativen übertroffen werden. Darüber hinaus bietet der vorgeschlagene Ansatz einen effizienten Weg zur Implementierung einer kontinuierlichen Preisgestaltung Mittels einfacher mathematischer Funktionen. Des Weiteren wird eine Feldstudie durchgeführt, welche bestätigte, dass der neue Modellierungsansatz zur Umsatzsteigerung von durchschnittlich 6% führt.

Der zweite Ansatz, ein Finite-Mixture-Modell mit kovariatenabhängigen Mixture-Wahrscheinlichkeiten, reduziert die Komplexität des verallgemeinerten additiven Modells, da keine hochdimensionalen Glättungsfunktionen zur Erfassung variabler Preisableitungen geschätzt werden müssen.

Im Vergleich zum verallgemeinerten additiven Modell, das eine einzige Buchungsintensität

mit zahlreichen Glättungsfunktionen modelliert, geht das Finite-Mixture-Modell davon aus, dass Schwankungen in der beobachteten Zahlungsbereitschaft der Fahrgäste auf die Heterogenität der Kunden zurückzuführen sind. Das Mixture-Modell wird für einen Datensatz von über einer Million täglicher Buchungen für 9.602 Linienflüge auf einer Kurzstrecke über zwei Jahre hinweg geschätzt. Die Schätzungen verdeutlichen eine umfangreiche latente Segmentierung der Fluggäste, welche sich in zahlreichen Kovariateneffekten deutlich unterscheiden. Das kalibrierte Modell kann die Nachfrage- und Preiselastizität für Flüge quantifizieren, die an verschiedenen Tagen vor dem Abflug gebucht werden. Da das Modell interpretierbar ist, können Prognosen auch unter unvorhersehbaren Szenarien erstellt werden. Obwohl unser Modell auf der Grundlage von Daten kalibriert ist, welche vor COVID-19 erhoben wurden, dürften viele empirische Erkenntnisse auch dann noch gültig sein, wenn sich der Flugverkehr in der Zeit nach COVID-19 normalisiert.

# Abstract

For service industries such as air travel, pricing drives demand. To maximize revenue, airlines have to predict the demand-elasticity of the price at the micro-level to optimize their pricing strategy such that the right customer buys the right product bundle at the right time.

This thesis contributes to the literature of operation research and statistics by presenting two models for estimating the passengers' price sensitivity for air travel with variable price derivatives at the daily booking and individual flight level. Furthermore, the models' applications to airline revenue management, particularly continuous pricing and customer segmentation, are discussed. Additionally, as airline revenue management systems control demand by price, price endogeneity is considered.

The first, an augmented generalized additive model, assumes a booking intensity as a function of booking and flight level covariates, including nonlinear effects modelled semi-parametrically using penalized splines. The application of monotonicity constraint ANOVA-type smooth interactions up to the bivariate level can identify substantial variations in price sensitivity and exceed state-of-the-art alternatives' predictive performance. The proposed approach offers a simple and efficient way to implement continuous pricing with a closed-form solution. Furthermore, a field study is conducted, which results in a revenue increase of 6% on average.

The second approach, a finite mixture model with covariate-dependent probabilities, reduces the generalized additive model's complexity by not estimating high dimensional smoothing functions to capture variable price derivatives.

Compared to the generalized additive model, which models a single booking intensity with numerous smoothing functions, the finite mixture model assumes fluctuations in the observed passenger willingness to pay to originate from customer heterogeneity. The mixture model is estimated for a unique dataset of over one million daily counts of bookings for 9,602 scheduled flights on a short-haul route over two years. A rich latent segmentation is uncovered, along with strong covariate effects. The calibrated model can quantify demand and price elasticity for flights booked on different days before departure. As the model is

interpretable, forecasts can be created even under unforeseeable scenarios. For instance, while our model is calibrated on data collected before COVID-19, many empirical insights will likely remain valid as air travel slowly recovers in post-COVID-19 times.

# Contents

# Chapter 1

# Introduction

This thesis is the result of the interdisciplinary project between the *AirABC*, specifically the *Resort of Revenue Management and Operation Research* and the *Department of Statistics* of the *Ludwig Maximilians University Munich*. The project's research question is whether integrating accurate willingness to pay estimates for airfares into AirABC's revenue management system can improve revenue.

## 1.1 Motivation

The different ways airlines change the fare of a flight depend upon their ability to discriminate between passenger requests wanting to travel from an origin- to a destination airport ($OD$) with a possible transfer point (Yeoman and McMahon-Beattie, 2010, p. 110). Embedding the work into the existing literature of operation research and applied statistics and motivating its relevance, it is crucial to understand the logic of how AirABC's revenue management system functions.

**Revenue Management**:

Even before the advent of COVID-19, profits in the airline industry were notoriously low. For example, the industry average net margin was only 3.1% in 2019 (IATA, b). Revenue management helps airlines increase their thin margin by about 4-5% (Talluri and van Ryzin, 2005, p. 10). Revenue management practices ensure products are sold to the right customer at the right price (Yeoman and McMahon-Beattie, 2010, p. 9).

Figure 1.1 illustrates a small network of three airports ($\mathcal{A}, \mathcal{B}$, and $\mathcal{C}$). Each dotted line is

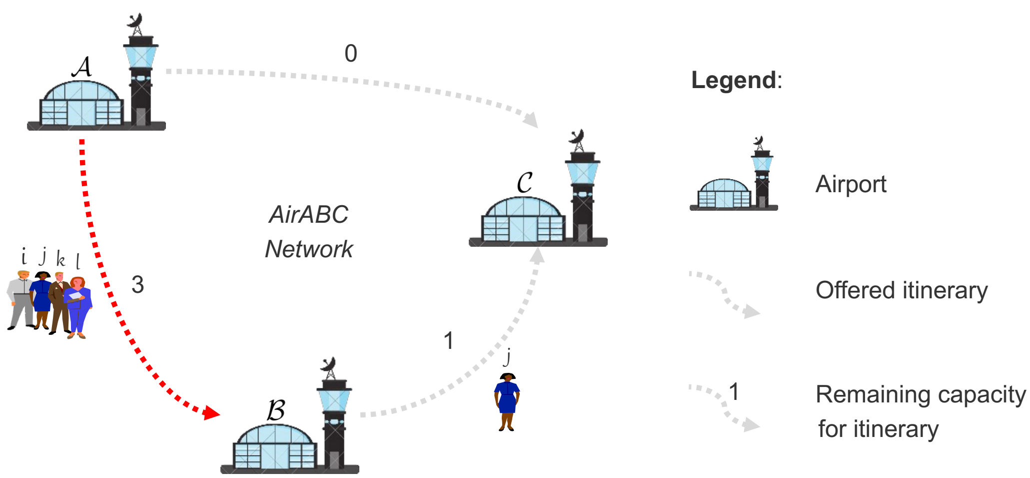



Figure 1.1: Example of an airline network with three airports ($\mathcal{A}, \mathcal{B}$, and $\mathcal{C}$). The remaining capacity is 3 for $\mathcal{A}$ to $\mathcal{B}$, 1 for $\mathcal{B}$ to $\mathcal{C}$, and 0 for $\mathcal{A}$ to $\mathcal{C}$. There are four passenger requests ($i, j, k$, and $l$), three for the $OD, \mathcal{A}$ to $\mathcal{B}$ and one for $OD, \mathcal{A}$ to $\mathcal{C}$.

one itinerary with the number giving its remaining capacity (available seats) as there is four demand for the $\mathcal{A}$ to $\mathcal{B}$ connection, passenger $j$ going from $\mathcal{A}$ to $\mathcal{C}$ via $\mathcal{B}$ and $i, k, l$

requesting $\mathcal{A}$ to $\mathcal{B}$, but only three remaining seats available, the airline needs to decline one of the passengers requests.

**The (simplified) Capacity Steering Problem of AirABC**:

To decide which passengers to accept/decline, the airline maximizes the revenue obtained from selling the limited capacity ($C$) of a perishable asset (once the flight departs, seats cannot be sold again) over a fixed selling horizon (typically one year). Typically, airlines organize their fares ($f$) in (booking-) classes (labeled/indexed by letters of the Alphabet), i.e., $f_A, \dots, f_Z$ and control the seat-inventory by their availability (Yeoman and McMahon-Beattie, 2010, p. 111). Assuming that the willingness to pay of passenger $i, j, k$ and $l$ is $w_i < w_j = w_l < w_k$ and that AirABC offers the fares $f_A = w_i = 30 < f_B = w_j = w_l = 70 < f_C = w_k = 150$, the airline wants to accept passenger $j, k$ and $l$. To that end, a minimum acceptable fare value is introduced, the bidprice ($\pi$) (Belobaba et al., 2015, p. 115). Thus, only passenger requests are accepted where the fare value is at least the sum of $\pi$ over all the requested flight legs, i.e., $f >= \sum \pi$. To maximize revenue, AirABC would choose its bidprice values for each flight-leg ($\mathcal{A}$-$\mathcal{B}$, $\mathcal{B}$-$\mathcal{C}$, $\mathcal{A}$-$\mathcal{C}$) as follows:

- $\pi_{\mathcal{A}\text{-}\mathcal{B}} = \min\{f_B, f_C\} = f_B$ (to decline passenger $i$, fare $f_A$ cannot be offered),
- $\pi_{\mathcal{B}\text{-}\mathcal{C}} = 0$ (there is no capacity constraint),
- $\pi_{\mathcal{A}\text{-}\mathcal{C}} = \infty$ (no more seats are available; the flight is closed for sale).

As $w_i < f_B = \pi_{\mathcal{A}\text{-}\mathcal{B}} = w_j = w_l < w_k$ and $w_j >= \pi_{\mathcal{A}\text{-}\mathcal{B}} + \pi_{\mathcal{B}\text{-}\mathcal{C}} = f_B + 0$, passenger $i$ is denied and $j, k, l$ are accepted. Mathematically, the bidprice results from an optimization problem, having a vast body of literature Pang et al. (2015); Talluri and van Ryzin (2005) and requires the airlines' knowledge of how much demand arrives at any given day before departure ($t$) for every $OD$. For illustrative purposes, we assume that the passenger arrival process $N(t)$ for a specific $OD$ follows a non-homogeneous Poison process with an arrival rate

$$\lambda(t, f) = \exp(\eta(t, f)) \tag{1.1}$$

where $\eta(t, f) = \beta_0 + \beta_1 t + \beta_2 f$ is the linear predictor. The optimization problem of a single flight-leg (for instance $\mathcal{A}$-$\mathcal{B}$) is

$$\begin{aligned} \max \quad & \lambda(t, f) f \\ \text{s.t.} \quad & \lambda(t, f) \leq C(t) \; \forall \; t \in \{0, \dots, 365\} \\ & f \geq 0 \end{aligned} \tag{1.2}$$

where we (ignoring cost for simplicity) maximize revenue ($\max \lambda(t, f) f$) under the constraints that we (ignoring overbooking) cant sell more than the available capacity ($\lambda(t, f) \leq C(t) \; \forall \; t \in \{0, \dots, 365\}$) and ensure that the offered fare is not negative ($f \geq 0$). Solving the Karush-Kuhn-Tucker conditions of Equation 1.2 yields the bid-price as:

$$\pi^* = \begin{cases} \log\left(\frac{D(t)}{C(t)}\right)\alpha & \text{if } C(t) \leq D(t) \\ 0 & \text{if } C(t) > D(t) \\ \infty & \text{if } C(t) = 0 \end{cases} \tag{1.3}$$

where $C(t)$ is the remaining seat-capacity at day $t$, $D(t) = \exp(\beta_0)(1 - \exp(\beta_1 t))$ the remaining demand to come, and $\alpha = -\left(\frac{\partial\eta(t,f)}{\partial f}\right)^{-1} = -\frac{1}{\beta_2} =$ is the negative reciprocal demand elasticity parameter. Note that the third condition of Equation 1.3 is added to ensure that no more passengers are accepted if all the seats are sold. It is important to highlight two key insights into the application of bidprice control:

1. The bidprice control on leg-level does not maximize the revenue as it only considers the capacity constraints on leg-level and not the passengers' willingness to pay on $OD$-level, see Belobaba et al. (2015, p. 61), Boyd and Bilegan (2003, p. 1371). For instance, passenger $j$ (going from $\mathcal{A}$ to $\mathcal{C}$ via $\mathcal{B}$) may not decide to buy every single leg ($\mathcal{A}$-$\mathcal{B}$ and $\mathcal{B}$-$\mathcal{C}$) separately but instead, select to purchase the whole journey $\mathcal{A}$-$\mathcal{C}$ at once.

2. The second condition of Equation 1.3 shows that the bid price is zero when there is less demand than capacity, which means that the airline would offer the lowest fare even though there may be 'yieldable demand,' i.e., passengers with a high willingness to pay.

3. Unlike the demand-model 1.1 using $f$ as an explanatory variable, all of the demand-models used for $OD$-control in practice assume that demand between classes is independent (Fiig et al., 2010, p. 4). Since the dawn of low-cost airlines (LCC), the segmentation criteria between classes have mostly vanished. The only difference between classes is the price, so the independence assumption between classes is violated in practice.

**The (simplified) Pricing Problem of AirABC**:

Based on the example of Figure 1.1, passenger $k$ buys (down to) fare $f_B = 70$ even though the willingness to pay is $w_k = f_C = 150$. Therefore, accepting passenger $k$ for $f_B$ results in a revenue loss of $f_C - f_B = 80$. To ensure that fare $f_B$ is not offered and to effectively (price-) discriminate passenger requests, the airline adjusts the fare values to ensure that $f_C \geq \pi_{\mathcal{A}\text{-}\mathcal{B}} > f_B$. Mathematically, the airline maximizes the margin over bidprice

$$\begin{aligned} \max_{f} \quad & \lambda(t, f)(f - \pi^*) \\ \text{s.t.} \quad & f \geq 0 \end{aligned} \tag{1.4}$$

If we assume that $\lambda(t, f)$ follows an exponential demand model, the (closed form) solution of Equation 1.4 is:

$$f^* = \alpha + \pi^* \tag{1.5}$$

where the $\alpha$-parameter is the margin and $\pi^*$ the bid-price as defined by Equation 1.3. Even if $\pi^* = 0$, the airline charges a fare value of at least $\alpha$. If airlines organize their fares in (booking-) classes $f_A, \ldots, f_Z$, selecting a fare value from a continuous range is impossible. As mentioned earlier, the concept of fare adjustments, as discussed by Fiig et al. (2010), ensures that the offered fare maximizes the revenue for the case where airlines need to offer fares from a discrete selection of classes. For the example of Figure 1.1, the airline calculates the revenue for each possible option. Offering the $\mathcal{A}$-$\mathcal{B}$-connection for a $f_A$ results in a revenue of $3f_A = 90$. For class $B$ the revenue is $2f_B = 140$ and $f_C = 150$ for class $C$. The adjustment fare ($f^{‘}$) is defined by the marginal revenue contribution of receiving an additional booking in a particular class and requires the airlines' knowledge of how much the demand decreases if the price increases (the price elasticity of demand). Mathematically, the fare adjustment for class $A$ is the revenue difference to the next higher class, i.e., $3f_A - 2f_B = 90 - 140 = -50$, divided by the difference in demand $3 - 2 = 1$, i.e., $f^{‘}_A = -50$. Therefore, the airline would not offer class $A$ as it loses 50 revenue. The adjusted fare for class $B$ is $f^{‘}_B = (2f_B - f_C)/(2-1) = (140-150)/1 = -10$. Thus, $B$ is not offered. Note that the highest fare $f_C$ is not adjusted. To ensure that passenger $k$ books its ticket for fare $f_C$, the availability control mechanism of AirABC uses the adjusted fares $f^{‘}$ instead of the actual fares $f$. As $f^{‘}_A, f^{‘}_B < \pi_{\mathcal{A}\text{-}\mathcal{B}}$, only class $C$ with fare $f_C$ is available, and the maximum revenue is obtained.

To apply fare adjustments, it is vital to highlight that the price derivative of the linear predictor of Equation 1.1 needs to change within the variables the airline wants to use for price discrimination. Remember, that the margin parameter $\alpha$ of the optimal fare $f^*$ 1.5 for $\eta(t, f) = \beta_0 + \beta_1 t + \beta_2 f$ is $\alpha = -\frac{1}{\beta_2}$. Therefore, the margin will not change even though the airline may suspect that the passengers' willingness to pay differs for requests at a specific time before departure and itinerary. The solution to this problem is the identification of relevant interaction effects. For instance, to discriminate passenger requests at the time of booking, the airline needs to add the function $f(t, f) = \beta_3 t f$ to $\eta(t, f)$, which results in $\alpha = -\frac{1}{\beta_2 + \beta_3 t}$.

**Disentanglement**:

To discriminate between passenger requests, applying fare adjustments and bidprice defines the AirABCs revenue management $OD$ system. The literature, see Pölt (2016, p. 238), Vinod (2016), and Doreswamy et al. (2015) acknowledges that a $OD$-system outperforms the leg-system in terms of revenue gain. Despite the superior revenue performance, $OD$-systems require unfeasible amounts of demand-forecasts (Boyd and Bilegan, 2003, p. 1377), which makes its application numerically unstable and results in revenue loss. With an average of 3.000 flights per day[1] × 366 days of booking × 20 fare-classes × 2 (inbound-/outbound-differentiation, about 43.920.000 demand forecasts need to be created every day. Regardless of the vast number of forecast units, the number increases even further if price discrimination) is performed for different demand segments, considering the passengers' length of stay. As the number of bookings per day is only a small percentage

[1]pre-COVID19-times there was an average of 90.000 flights per month

of the number of forecast units (and specifically does not change if the airline increases the demand forecast's granularity), the amount of bookings per demand estimate goes to zero. Bartke (2014, Chapter 7), who termed this issue *the problem of small numbers*, shows that the mean-squared estimation error becomes arbitrarily large. To avoid obtaining a forecast on $OD$-level but maintaining its benefits, Pölt et al. (2018) proposed the idea of disentanglement.

Instead of estimating demand on $OD$-level for capacity control, a model for network contribution depending on the bid price, time to departure $t$, and itinerary information such as travel time, departure day of the week, and departure time is estimated. Consequently, airlines can solve the pricing- and capacity-steering problem independently. Specifically, to solve the pricing problem, disentanglement allows (1) application aggregation techniques to counter the aforementioned 'problem of small numbers' and (2) not having to maintain a complete data set for every passenger reservation. Hence, data from staff travel or redemption can be removed if modeling the price elasticity of demand.

## 1.2 Outline

This thesis's contribution focuses on the pricing problem of AirABC, specifically identifying demand models that allow AirABC to price discriminate effectively. To that end, the demand model needs variable price derivatives for any dimension where the passengers' willingness to pay shows significant fluctuations. This thesis proposes two modeling approaches of retail demand for air travel and ticket price elasticity at the daily booking and individual flight level. Both models assume that daily bookings are modeled as a nonhomogeneous Poisson process concerning the time to departure.

The first approach, an augmented generalized additive model (e.g. Wood, 2017), is described in Chapter 2 consisting of five sections. The augmentation concerns two aspects. Firstly, to estimate a full factorial model with all possible uni- and bivariate smooth functions of covariates, ANOVA type interactions (Lee and Durbán, 2011) are used. Secondly, to ensure that demand decreases in price, smooth functions containing price must be constraint (Pya and Wood, 2014). The outline of Chapter 2 is as follows. Section 2.2 briefly introduces the statistical methodology and reviews related literature on demand estimation and dynamic pricing. Section 2.3 outlines the model. Section 2.4 discusses the results of applying the model to empirical airline data and presents a forecasting benchmark. Section **??** highlights the practical applicability within the airline domain by demonstrating how the proposed algorithm can return continuous or discrete prices. Section **??** also documents a field study implementing our approach and dynamic pricing. The chapter concludes with section 2.6, giving airline managers insights into how adjustments to the augmented generalized additive model allow describing changes to passenger demand and price elasticity at the aggregated level due to COVID-19.

The second modeling approach is presented in Chapter 3. Here, it is assumed that fluctuations in willingness to pay and price derivatives originate from customer heterogeneity. Therefore, the booking intensity is captured as a function of booking and flight level covariates, including nonlinear effects modeled semi-parametrically using penalized splines. The customer heterogeneity is incorporated using a finite mixture model, where the latent segments have covariate-dependent probabilities. Chapter 3 is organized into four sections. Section 3.2 provides a brief overview of part of the extensive literature on modeling airline passenger demand considering customer heterogeneity. Section 3.3 introduces the employed dataset, while Section 3.4 outlines the flexible Poisson model. The latter includes the mixture model, penalized spline smoothing, penalized maximum likelihood estimation, and an endogeneity correction approach. Section 3.5 contains the empirical analysis.

## 1.3 Disclaimer

The methodology and the presentation of its results depend on data gathered from 2012 to 2015. Since the beginning of the coronavirus pandemic in Feb 2020, the entire airline industry has been in the midst of an unprecedented crisis. Though many of the results presented in this thesis only apply to pre-corona times, we outline the methodological applicability to post-corona times in the respective chapters.

# Chapter 2

# Modelling Price Sensitive Demand in Turbulent Times: An Application to Continuous Pricing

**Contributing article**:
The content of Chapter 2 is an alternative modelling approach of the article (see, Chapter 3):

Meyer JF, Kauermann G, Smith MS. Interpretable modelling of retail demand and price elasticity for passenger flights using booking data. Statistical Modelling. 2022;0(0).
doi:10.1177/1471082X221083343,
available at https://journals.sagepub.com/doi/full/10.1177/1471082X221083343.

**Author contributions**:
Jan Felix Meyer conceived the research question and was responsible for the data management and implementation. He prepared the first draft, including working examples and visualisations. Göran Kauermann, Christopher Alder, and Catherine Cleophas provided valuable inputs to all sections of the article. All author revised and and proofread the manuscript.

## 2.1 Abstract

Pricing drives demand for service industries such as air transport, hotels, and car rentals. To optimise the price, firms have to predict real-time customer demand at the micro level and optimise the price. This paper contributes to revenue management by introducing a nonparametric statistical approach to predict price-sensitive demand and its application to continuous pricing.

Continuous pricing lets service companies maximise revenue by using customers' willingness to pay. However, it requires accurate demand estimations, particularly of customers' price sensitivity. This paper introduces an augmented generalised additive model to estimate price sensitivity, which identifies substantial variations in price sensitivity, exceeds the predictive performance of state-of-the-art alternatives, and controls for price endogeneity. In addition, the demand model has variable price derivatives enabling continuous pricing.

The proposed approach offers a simple and efficient way to implement continuous pricing with a closed-form solution. Our research also highlights the relevance of considering the problem of price endogeneity when estimating price-sensitive demand based on observations that resulted from prior pricing decisions.

We demonstrate how continuous pricing is applied using empirical airline ticket data. We document a field study, which shows a revenue increase of 6% on average and outline how the approach applies to turbulent market conditions caused by the COVID-19 pandemic, the surge in inflation since mid-2021, and the start of the Ukraine war in April 2022.

## 2.2 Introduction

### 2.2.1 The Revenue Management Environment

As outlined by IATA (a), recent efforts by the International Air Transport Association (IATA) revolutionize how airlines retail, distribute, and sell their products. In particular, reducing distribution costs and increasing control of the offered content is in the focus (Bingemer, 2018). From the passengers' point of view, the airline initiative leads to a new shopping experience with more personalized offers (Wittman and Belobaba, 2017), increased accuracy in pricing (Wittman and Belobaba, 2019), and more convenient customer touch points (Sankaranarayanan and Lalchandani, 2019) to sell new products.

Viewing the changes within the airline industry from a revenue management point of view outlines that the classical practice of revenue management (availability control of booking classes) needs to change accordingly to tap into the newfound revenue potential to increase the airlines‘ profitability. One of the potentials that airlines discovered recently is continuous pricing, where the airline decides to offer one of many possible price fixed price points; within continuous pricing, the price moves continuously from a lower to an upper bound.

Besides implementing new revenue management practices, past years have shown that airlines need to be capable of quickly adapting to changing market conditions. Due to the global COVID-19 pandemic, since the beginning of 2020, the global aviation industry has been struck by an unprecedented crisis. Furthermore, the surge in inflation since mid-2021 and the start of the Ukraine war in April 2022 have added additional complexity and uncertainty to the market, making traditional methods of price elasticity estimation within airline revenue management even more challenging. These events underscore the need for alternative approaches and careful consideration when analyzing current market conditions. From an airline operations perspective, all the revenue management systems depending on demand-forecasting models for price optimization need to be corrected as they rely on data recorded before the start of market turbulences, such as the COVID-19 pandemic. This section gives airline managers insights into how passenger demand and price elasticity change on an aggregated level by analyzing data gathered during the COVID-19 pandemic.

### 2.2.2 Motivation and Research Goals

Similar to related work of Arandia (2013), Wu and Akbarov (2012), and Zhang and Kou (2010), we model customer arrivals by a nonhomogeneous Poisson process (NHPP). Thus, the Poisson intensity defines the expected number of arrivals per day. The Poisson intensity is modeled as a function of covariates and accounts for variable price sensitivity.

The proposed approach considers multivariate functional dependencies between price and additional covariates to capture the unknown structure of price sensitivity. However, it does not assume a known functional form but approximates functional components via penalized splines. A penalized spline approximates an unknown function by a linear combination of basis functions. Common bases are B-splines (de Boor, 1978), cubic (Gu, 2002, p.2) or, thin-plate splines (Wood, 2017, p.150). We refer to Wood (2017) for technical details and to Marx et al. (2016) for a general overview.

The approach presented here includes regularisation parameters (e.g. Eilers and Marx, 1996) to control the trade-off between flexibility and generality. Similar to Blundell et al. (2012) or Brezger and Steiner (2008), we additionally impose a monotonicity constraint on functions that contain price such that demand decreases when price increases. As pointed out by Tutz and Leitenstorfer (2007), imposing monotonicity also improves the validity of price-sensitivity estimates. To this end, we combine the penalized smoothing spline ANOVA type interaction model of Lee and Durbán (2011) with the shape-constrained generalized additive model framework of Pya and Wood (2014). Considering smoothing spline ANOVA-type interactions allows us to disentangle changes in the volume of demand and related pricing effects.

Our work contributes to the literature on service demand estimation and pricing. Firstly, it extends the class of nonparametric models by augmenting generalized additive model framework (e.g. Wood, 2017), combined with monotonicity constraints (Pya and Wood, 2014) and ANOVA type interactions (Lee and Durbán, 2011). Secondly, we extend the model and correct for price endogeneity.

To illustrate the use of the resulting demand estimates, we will also present a dynamic pricing approach. To do so, given opportunity costs, we assume to define the optimal price via a closed-form solution as the sum of a margin and a cost component. Based on empirical airline data, we show that the proposed approach outperforms a selection of state-of-the-art demand estimation routines regarding forecasting accuracy. These competing approaches include the parametric model of Fiig et al. (2014), the nonparametric model of Vulcano et al. (2012), and the heuristic of Weatherford and Pölt (2002). Finally, we document a field study that verifies the model's ability to increase revenue by 6% on average in the airline setting. In the remainder of this paper, Section 2.2.3 reviews related literature on demand estimation and dynamic pricing. Section 2.3 outlines the model. Section 2.4 discusses applying the model to empirical airline data and presents results from benchmarking our approach to alternatives. Section **??** presents a dynamic pricing algorithm. The proposed algorithm can return continuous or discrete prices. The practical applicability of our approach within the airline industry is highlighted within section **??**, where the application of dynamic pricing is demonstrated through a field study. Finally, Section 2.7 concludes the paper.

### 2.2.3 Literature Review on Demand Estimation and Dynamic Pricing

The literature on demand estimation given functional structures can be categorized into two groups. The first group includes parametric and linear models, which implicitly assume constant price sensitivity. An example is the parametric, multiplicative, and non-linear forecast model (FCST) of Fiig et al. (2014). In FCST, the upsell probability captures price sensitivity and is assumed to be independent of other confounders. To model a customer choice, multinomial logit (MNL) models are proposed by Vulcano et al. (2010); Newman et al. (2014); Dai et al. (2014), and Xie et al. (2016). All these approaches assume a linear relationship between covariates and the utility defining the choice probability.

The second group represents nonparametric techniques that allow for more complex concepts of price sensitivity. Relaxing the linearity assumption, Vulcano et al. (2012) introduces a nonparametric approach based on the expectation maximization (EM) algorithm. They employ a mixture of a Poisson- and MNL distribution, where a multinomial utility choice model links the booking decision to a choice probability. The model also differentiates primary and secondary demand, i.e., assuming customers‘ first choice to be available.

Several contributions focus on NHPP and propose data-driven techniques to capture a dynamic arrival rate while abandoning pre-defined functional structures. To model the NHPP's cumulative intensity function nonparametrically, Leemis (1991) proposes piecewiselinear interpolation. Zhang and Kou (2010) analyze the doubly stochastic Poisson process (or Cox process) and show that nonparametric kernel functions improve the fit for high arrival rates. Beyond parametric or nonparametric demand functions, heuristics use imputation when demand observations are censored. In revenue management, this challenge arises when the product is not offered during part of the sales horizon. In such situations, Weatherford and Pölt (2002) proposed substituting demand with the mean number of bookings.

The (dynamic) pricing literature also discusses parametric versus nonparametric demand estimation approaches. Given Poisson-distributed customer arrivals, where a Bernoulli variable defines the purchase probability, Avramidis (2013) proposes a way to estimate arrival rates and purchase probabilities. The authors show their approach can outperform the estimator introduced in Besbes and Zeevi (2009). Given a linear demand-price relationship, Keskin and Zeevi (2014) proposes the greedy iterative least squares (GILS) approach. Extending this work, Besbes and Zeevi (2015) shows that assuming a linear price-demand relationship does not significantly diminish revenue under reasonably general conditions. Including additional covariates such as market expenditures, geographical information, and socio-economic attributes, Qiang and Bayati (2016) extend the GILS approach. Also, assuming linearity for every covariate, the authors show asymptotically optimal performance.

Abandoning the linearity assumption, Farias et al. (2013) captures customers' choice behavior to predict revenue gains. The authors conclude that nonparametric techniques are better suited for large-scale automatization as they rely on something other than expert information. Proposing a nonparametric method that uses B-splines for approximation, Chen et al. (2014) claims that nonparametric techniques are asymptotically robust if the demand function is sufficiently smooth. The authors also show that misspecified parametric methods can cause substantial revenue losses.

Here, we propose differentiating models based on their demand function as seen in den Boer (2015). The demand function may be static or dynamic. Unlike static demand functions, where changes in pricing are motivated by limited capacity, we model demand via uni-, and bivariate functions of price and other covariates. This setting renders the demand function dynamics for confounding variables such as time. We contribute to nonparametric dynamic pricing by non-linear relationships of demand and price to confounding variables, which can also influence price sensitivity. We achieve this goal via a bivariate and penalized B-spline setting. The resulting dynamic pricing algorithm relies on price derivatives, which change dynamically over confounding variables. To consider scarce capacity, we assume known opportunity cost to evaluate the margin and cost component separately. This assumption enables subsequent optimization steps to consider separate capacity allocation and revenue maximization. Revenue maximization solves a Cournot-type price optimization problem. Pölt et al. (2018) describes a similar concept, concentrating on capacity allocation.

Our work also contributes to research on price endogeneity. Price endogeneity causes biased sensitivity estimates: Mumbower et al. (2014); Lo et al. (2015), and Petrin and Train (2010) show that ignoring price endogeneity means underestimating the price coefficient. Typically, instrument variables are suggested to cure endogeneity. Usually, the instrument variable is a linear combination of (presumably) exogenous variables. Given a fitting instrument variable, unbiased estimates result from a two-staged estimation procedure (Davidson and MacKinnon, 1999).

In recent times, driven by the COVID-19 pandemic, the surge in inflation since mid-2021, and the start of the Ukraine war in April 2022, the models' capability to adjust to changing market conditions has become an important feature. As described by Yeoman (2021), airlines need to accept COVID-19 as the new norm until the coronavirus pandemic disappears from this planet. Yeoman (2022), analyses the impact of the Ukraine war and inflation on price sensitivity and demand. To adjust how revenue management is performed, Vinod (2021) suggests monitoring key revenue management metrics and taking corrective action with demand and supply levers to make the revenue plan happen. To perform corrective actions using the presented model framework of this paper, Bonciolini (2022) presented a price elasticity monitoring method that automatically adjusts price elasticity estimates. Similarly to adjustments to price elasticity, Gatti Pinheiro et al. (2022) introduced a shock detector to identify positive and negative shocks in demand volume and willingness-to-pay. Besides applying adjustments as proposed by Bonciolini (2022), our work shows that the changing market conditions caused by the COVID-19 pandemic can be considered by adding additional functions to the model.

## 2.3 Statistical Model and Estimation

This section outlines the statistical model class and introduces parameter estimation based on a penalized likelihood procedure. We also describe the two-staged estimation approach.

### 2.3.1 Model Development

Let index $i = 1, \dots, M$ describe a flight connection between two cities. Let $N_i(t)$ define the number of accumulated bookings for flight $i$ at the time $t \in \left[t_i^{\text{open}}, t_i^{\text{close}}\right]$, where the interval represents the sales horizon during which product $i$ is offered. Indexing $t_i^{\text{open}}$ and $t_i^{\text{close}}$ of the sales horizon by $i$ allows for individual sales horizons per product. We assume that $N_i(t_i^{\text{open}}) \equiv 0$, i.e., there are no observed bookings at the beginning of the sales horizon.

We model the accumulated bookings $N_i(t)$ as a Poisson process, such that the increments, $N_i(t) - N_i(t-1)$ are Poisson-distributed with

$$P\Big(\big\{N_i(t) - N_i(t-1)\big\} = y_{i,t}\Big) = \frac{\lambda(t)^{y_{i,t}}}{y_{i,t}!} \exp\big(-\lambda(t)\big). \tag{2.1}$$

Equation (2.1), also includes the possibility of observing no bookings ($y_{i,t} = 0$, referred to as non-bookings). The Poisson intensity $\lambda(t)$ accounts for changes in booking intensity and depends on price and additional observable covariates. The covariates for product $i$ at time $t$ with price $p_{i,t}$ are given by a covariate vector $(\mathbf{x})_{i,t} = (x_{1,i,t}, \dots, x_{K_2,i,t}, p_{i,t}, z_{1,i,t}, \dots, z_{K_1,i,t})$. The index-sets $I_1 = \{1, \dots, K_1\}$ and $I_2 = \{1, \dots, K_2\}$ give the positions of the categorical covariates $x_{k,i,t}, (k \in I_1)$ and continuous covariates $z_{k,i,t}, (k \in I_2)$. For categorical covariates in $I_1$, the $k^{\text{th}}$ variable takes values from the set $J_k = \{1, \dots, G_k\}$. This leads to the model:

$$\lambda((\mathbf{x})_{i,t}, t) = \lambda\left(x_{1,i,t}, \dots, x_{K_2,i,t}, p_{i,t}, z_{1,i,t}, \dots, z_{K_1,i,t}\right). \tag{2.2}$$

The model aims to quantify the effect of price on booking intensity given covariates $z_{k,i,t}, (k \in I_2)$ and $x_{k,i,t}, (k \in I_1)$. Thus, the effects of covariates on booking intensity need to be specified. To this end, the model captures all bivariate interaction effects of the continuous covariates by setting

$$\begin{aligned} \log\Big(\lambda((\mathbf{x})_{i,t}, t)\Big) = \beta_0 &+ \sum_{k \in I_1} \sum_{j \in J_k} (\mathbf{1})_{\{x_{k,i,t}=j\}} \beta_{k,j} \\ &+ f_p\left(p_{i,t}\right) + f_{p,t}\left(p_{i,t}, t\right) + \sum_{k \in I_2} f_{p,k}\left(p_{i,t}, z_{k,i,t}\right) \\ &+ f_t\left(t\right) + \sum_{k \in I_2} f_k\left(z_{k,i,t}\right) + \sum_{k \in I_2} f_{t,k}\left(t, z_{k,i,t}\right) + \sum_{\substack{k_1<k_2 \\ k_1,k_2 \in I_2}} f_{k_1,k_2}\left(z_{k_1,i,t}, z_{k_2,i,t}\right). \end{aligned} \tag{2.3}$$

Here, $(\mathbf{1})_{\{x_{k,i,t}=j\}}$ is an indicator function that equals one if the categorical covariate $x_{k,i,t} = j \in J_k$. The coefficient-vector $(\beta)_{k \in I_1} = (\beta_{k,1}, \dots, \beta_{k,G_k})$ quantifies the effect of the $k^{\text{th}}$ categorical variable on the booking intensity. Similar to a full factorial design, which analyses the effect of each covariate as represented by the univariate function $f(\cdot)$, bivariate function $f(\cdot,\cdot)$ captures all interactions between covariates. For example, $f_p\left(p_{i,t}\right)$ determines the general level of price-sensitivity, and $f_t\left(t\right)$ describes the dynamics in the arrival of bookings. The interaction effect $f_{p,t}\big(p_{i,t}, t\big)$ quantifies how price-sensitivity changes over the sales horizon.

To describe price sensitivity, we first isolate demand components unrelated to price. These are captured by price-independent functions $f_t(\cdot), f_{k \in I_2}(\cdot)$, and $f_{t,k \in I_2}(\cdot,\cdot), f_{\substack{k_1<k_2 \\ k_1,k_2 \in I_2}}(\cdot,\cdot)$. In contrast, $f_p(\cdot), f_{p,t}(\cdot,\cdot)$, and $f_{p,k \in I_2}(\cdot,\cdot)$ amend the slope of price, representing price-sensitivity. When demand is price-sensitive, fewer bookings occur if the price is high, and more bookings occur if the price is low. Monotonicity constraints as proposed in Pya and Wood (2014) for functions $f_p(\cdot), f_{p,t}(\cdot,\cdot)$, and $f_{p,k \in I_2}(\cdot,\cdot)$ ensure this. Appendix A.1 provides a detailed description and a technical discussion of the penalty setup. Furthermore, Appendix B.2 outlines the implementation to consider price endogeneity during the estimation process.

## 2.4 Empirical Data

To demonstrate the proposed approach, we apply it to airline booking data.

The empirical data set includes $1,708,236$ bookings ($y_{i,t} > 0$) for the economy compartment as offered on eight European city pairs between April 1, 2012, and December 31, 2015. Eight pairs of origin and destination (OD) create 16 point-to-point (P2P) connections, as passengers can travel each connection in two directions. Each flight $i$ is described by the continuous covariates departure day of the year (YDAY, taking values from $1, \ldots, 365$) and departure time (DTIME, taking values between 0 and 24). The dataset also includes daily snapshots of the lowest available fare ($p_{i,t}$). Each booking is attributed to one of 16 P2P connections and one of the 10 to 14 fares offered.

When the airline records no bookings for flight $i$ on the day $t$, a non-booking entry ($y_{i,t} = 0$) reports the offered price. Entries are further described by the categorical covariates booking weekday (BDAY, taking the values Monday, ..., Sunday), the departure weekday (DDAY, taking the values Monday, ..., Sunday), and the P2P connection. Flight, booking, and availability data create a complete record of available and booked fares for the entire booking horizon. The daily snapshot only covers price changes within one booking day. Therefore, when multiple bookings for different prices occur on the same day, $p_{i,t}$ is observed at a finer resolution than per day $t$. Appendix C.3 describes how an extension of model (2.3) compensates for this.

We exclude flight departures on public or school holidays, major fairs, exhibitions, or conferences. If a flight is canceled or re-scheduled, we maintain data from before the adjustment and treat data collected after the change as a new flight. In consequence, we consider 70,283 flights and 3,225 departure days.

Our analysis excludes bookings of fares with fewer restrictions than the lowest available fare. From a pricing point of view, the revenue gain from such upselling could be modeled separately. Here, we aim only to find the best price for the basic fare, as ancillary features could complicate the estimation of the price sensitivity.

The analysis only considers ticketed bookings. It focuses on return tickets, which represent 95.5% of all bookings. Finally, to reduce the number of non-bookings, it only considers the slice of the sales horizon that captures 99% of bookings. Table 2.1 summarises the analyzed data set.

The first row lists the number of days in the sales horizon considered. The second row describes the number of daily services. Row three and four report the number of bookings ($y_{i,t} > 0$) and non-bookings ($y_{i,t} = 0$).

| Origin destination pair | Booking horizon | Daily services | Bookings | Non-Bookings |
|---|---|---|---|---|
| OD1 | 291 | 6 | 101,905 | 1,193,193 |
| OD2 | 285 | 8 | 192,764 | 2,007,423 |
| OD3 | 254 | 12 | 280,130 | 3,339,491 |
| OD4 | 208 | 4 | 36,076 | 887,384 |
| OD5 | 282 | 5 | 99,077 | 1,339,820 |
| OD6 | 110 | 15 | 406,884 | 621,838 |
| OD7 | 110 | 16 | 320,067 | 448,388 |
| OD8 | 106 | 17 | 271,333 | 425,446 |

Table 2.1: Summary of data for the origin-destination pairs OD1 to OD8.

### 2.4.1 The Airline's Pricing Model

As discussed in Appendix B.2, a prediction model for the airline's pricing and a suitable instrument variable are required to account for price endogeneity. As airline revenue management assumes fixed and variable costs to be constant across flights operating on the same route, these cost types are unsuitable instruments. Instead, we propose to use the opportunity cost of capacity as an instrument variable.

Airline revenue management often employs the opportunity cost of capacity as a bidprice for capacity allocation. The bidprice is known for future departure days and varies across bookings, as each booking increases the bidprice to acknowledge scarce capacity. This mechanism introduces simultaneity between demand and price, i.e., price influences demand and vice versa. Furthermore, when flights are part of a transfer itinerary, the bidprice accounts for network effects. Therefore, the bidprice is not entirely determined by the demand for a single flight.

However, this instrument variable only varies if the capacity is scarce. Therefore, it is only useful when demand exceeds capacity. When the natural logarithm of the bidprice is the instrument $\mathrm{IV}_{i,t}$, the functional structure of the first-stage regression equation (6) is:

$$\eta_{i,t} = \theta_0 + \theta_1 \log(\mathrm{IV}_{i,t}) + \sum_{j=1}^{6} (\mathbf{1})_{\mathrm{BDAY}_{i,t}=j} \theta_{1,j} + s_t(t) + s_1(\mathrm{DTIME}_{i,t}) + s_2(\mathrm{YDAY}_{i,t}). \quad (2.4)$$

### 2.4.2 The Airline's Demand Model

The empirical data set records three categorical covariates: DDAY, BDAY, and P2P. Segmenting by the DDAY and P2P induces a full interaction between these variables and any other covariate that describes the demand model (2.2). Therefore, we estimate $7 \times 16 = 112$ separate models at this level. As the categorical variable BDAY is closely associated with the booking time $t$, the model includes it to capture potential changes in booking intensity over the sales horizon. Incorporating BDAY as the only categorical variable (taking BDAY=Sunday as the reference category) yields the first index set as $I_1 = \{1\}$ with $J_1 = \{1, \ldots, 6\}$, where 1 represents Monday, 2, Tuesday,...,6 Saturday.

In addition to price $p_{i,t}$ and booking time $t$, the data records continuous covariates DTIME and YDAY. $I_2 = \{1, 2\}$ defines the second index set. The covariate vector is, therefore $(\mathbf{x})_{i,t} = (\mathrm{BDAY}_{i,t}, p_{i,t}, \mathrm{DTIME}_{i,t}, \mathrm{YDAY}_{i,t}, t)$. Thus, a typical data row for flight $i$ with booking day Monday (reported as 1), ticket price \$50, departure time 6:00 am (reported as 6), and departure date on January 2 (reported as 2) at the day of departure ($t = 0$) is $(\mathbf{x})_{i,t} = (1, 50, 6, 2, 0)$. The airline's demand model is as follows:

$$\begin{aligned} \log\Big(\lambda((\mathbf{x})_{i,t}, t)\Big) &= \beta_0 + \beta_{\widehat{\xi}}\widehat{\xi}_{i,t} + \sum\nolimits_{j=1}^{6} (\mathbf{1})_{\mathrm{BDAY}_{i,t}=j}\beta_{1,j} + f_p\left(p_{i,t}\right) \\ &+ f_{p,t}\big(p_{i,t}, t\big) + f_{p,1}\left(p_{i,t}, \mathrm{DTIME}_{i,t}\right) + f_{p,2}\big(p_{i,t}, \mathrm{YDAY}_{i,t}\big) \\ &+ f_t\left(t\right) + f_1\left(\mathrm{DTIME}_{i,t}\right) + f_2\left(\mathrm{YDAY}_{i,t}\right) \\ &+ f_{t,1}\left(t, \mathrm{DTIME}_{i,t}\right) + f_{t,2}\big(t, \mathrm{YDAY}_{i,t}\big) + f_{1,2}\big(\mathrm{DTIME}_{i,t}, \mathrm{YDAY}_{i,t}\big). \end{aligned} \tag{2.5}$$

Here, coefficient-vector $(\beta)_1 = (\beta_{1,1}, \ldots, \beta_{1,6})$ quantifies the effect of the booking day on the demand intensity.

### 2.4.3 Estimation Results

We look at one route, OD8, and DDAY = Thursday, to demonstrate the estimation results. Figure 2.1 shows the smooth components $s_t(t), s_1(\mathrm{DTIME}_{i,t})$, and $s_2(\mathrm{YDAY}_{i,t})$ from the first-stage. The left panel shows the estimate of $s_t(t)$, which indicates that prices increase over the sales horizon. The estimate of $s_1$(DTIME) (middle panel) shows there exists a strong departure time pattern. Flights departing around 8:00 am and 7:00 pm are the most expensive. The right panel shows the estimate of $s_2$(YDAY). The lack of a strong pattern indicates little price variation over the days of the year.

Figure 2.2 shows second-stage estimates of demand model (2.5). Our analysis concentrates on four aspects. All covariates except the one indicated in the caption are fixed in each panel. The panels on the left reveal how price and demand vary over departure time (a) and year day (c). Panels (b) and (d) show how demand varies independently of the price. Panel (a) indicates that price sensitivity depends on departure time. Peak flights at 7:00 am and 7:00 pm show a smaller price slope than midday flights. Panel (c) accounts for seasonal price sensitivity changes but only shows moderate fluctuations. Panel (b) confirms that demand varies along the time-to-departure axis by departure time. The step-like pattern in panel (b) arises from the influence of BDAY: fewer bookings occur on weekends than on weekdays.

Table 2.2 reports the parameter estimates of the first (2.4) and second-stage (2.5), as well as the estimated standard errors (in brackets). The parameter estimates for the second stage show fewer bookings on weekends than on weekdays. When the second stage considers price endogeneity, there are about ten times more bookings for BDAY=Monday than on the reference category BDAY=Sunday ($\exp(-2.01) \approx 0.13$ to $\exp(-2.01+2.29) \approx$

1.32). For the first stage, BDAY estimates are negative for weekdays and positive for the weekend. Thus, weekdays combine high demand with low prices, while weekends combine low demand with high prices. Apart from the BDAY effect, we also report the estimates of IV (only relevant for the first stage) and first-stage residual $\widehat{\xi}$ (only suitable for the second stage). In the first stage, IV is significantly different from zero (with t-statistic = 0.07/0.0004 = 175), which satisfies one requirement of the two-staged procedure. The parameter estimate for the first-stage residual $\widehat{\xi}$ represents the working parameter of a univariate spline with one inner knot-interval and a monotonicity restriction according to Pya (2010, p. 45). The BDAY estimates show that considering the first-stage residual $\widehat{\xi}$ does not change the parametric coefficients of BDAY. As the second stage, which controls for price endogeneity, shows smaller values for AIC and BIC, the remainder of the paper focuses on the corresponding model setup with endogeneity considered. Appendix D.4 discusses the models' prediction accuracy assessment.

## 2.5 Dynamic Pricing

This section demonstrates the demand model's applicability to dynamic pricing. The resulting pricing leads to an offered price that maximizes the revenue gained from sales.

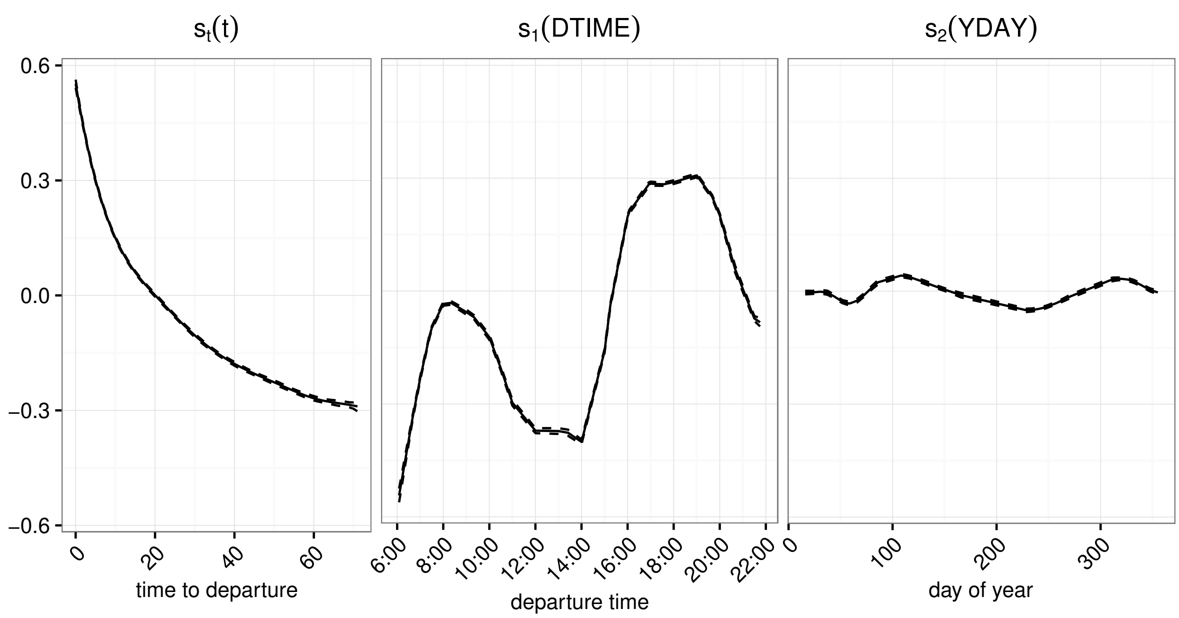



Figure 2.1: Estimate of the three smooth functions $s_t(t), s_1(\text{DTIME})$, and $s_2(\text{YDAY})$ for the first-stage model (2.4). Solid line = estimate, dotted lines = 99% confidence band.

### 2.5.1 Continuous Prices

This paper assumes dynamic pricing with support on $\mathbb{R}_+$ is possible. With revenue gain $r_{i,t}$ from selling a ticket at a price $p_{i,t}$, we calculate the total revenue as $\lambda((\mathbf{x})_{i,t}, t)r_{i,t}$. The opportunity cost of selling capacity to $\lambda((\mathbf{x})_{i,t}, t)$ many passengers is $\lambda((\mathbf{x})_{i,t}, t)\pi_{i,t}$. When the capacity constraint is irrelevant, we can set $\pi_{i,t} = 0 \; \forall \; i, t$. According to equation (2.2), the proposed demand model includes a multiplicatively separable demand rate with expo-

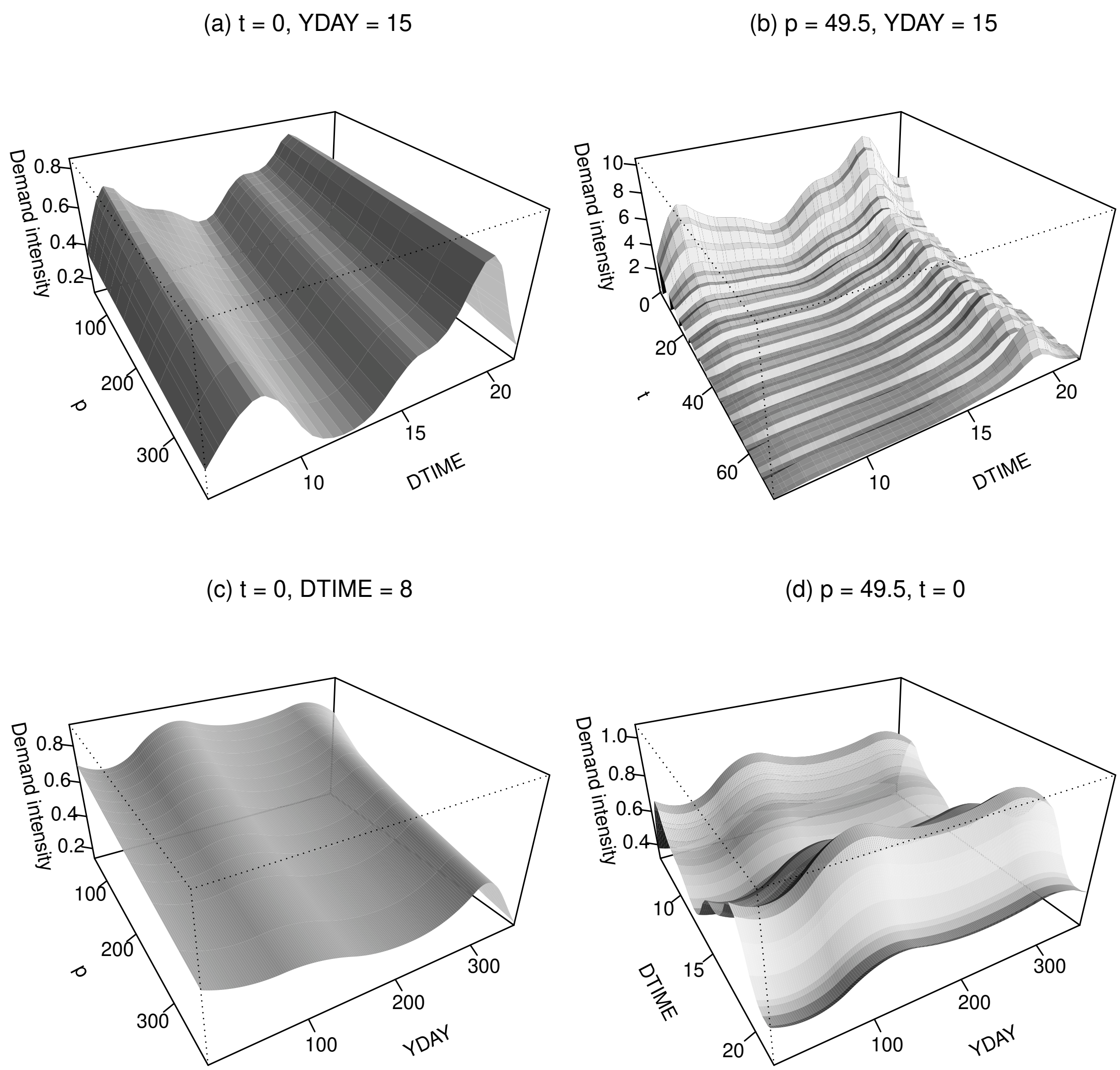



Figure 2.2: Estimates of the conditional demand intensity for OD8 and DDAY = Thursday. The variables appearing in each panel's title are fixed to a specific value (e.g., t is fixed to 0 and YDAY to 15 for panel (a). Every panel shows how the demand intensity (vertical axis) changes within the covariates that are shown at the horizontal axis (e.g. the demand intensity within the panel (a) changes in departure time and peaks at 8:00 am and 7.00 pm).

| | Stage | | |
|---|---|---|---|
| | First | Second | |
| | | Endogeneity considered | |
| | | yes | no |
| Intercept | 4.8223(0.0022) | −2.0104(0.3877) | −1.8381(0.1084) |
| log(IV) | 0.0700(0.0004) | - | - |
| $\widehat{\xi}$ | - | −0.0020(0.0013) | - |
| BDAY = Monday | −0.0253(0.0025) | 2.2966(0.1123) | 2.2946(0.0358) |
| BDAY = Tuesday | −0.0293(0.0025) | 2.2116(0.1311) | 2.2079(0.0452) |
| BDAY = Wednesday | −0.0263(0.0025) | 2.2363(0.1406) | 2.2326(0.0498) |
| BDAY = Thursday | −0.0224(0.0025) | 2.1077(0.1366) | 2.1019(0.0504) |
| BDAY = Friday | −0.0129(0.0025) | 2.1083(0.1273) | 2.1050(0.0460) |
| BDAY = Saturday | 0.0064(0.0028) | −0.3627(0.1692) | −0.3627(0.0516) |
| AIC | 1457362 | 106957.9 | 107131.0 |
| BIC | 4741586 | 107920.4 | 108034.5 |

Table 2.2: Parameter estimates for OD8 and DDAY=Thursday. Standard errors in brackets.

nential price sensitivity. From this, we derive the optimal price $p^{\star}_{i,t}$ by solving

$$\max_{p_{i,t}} \Big\{\lambda((\mathbf{x})_{i,t}, t) r_{i,t} - \lambda((\mathbf{x})_{i,t}, t)\pi_{i,t}\Big\} \Leftrightarrow \frac{\partial\Big(\lambda((\mathbf{x})_{i,t}, t) r_{i,t} - \lambda((\mathbf{x})_{i,t}, t)\pi_{i,t}\Big)}{\partial p_{i,t}} \overset{!}{=} 0. \tag{2.6}$$

To efficiently calculate the derivative of $\lambda((\mathbf{x})_{i,t}, t)$ concerning $p_{i,t}$, we can take advantage of the fact that the derivative of a B-spline is a linear combination of lower order B-splines (Marsh and Marshall, 1999). Solving (2.6), i.e., maximizing the total revenue gain over cost, yields the optimal value for $r_{i,t}$, say $r^{\star}_{i,t}$. To calculate the corresponding optimal price-value $p^{\star}_{i,t}$, we have to derive the amount of variable cost (including fuel, onboard service, and taxes) that has to be added to $r_{i,t}$ to achieve price $p_{i,t}$ (the amount is defined by the difference $p_{i,t} - r_{i,t}$). We do so by solving the regression problem

$$\begin{aligned} p_{i,t} - r_{i,t} &= \alpha_0 + \alpha_1 p_{i,t} + \epsilon_{i,t} \\ r_{i,t} &= \underbrace{-\alpha_0}_{\equiv\gamma_0} - \underbrace{(1-\alpha_1)}_{\equiv\gamma_1} p_{i,t} + \epsilon_{i,t}. \end{aligned} \tag{2.7}$$

The difference between $r^{\star}_{i,t}$ and $p^{\star}_{i,t}$ is determined by $\gamma_0 = -\alpha_0$ and $\gamma_1 = 1 - \alpha_1$. The parameter $\alpha_0$ represents the constant variable cost amount that does not depend on the ticket price, e.g., fuel or onboard service. The parameter $\alpha_1$ gives the variable cost factor that increases the price, e.g., taxes. If model (2.3) is linear in $p_{i,t}$, the maximization problem (2.6) has the closed-form solution

$$p^{\star}_{i,t} = \underbrace{-\frac{1}{(\mathbf{1})_{2\mathrm{s}}\beta_{\widehat{\xi}} + f'_p + f'_{p,t}(t) + \sum_{k\in I_2} f'_{p,k}(z_{k,i,t})}}_{\text{profit margin}} + \underbrace{\frac{\alpha_0}{1-\alpha_1} + \frac{\pi_{i,t}}{1-\alpha_1}}_{\text{cost margin}}. \tag{2.8}$$

A detailed description of how to derive (2.8) by solving (2.6) is found in Appendix F.6. Here, $f'_p, f'_{p,t}(t)$, and $f'_{p,k}(z_{k,i,t})$ correspond to the derivative of $f_p, f_{p,t}(t), f_{p,k}(z_{k,i,t}), k \in I_2$ with respect to $p_{i,t}$, e.g., $f'_p = \frac{\partial f_p(p_{i,t})}{\partial p_{i,t}}$. Assuming linearity of $p_{i,t}$, $f'_p$ is only a scalar. Factor $\beta_{\widehat{\xi}}$ only appears in (2.8) if $(\mathbf{1})_{2s}$ indicates that price endogeneity is considered, where $(\mathbf{1})_{2s} = 1$ if the estimation is two-staged and $(\mathbf{1})_{2s} = 0$ otherwise. The influence of $\widehat{\xi}$ is constrained such that $\beta_{\widehat{\xi}} < 0$.

By monotonicity of (2.2) within $p_{i,t}$, every derivative $f'_p, f'_{p,t}(t), f'_{p,k}(z_{k,i,t})$ is strictly negative, rendering the profit margin positive. The smooth components $f_t, f_k, f_{t,k}, f_{k_1,k_2},\ k, k_1, k_2 \in I_2, k_1 < k_2$ and parametric parts $x_{k,i,t}, (k \in I_2)$ are not relevant to price-sensitivity and vanish when the derivative of $p_{i,t}$ is calculated. Thus $\frac{\partial \lambda((\mathbf{x})_{i,t},t)}{\partial p_{i,t}} = \lambda'((\mathbf{x})_{i,t}, t) = (\mathbf{1})_{2s}\beta_{\widehat{\xi}} + f'_p + f'_{p,t}(t) + \sum_{k \in I_2} f'_{p,k}(z_{k,i,t}) < 0$.

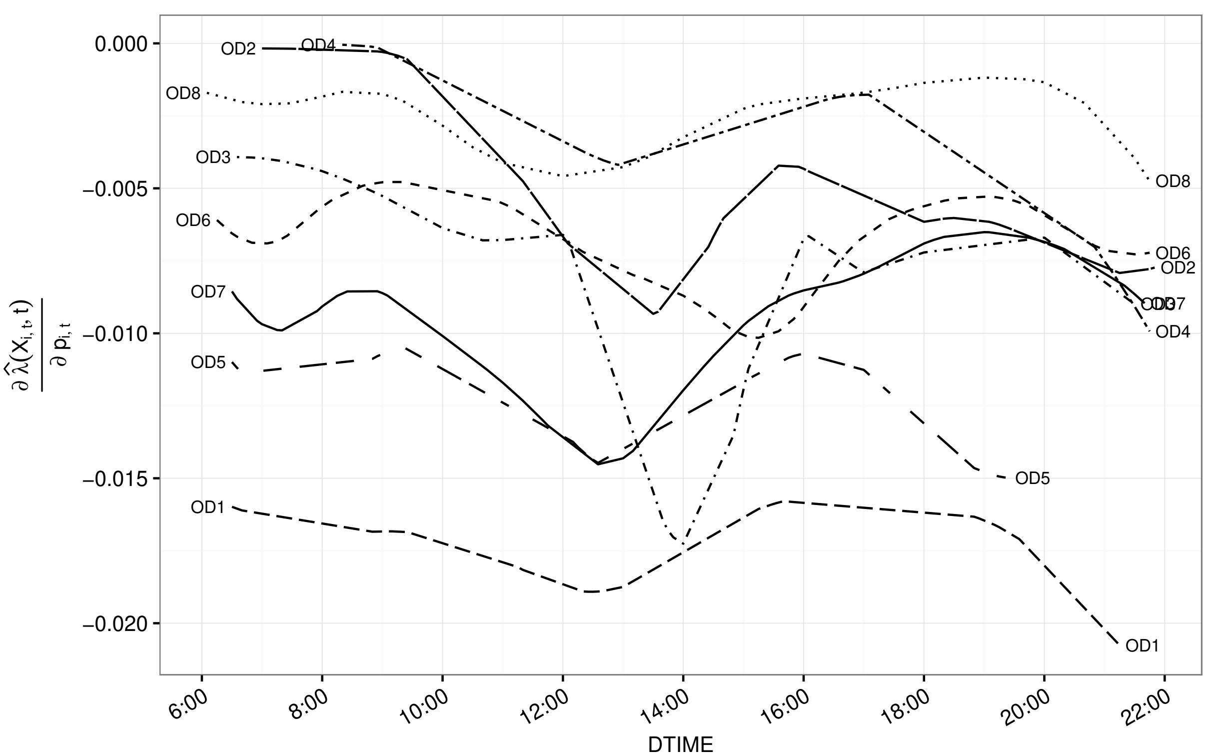



Figure 2.3: Estimated derivative $\widehat{\lambda}'((\mathbf{x})_{i,t}, t)$ for DDAY=Thursday, t=0 and YDAY=15 for different DTIME values. Each line shows the changes of the derivative $\widehat{\lambda}'((\mathbf{x})_{i,t}, t)$ for a specific OD.

Figure 2.3 shows how each OD's derivative $\lambda'((\mathbf{x})_{i,t}, t)$ evolves over DTIME for DDAY=Thursday, $t = 0$, and YDAY=15. Note that decreasing price-sensitivity goes along with increasing $\lambda'((\mathbf{x})_{i,t}, t)$. Aside from the amplitude of the pattern, every OD experiences a decline in price-sensitivity during the morning and evening ($|\lambda'((\mathbf{x})_{i,t}, t)|$ becomes smaller), meaning that passengers are willing to pay more if traveling during these hours of the day.

As (2.8) shows, $p^{\star}_{i,t}$ is the sum of the profit margin and cost. Whenever price-sensitivity decreases, the value of $\lambda'((\mathbf{x})_{i,t}, t)$ increases. Therefore, the optimal price $p^{\star}_{i,t}$ increases by the increasing value of the profit margin. Thus, the offered price $p^{\star}_{i,t}$ inversely depends on the estimated price sensitivity. The lower bound of $p^{\star}_{i,t}$ is the cost of production. Depending on the application, this bound is also a function of opportunity cost $\pi_{i,t}$.

Figure 2.4 shows $p^{\star}_{i,t}$ for $t = 0$, DDAY=Thursday, and day of the year 15 for OD6 and OD8 with opportunity cost $\pi_{i,t} \in \{1, 20, 100\}$. Appendix G.7 induces supplementary illustrations for the other OD. The general shape of each curve is defined by the derivative function

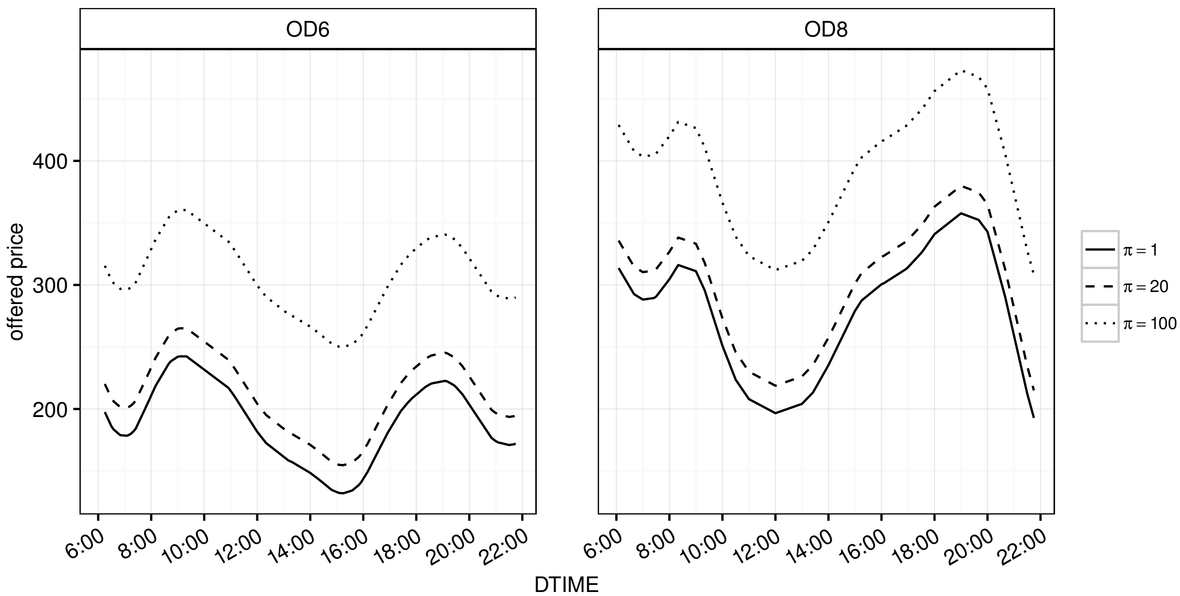



Figure 2.4: Optimal price values for OD6, OD8 at DDAY=Thursday for opportunity cost $\pi_{i,t} = 1, 20, 100$.

$\lambda'((\mathbf{x})_{i,t}, t)$ as shown by Figure 2.3. For OD6, optimal price $p^{\star}_{i,t}$ increases if the value of the derivative function $\lambda'((\mathbf{x})_{i,t}, t)$ increases, indicating a decrease in price-sensitivity. This mechanic leads to a pricing policy where prices inversely depend on price sensitivity, i.e., the price increases if the price sensitivity decreases and vice versa. For OD8, this yields higher prices for peak flights, as an increase within the derivative function $\lambda'((\mathbf{x})_{i,t}, t)$ indicates a decrease in price sensitivity. Additionally, a change in the opportunity cost $\pi$ shifts the entire curve as the cost of selling a seat increases, regardless of price sensitivity.

## 2.5.2 Discrete Prices

For most airlines, technological hurdles still call for offering prices from a countable and finite set of discrete values $\Omega_p = \{p_1, \ldots, p_J\}$ with revenue gain $r_j,\ j \in \{1, \ldots, J\}$ (compare Fiig et al., 2015). As suggested by Fiig et al. (2010), we also propose only to offer the subset of prices $\Omega'_p \subset \Omega_p$ for which the marginal revenue gain exceeds the marginal cost (bid-price). As it turns out (proof given in Appendix E.5), the solution of the continuous problem (2.8)

| | OD1,OD2, and OD4 | | | | | | | OD3 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Outbound direction | | | | | | | |
| CW | Mo | Tu | We | Th | Fr | Sa | Su | Mo | Tu | We | Th | Fr | Sa | Su |
| 25 | o | x | o | x | o | x | o | x | o | x | o | x | o | x |
| 24 | x | o | x | o | x | o | x | o | x | o | x | o | x | o |
| 23 | o | x | o | x | o | x | o | x | o | x | o | x | o | x |
| 22 | x | o | x | o | x | o | x | o | x | o | x | o | x | o |
| | | | | | | | Inbound direction | | | | | | | |
| 25 | x | o | x | o | x | o | x | o | x | o | x | o | x | o |
| 24 | o | x | o | x | o | x | o | x | o | x | o | x | o | x |
| 23 | x | o | x | o | x | o | x | o | x | o | x | o | x | o |
| 22 | o | x | o | x | o | x | o | x | o | x | o | x | o | x |

Table 2.3: Layout of the test-pattern. Abbreviations: x = influenced, o = non-influenced.

also defines the boundary point for price points that belong to the set $\Omega'_p$. Therefore, it is optimal to offer the next highest price point to the optimal price $p^{\star}_{i,t}$.

### 2.5.3 Performance Evaluation in a Field Study

In cooperation with a European network airline, we test our model in a field study. To that end, we choose four OD to reflect specific market characteristics, e.g., the dominance of business vs. leisure travelers or high versus low competition. OD1 and OD4 represent leisure routes, whereas OD2 and OD3 represent business routes. Applying our approach to every combination of OD, DDAY, and P2P created $4 \times 7 \times 2 = 56$ separate model estimates. For every model, we offer (given the bidprice) the next highest price, the optimal price $p^{\star}_{i,t}$.

The field study includes departure dates from 2016-05-30 until 2016-06-26. For every flight within that departure period and booking dates between 2016-04-04 and 2016-05-29, the available price is chosen according to the optimal price $p^{\star}_{i,t}$ (see 2.5.2). Table 2.3 describes the test pattern.

The checkerboard pattern aims to capture and control systematic differences in DDAY and OD directions. The airline's regular revenue management system (non-influenced) calculates the bidprices and lets market analysts set the prices. Both the traditional approach and our model relied on the same bidprice information. Thus, the observed revenue difference originates from the passengers' price sensitivity estimates. The overall revenue gain is aggregated separately for influenced and non-influenced departure days for each OD. The difference is given in percent of the regular system's performance. As the results in Table 2.4 show, our approach successfully increased the revenue between 1.64% to 8.03% and the margin by 11.08% to 36.39% across markets. The margin defines the gaps between revenue earned and opportunity cost $\pi$. Due to the positive outcome, the airline plans to apply our approach to every OD of the European network by the end of

| Market | OD1 | OD2 | OD3 | OD4 | overall |
|---|---|---|---|---|---|
| Revenue | 1.64% | 6.31% | 8.44% | 8.03% | 6.37% |
| Margin | 36.39% | 18.96% | 13.91% | 11.08% | 15.36% |

Table 2.4: Results of the live test in percent

2017.

## 2.6 Model adjustments due to market turbulences

Since 2017 and due to the success of the field study, see Section 2.5.3, AirABC decided to increase the models' scope to about 360 ODs. In May 2019, an intervention analysis examining the revenue impact of the models' rollout resulted in an average revenue increase of +2.37%, confirming the field study's positive result. Since the models' rollout, the price elasticity estimates have regularly been updated using the past two departure years. Even though frequent updates of the demand model would eventually catch up with a changing market environment, due to the functional structure of the predictor (2.5), the price-elasticity estimates for the next selling date only differ from the previous selling date by variables that describe the flight (DTIME, YDAY) and the time to departure $t$. Specifically, structural changes within the market happening from one selling date to another cannot be captured accurately. To allow for more flexibility for demand- and price-elasticity predictions to describe changes in the market environment, four additional covariates are introduced:

1. Booking day of the year (BYDAY), taking values from $1, \dots, 365$
2. Selling Date Index (SDI) as
   (a) SDI = -630 for April 12, 2019
   (b) ...
   (c) SDI = -2 for December 30, 2019
   (d) SDI = -1 for December 31, 2019
   (e) SDI = 0 (ease of interpretation) for the selling date January 1, 2020
   (f) SDI = 1 for January 2, 2020
   (g) SDI = 2 for January 3, 2020
   (h) ...
   (i) SDI = 100 for April 11, 2021
3. Daily new cases of COVID-19 within the country of the origin airport (OCOV)

4. Daily new cases of COVID-19 within the country of the destination airport (DCOV)

The data for daily new cases of COVID-19 is gathered from (OWID). Figure 2.5 shows how the daily new cases of COVID-19 evolved over the period 2020-01-24 to 2021-04-11. For every country, the first wave of COVID-19 cases occurred from 2020-03-01 to 2020-05-01 and was followed by another wave starting between 2020-08-01 and 2020-10-01. Within 2021-05-01 and 2021-08-01, the number of cases was reduced by severe lock-down measures, which were (depending on the country) partially lifted during the summer months, i.e., between 2020-06-01 to 2020-08-01 and re-implemented afterward. The adjustment of model (2.5), including the additional covariates (in black), is defined by Equation (2.9). Table 2.5 describes the nine functions in more detail.

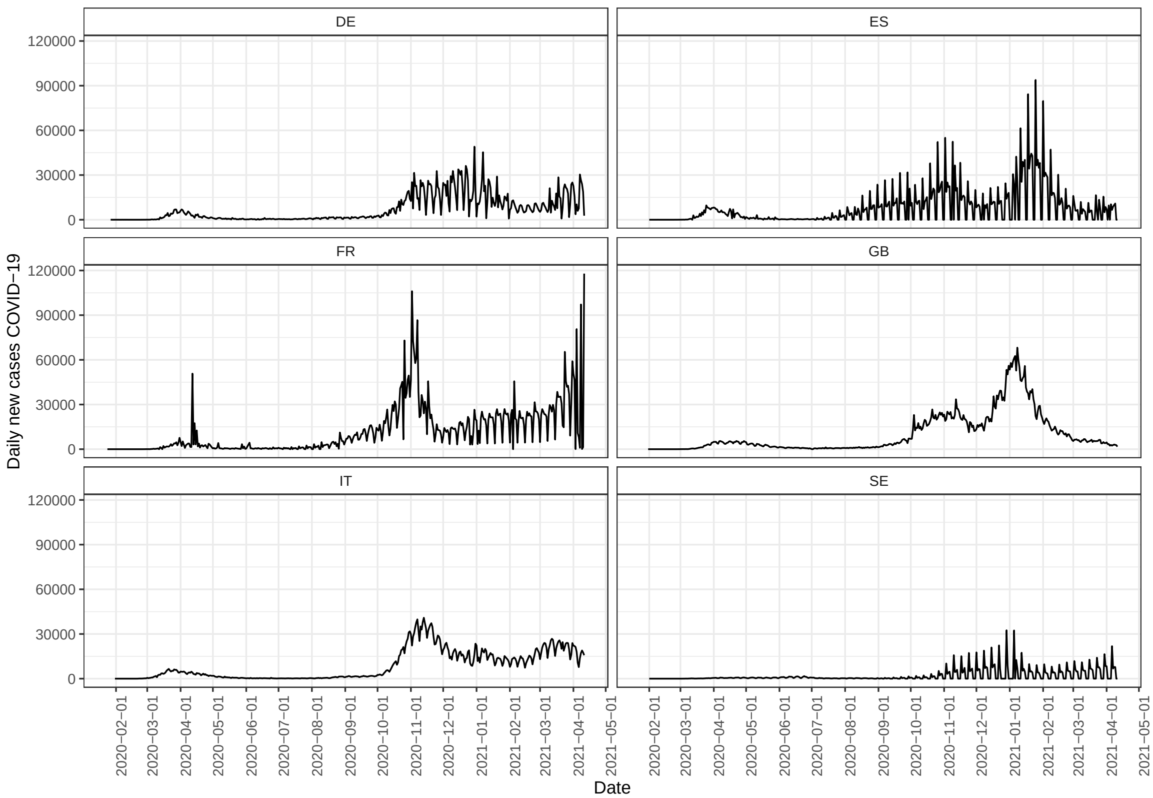



Figure 2.5: Daily new cases of COVID-19 for the countries Germany(DE), Spain (ES), France (ES), United Kingdom (GB), Italy (IT), and Sweden (SE) from 2020-01-24 to 2021-04-11.

$$
\begin{aligned}
\log\Big(\lambda((\mathbf{x})_{i,t},t)\Big)^{\text{adj}} &= \log\Big(\lambda((\mathbf{x})_{i,t},t)\Big) \\
&+ f_3\left(\text{BYDAY}_{i,t}\right) + f_4\left(\text{SDI}_{i,t}\right) + f_5\left(\text{OCOV}_{i,t}\right) + f_6\left(\text{DCOV}_{i,t}\right) \\
&+ f_{p,3}\left(p_{i,t},\text{BYDAY}_{i,t}\right) + f_{p,4}\left(p_{i,t},\text{SDI}_{i,t}\right) + f_{p,5}\left(p_{i,t},\text{OCOV}_{i,t}\right) + f_{p,6}\left(p_{i,t},\text{DCOV}_{i,t}\right) \\
&+ f_{t,5}\left(t,\text{SDI}_{i,t}\right).
\end{aligned} \tag{2.9}
$$

where $\log\Big(\lambda((\mathbf{x})_{i,t}, t)\Big)$ corresponds to the (unadjusted) model (2.3). Note that the adjusted model (2.9) is different from the full factorial design of the (general) model (2.3). As the research goal of this section focuses on an aggregated level, all bivariate functions between the covariates DTIME, YDAY, OCOV, DCOV, and SDI are omitted. Not all combinations of continuous covariates are included, as the adjustment intends to answer the following questions:

Q1: Does the impact of COVID-19 on passenger demand change over time, i.e., is it different from one COVID-19 wave to another?

Q2: Does the impact of COVID-19 on passenger price elasticity change over time, i.e., is it different from one COVID-19 wave to another?

Q3: Do passengers book more spontaneously, i.e., closer to departure than pre-COVID-19 times?

The two functions $f_3\left(\text{BYDAY}_{i,t}\right)$ and $f_{p,3}\left(p_{i,t}, \text{BYDAY}_{i,t}\right)$ are added to ensure that the impact of the selling-date-index (SDI) on passenger demand and price elasticity may not be affected by a booking seasonality.

**Data Adjustments**:

To ensure that the estimations for the departure date seasonality are unbiased, the results of section 2.4.3 depend on data where the entire booking period is observable for every flight. Suppose data for flights with an incomplete booking period is considered. In that case, demand estimates for future departure months can be lower, which may give a wrong

| Function | Description |
|---|---|
| $f_{p,3}\left(p_{i,t}, \text{BYDAY}_{i,t}\right)$ | Booking yearday seasonality of price elastictiy |
| $f_3\left(\text{BYDAY}_{i,t}\right)$ | Booking yearday seasonality of demand |
| $f_{p,4}\left(p_{i,t}, \text{SDI}_{i,t}\right)$ | Daily change in price elasticity |
| $f_4\left(\text{SDI}_{i,t}\right)$ | Daily demand changes |
| $f_{p,5}\left(p_{i,t}, \text{OCOV}_{i,t}\right)$ | Change of price elasticity by daily new COVID19 cases at the origin airport |
| $f_5\left(\text{OCOV}_{i,t}\right)$ | Demand changes by daily new COVID19 cases at the origin airport |
| $f_{p,6}\left(p_{i,t}, \text{DCOV}_{i,t}\right)$ | Change of price elasticity by daily new COVID19 cases at the destination airport |
| $f_6\left(\text{DCOV}_{i,t}\right)$ | Demand changes by daily new COVID19 cases at the destination airport |
| $f_{t,5}\left(t, \text{SDI}_{i,t}\right)$ | Daily change of the demand booking curve |

Table 2.5: Additional model functions and covariates allowing for more flexibility for demand- and price-elasticity predictions

impression of how demand changes during the departure year. The models' demand predic-

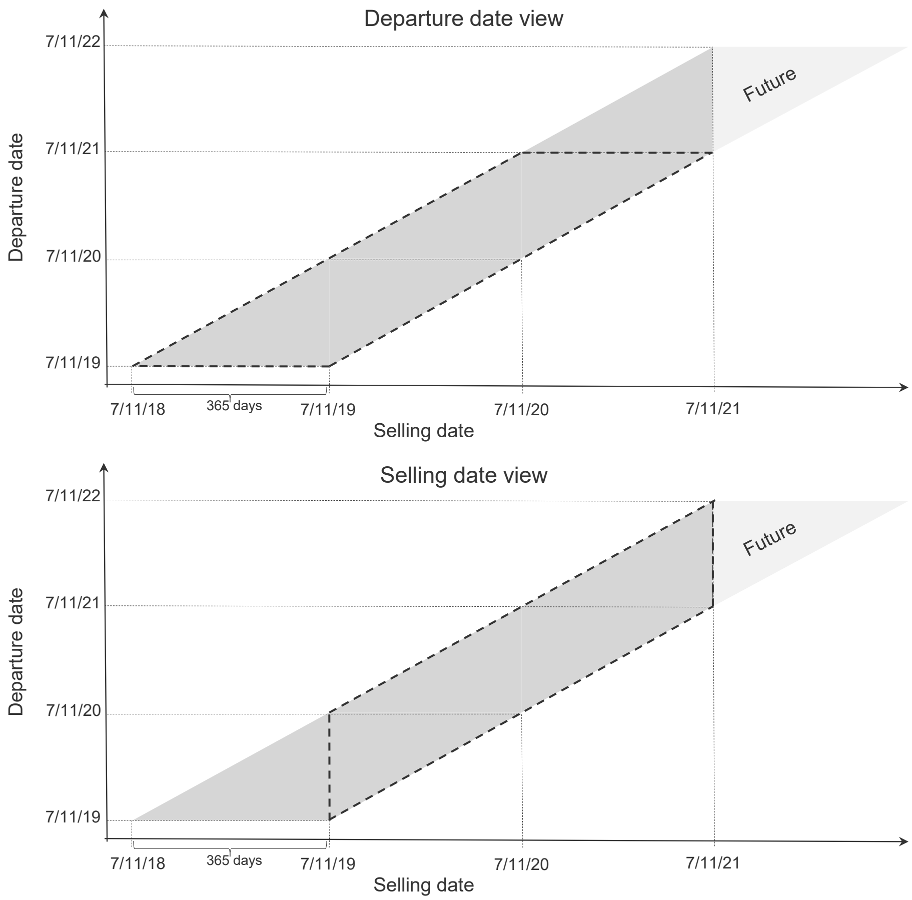



Figure 2.6: Assuming today's estimation date is 7/11/21, the top graph describes the available data history for the departure date view. The departure dates range from 7/11/18 until 7/11/19, each with a 365-day selling date history. The dotted area of the bottom graph describes the data for the selling date view, where each selling date includes the booking history of past and future departure dates.

tions using post-COVID-19 data no longer apply to the market environment of COVID-19 times. As the models' ability to react to recent changes in previous selling dates depends on the input data, the data scope changes from a departure-date view to a selling-date view. The departure-date view (the dotted area within the top graph of Figure 2.6) defines the data history of flights with complete selling-date history. The selling-date view (dotted area of the bottom graph of Figure 2.6) describes the data set, where every selling date contains the entire (future) departure-date history. Assuming today's estimation date is 7/11/21, Figure 2.6 shows that the data for departure date 7/11/22 is not available for the departure-date view until the models' estimation date reaches 7/11/22, whereas the selling-date view includes already all future fights until 7/11/22. Therefore, the benefit of having more up-to-date data and an accurate booking-behaviour representation (as captured by the functions $f_3\left(\mathrm{BYDAY}_{i,t}\right), f_{p,3}\left(p_{i,t}, \mathrm{BYDAY}_{i,t}\right), f_4\left(\mathrm{SDI}_{i,t}\right), f_{p,4}\left(p_{i,t}, \mathrm{SDI}_{i,t}\right)$, and $f_{t,5}\left(t, \mathrm{SDI}_{i,t}\right))$ outweighs the risk of having a biased seasonality estimate for departure dates.

**Results**:

To answer the questions Q1, Q2, and Q3, the exp() is applied to both sides of the adjusted equation of model (2.9), which gives

$$\begin{aligned}
&\lambda((\mathbf{x})_{i,t}, t)^{\mathrm{adj}} = \lambda((\mathbf{x})_{i,t}, t) \\
&\times \underbrace{\exp\Big(f_3\left(\mathrm{BYDAY}_{i,t}\right) + f_4\left(\mathrm{SDI}_{i,t}\right) + f_5\left(\mathrm{OCOV}_{i,t}\right) + f_6\left(\mathrm{DCOV}_{i,t}\right)\Big)}_{\mathrm{Q1}^f} \\
&\times \underbrace{\exp\Big(f_{p,3}\left(p_{i,t}, \mathrm{BYDAY}_{i,t}\right) + f_{p,4}\left(p_{i,t}, \mathrm{SDI}_{i,t}\right) + f_{p,5}\left(p_{i,t}, \mathrm{OCOV}_{i,t}\right) + f_{p,6}\left(p_{i,t}, \mathrm{DCOV}_{i,t}\right)\Big)}_{\mathrm{Q2}^f} \\
&\times \underbrace{\exp\Big(f_{t,5}\left(t, \mathrm{SDI}_{i,t}\right)\Big)}_{\mathrm{Q3}^f}.
\end{aligned} \tag{2.10}$$

The underlined elements $\mathrm{Q1}^f, \mathrm{Q2}^f$, and $\mathrm{Q3}^f$ decrease or increase the unadjusted demand model (2.3), i.e., $\log\Big(\lambda((\mathbf{x})_{i,t}, t)\Big)$ by a factor that depends on the additional covariates BYDAY, SDI, OCOV, and DCOV. Figure 2.7 shows how the factor $\mathrm{Q1}^f$ changes the demand over the selling dates ranging from 2019-02-01 to 2021-04-11. As for the model (2.3), we segment the data by DDAY and P2P. Every line shows the $\mathrm{Q1}^f$ estimate for a specific DDAY and P2P value. To answer the research question Q1, a steep drop in demand starting 2020-03-01, shortly before the first wave of COVID-19, is observed for every country. For some countries, specifically DE and GB, there is an increase in demand during the summer months between 2020-05-01 and 2020-09-01. For GB, we can observe a demand increase during the winter period. Therefore, depending on regulations and restrictions, every country reacts differently and highlights the importance of capturing these effects within the demand model.

Figure 2.8 describes how the factor $\frac{\partial \log(Q2^f)}{\partial p_{i,t}}$, i.e., the price derivative of the logarithm of the factor $\mathrm{Q2}^f$ changes the price elasticity over the selling dates ranging from 2019-02-01 to 2021-04-1. Suppose the adjusted demand model is used to derive optimal price values described in section 2.5. In that case, the factor appears within the denominator of the equation (2.8) and makes the profit margin depend on the additional covariates BYDAY, SDI, OCOV, and DCOV. The graph shows no significant changes in price elasticity, i.e., there is no visible trend and no change of price elasticity during the periods of first wave 2020-03-01 to 2021-05-01 (first wave) and 2020-08-01 and 2020-10-01 (second wave) implying that there is no connection to the number of new cases of COVID-19. Particularly for the countries DE and GB that show an increase in demand during the summer

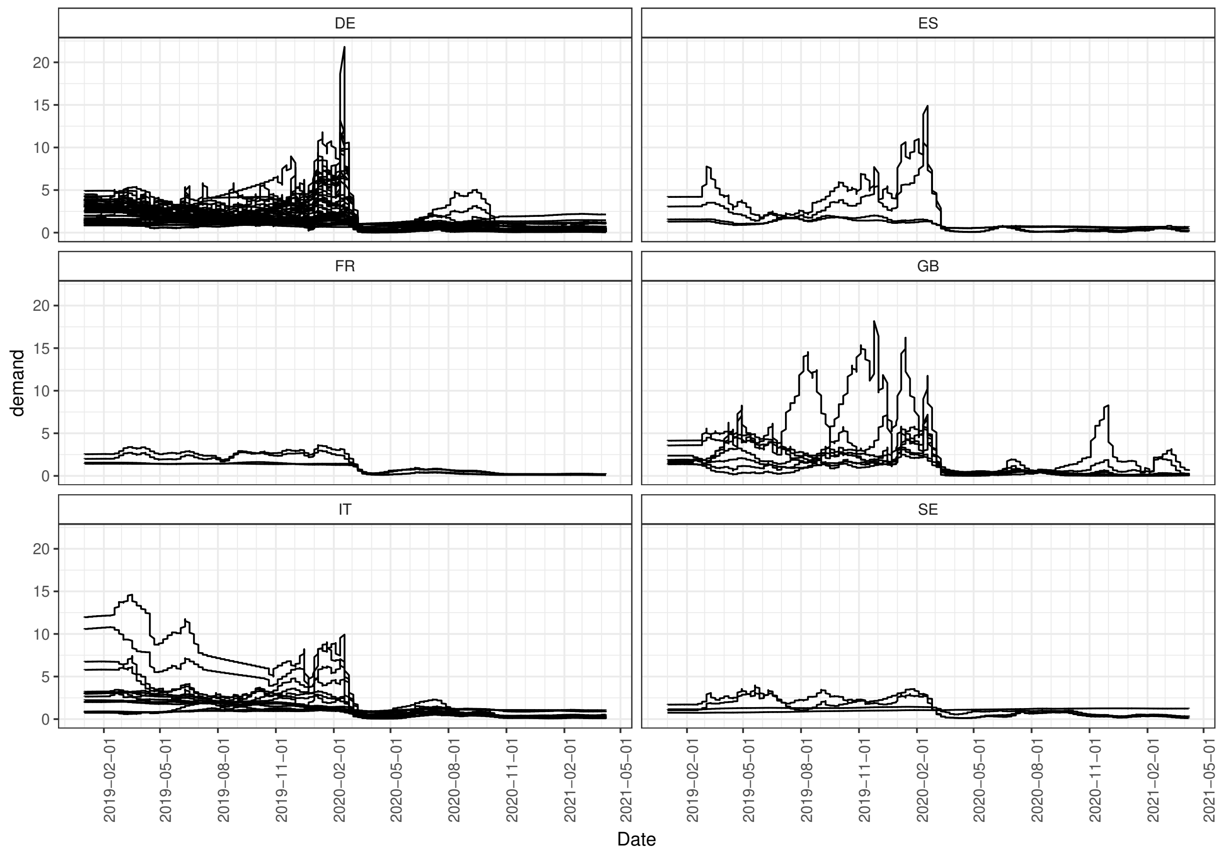



Figure 2.7: For the countries DE, ES, FR, GB, IT, and SE, every line corresponds to an estimate of the factor $Q1^f$ from the adjusted model (2.3) for a specific DDAY and P2P value for selling dates ranging from 2019-02-01 to 2021-04-11

months between 2020-05-01 and 2020-09-01, no visible change in price elasticity is observed. Therefore, there is no evidence that the passengers' price elasticity has changed over time between the two waves of COVID-19. Moreover, these results imply that travel will likely not recover by offering cheaper tickets but by conditions that ensure travel safety.

Finally, Figure 2.9 describes the change in the booking behavior over a 100-day booking period before departure from January 2021 to April 2021 for DE. The countries' ES, FR, GB, IT, and SE figures are in Appendix G.7. Every graph shows that the booking behavior of passengers has drastically changed in every country since the beginning of 2020. Whereas passenger demand steadily increased towards the day of departure, the booking curve is flat from 2020 until the end of the observed data range of April 2021. With the observations for 2.7 where a demand increase is observed for a certain period during the COVID-19 pandemic, airlines can no longer assume that passengers booking behavior follows pre-COVID-19 patterns. Passenger demand arrives not necessarily spontaneously, with a booking curve only increasing shortly before departure and otherwise flat but sporadically throughout the booking period. This effect may be emphasized by the airlines'

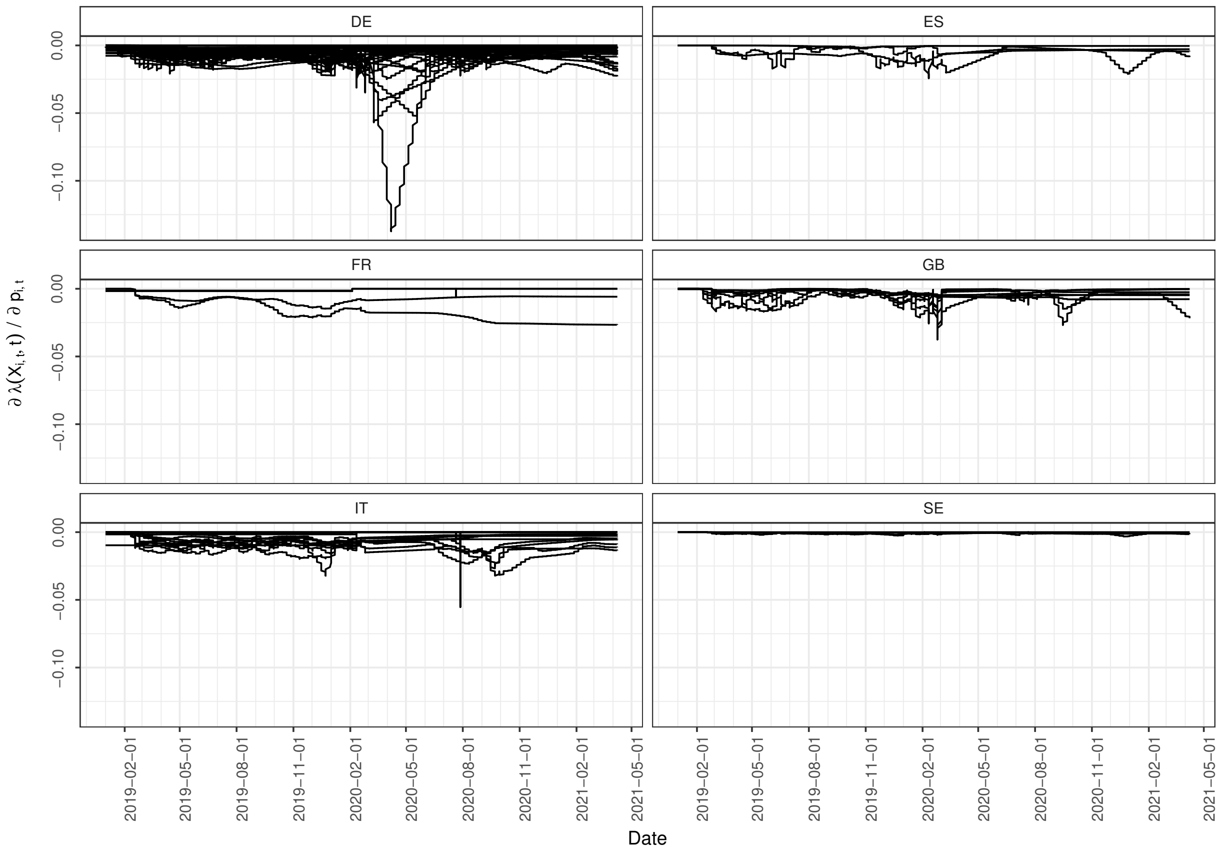



Figure 2.8: For the countries DE, ES, FR, GB, IT, and SE, every line corresponds to an estimate of the factor $Q2^f$ from the adjusted model (2.3) for a specific DDAY and P2P value for selling dates ranging from 2019-02-01 to 2021-04-11

offering cheap tickets together with a waiver of cancellation and rebooking constraints to counter the increased uncertainty of passengers.

## 2.7 Conclusion

This paper introduced a demand estimation approach combining smoothing spline ANOVA interaction components and monotonicity constraints. The functional structure of the proposed model predicts demand by nonparametric and bivariate functions, which dynamically detect and predicts changes in price sensitivity. We derived a closed-form solution to determine the optimal price as the sum of costs and profit margin. This setting is beneficial when the offered price depends on a profit margin and a cost part, which are evaluated independently. The resulting approach efficiently limits the computational effort of dynamic pricing.

To demonstrate and test the proposed approach, we applied it in an airline revenue management setting. Using the opportunity cost of capacity as an instrument variable, we

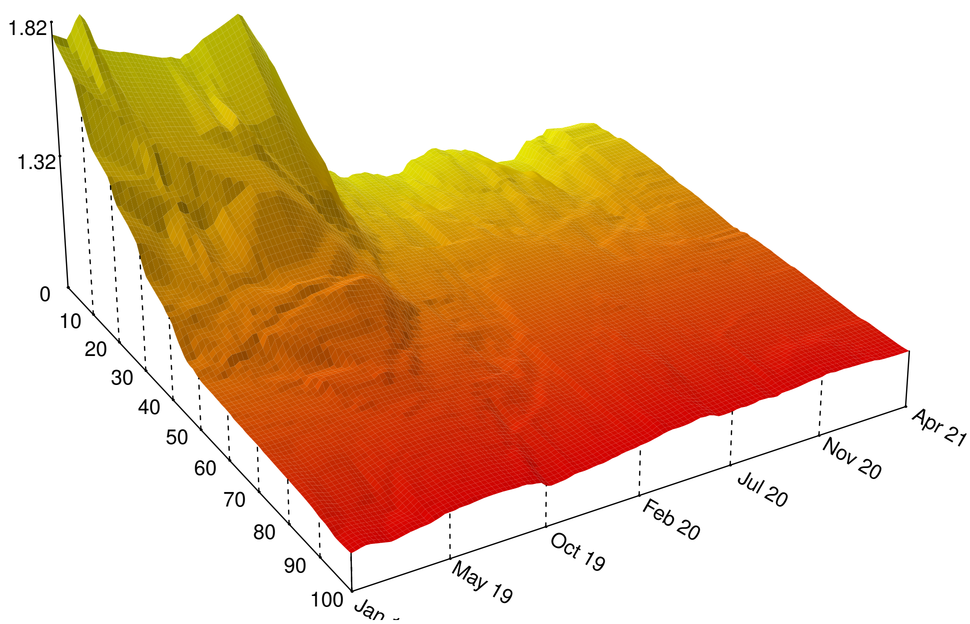



Figure 2.9: For the country DE, the graph shows how the factor $Q3^f$ changes for different days to departure, ranging from 100 days pre-departure to the day of departure and over the selling dates from January 2019 to April 2021.

employed the two-staged approach of Marra and Radice (2011) to account for price endogeneity. A benchmarking study favourably compared the performance of our approach to that state-of-the-art demand estimates. A field study demonstrated a revenue gain between 1.64% to 8.44% and an increase in the margin between 11.08% and 36.39%. Furthermore, we highlight the importance and flexibility of the proposed approach by adjusting the models' structure to consider changing market conditions caused by the COVID-19 pandemic. The model estimates suggest that the behaviour of passenger demand has changed drastically during the COVID-19 pandemic; specifically, bookings occur more sporadically.

As the following steps, we recommend investigating whether possible dependencies among customer segments could improve prediction accuracy. For segmentation, one could examine model-based partitioning (e.g. Zeileis et al., 2008). On another note, we explicitly excluded bookings of fares with fewer restrictions than the lowest available fare. As such features imply upsell, considering upsell-related passenger behaviour may further improve the revenue gain from dynamic pricing.

# Chapter 3

# Interpretable Modeling of Retail Demand and Price Elasticity for Passenger Flights using Booking Data

**Contributing article**:
Meyer JF, Kauermann G, Smith MS. Interpretable modelling of retail demand and price elasticity for passenger flights using booking data. Statistical Modelling. 2022;0(0).
doi:10.1177/1471082X221083343,
available at https://journals.sagepub.com/doi/full/10.1177/1471082X221083343.

**Author contributions**:
Jan Felix Meyer conceived the research question and was responsible for the data management and implementation of the Mixture Model. He prepared the first draft, including working examples and visualisations. Göran Kauermann and Michael Stanley Smith provided valuable inputs to all sections of the article. Michael Stanley Smith conducted the implementation for the Copula Model. All author revised and and proofread the manuscript.

## 3.1 Abstract

We propose a model of retail demand for air travel and ticket price elasticity at the daily booking and individual flight level. Daily bookings are modeled as a nonhomogeneous Poisson process with respect to the time to departure. The booking intensity is a function of booking and flight level covariates, including nonlinear effects modeled semi-parametrically using penalized splines. Customer heterogeneity is incorporated using a finite mixture model, where the latent segments have covariate-dependent probabilities. We fit the model to a unique dataset of over one million daily counts of bookings for 9,602 scheduled flights on a short-haul route over two years. A control variate approach with a strong instrument corrects for a substantial level of price endogeneity. A rich latent segmentation is uncovered, along with strong covariate effects. The calibrated model can be used to quantify demand and price elasticity for different flights booked on different days prior to departure and is a step towards continuous pricing; something that is a major objective of airlines. As our model is interpretable, forecasts can be created under different scenarios. For instance, while our model is calibrated on data collected prior to COVID-19, many of the empirical insights are likely to remain valid as air travel recovers in the post-COVID-19 period.

## 3.2 Introduction and Literature Review

The International Air Transport Association (IATA) estimates that in 2019 there were over 4.54 billion passengers on scheduled flights worldwide, generating revenues of $838 billion dollars (IATA, b). However, profits in the airline industry were notoriously low, even before the advent of COVID-19. For example, the industry average net margin was only 3.1% in 2019 (IATA, b). This forces airlines to seek ever greater competitiveness, including the development of improved revenue management methodologies (Talluri and van Ryzin, 2005). Increasing the accuracy of short term forecasts of passenger demand, along with estimates of its price elasticity, is one such operational efficiency. In particular, the availability of complete booking databases opens up the possibility of computing both demand forecasts and price elasticities for each individual flight and cabin class in real time. Yet, there is surprisingly little work in the statistical or econometric literatures on the modeling of passenger demand at such a disaggregate level — in part because the databases required are large, complex and proprietary. In this paper, we do so using a novel flexible statistical model, which we apply to a new and unique dataset of 1,333,712 daily counts of retail bookings for flights on a busy short-haul route. This approach allows us to compute the price elasticity of demand for this route at a daily and flight level resolution.

The data is sourced from the booking and flight databases of a large Western airline, and are a complete and accurate record of bookings. Therefore our data are free from the complex biases that can occur in booking datasets constructed using web crawlers or surveys. The airline wishes to remain anonymous, so that throughout this paper we refer to it as 'AirABC', and do not identify the origin and destination cities of the route. Only AirABC services this route, with alternatives restricted to other modes of transport or indirect flights, so that it is reasonable to consider these bookings in isolation of those for other airlines. Thus, our data are similar to those obtained from a controlled experiment. Tickets for different cabin classes (i.e. economy or business) and route directions are effectively separate products, and in our empirical work we consider bookings in one direction (so called half-return journey) for the main economy class cabin; although the model can be employed directly for other cabin classes or return journeys.

We model the booking process for each flight as an nonhomogeneous Poisson process with respect to the (decreasing) number of days to departure. The booking intensity has both a baseline component and a ticket price adjustment. The baseline component is modelled as additive in covariates, including smooth unknown functions of the flight departure time and day to departure. The price adjustments follow a finite mixture modeled using a multinomial logistic regression (MNL) with probabilities that are additive in covariates, including smooth unknown functions of the flight departure time and day to departure. Such a model is similar to the 'mixture-of-experts' models that are popular in the machine learning literature (Jordan and Jacobs, 1994), where each mixture component is called an 'expert'.

The unknown smooth functions in the baseline intensity and mixture probabilities are modeled semi-parametrically with penalized splines (Wood, 2017, chap. 5). This is important because prior research (Wen and Chen, 2017) and our empirical analysis suggests the effects of the key covariates 'flight departure time' and 'time to departure' can be highly non-linear. A quadratic penalty is used to ensure smoothness of each penalized spline, with the smoothing parameter selected by minimizing the BIC as in Ruppert et al. (2003) and Kauermann et al. (2009). The inclusion of covariates in this way means that each expert is a semi-parametric Poisson regression, and the MNL is also semi-parametric.

From a marketing perspective, the model provides a latent segmentation that accounts for customer heterogeneity (Wedel and Kamakura, 2012) at the daily booking count and flight level. Teichert et al. (2008) highlight the importance of identifying different segments to account for customer heterogeneity in airline passenger demand. They found more than two latent segments, which is consistent with our empirical work where we find up to seven segments. From a revenue management perspective, because the probability of latent class membership varies at the booking day and flight level, so does the ticket price elasticity. This is a key input into variable pricing frameworks. From a regulatory perspective, segmentation at the daily booking and flight level, as opposed to the customer level, avoids the need to collect individual level data. This is an advantage because the collection of such information can either be a concern to breach data privacy provisions, such as the EU General Data Protection Legislation, or is not available to practitioners. In particular data containing socio-economic and trip characteristics of air travelers as revealed by a preference survey (Wen and Chen, 2017; Teichert et al., 2008) is generally unavailable to the airline, nor can it be used by today's revenue management systems

(Hetrakul and Cirillo, 2014).

A central problem in the estimation of price elasticity using realized demand is that price is likely to be endogenous (Petrin and Train, 2010; Li et al., 2014). We address this using a control function approach similar to that suggested by Marra and Radice (2011) for generalized linear models. We employ the 'bid-price' (Talluri and van Ryzin, 2004, pp. 31) as an instrumental variable, which is an airline industry displacement measure that varies at both the flight and daily levels. We find strong evidence of all aspects of our proposed model— non-linear covariate effects, customer heterogeneity and price endogeneity—in our empirical analysis of passenger demand. A detailed overview of prior studies of retail demand for passenger flights in the revenue management literature that have features closest to ours is given in Section 1 of the Web Appendix.

Deep models from machine learning are also increasingly used to forecast complex time series with nonlinear serial dependencies (Diaconescu, 2008), including in transportation; see Ke et al. (2017), Lin et al. (2018) and Xu et al. (2018) for recent examples. Our proposed nonhomogenous Poisson model has the advantage of being interpretable and provides insights into customers' behavior that can be used in different scenarios. We mention this point explicitly since the airline industry market is experiencing dramatic changes through the COVID-19 pandemic (see e.g. Peterson and Thankom (2020) or IATA (c)). Even though the analysis in this paper uses data from prior to the pandemic, many of the empirical insights obtained in the nature and form of the key drivers of demand and price elasticity, as well latent segmentation, are likely to remain valid when air travel recovers post COVID-19. It also has the potential to provide forecasts under different scenarios. For example, baseline intensity can be adjusted to account for new realities in future passenger demand, while retaining the remaining aspects of the calibrated model, to produce flight-level daily demand forecasts.

To account for any unexplained intraday dependence between bookings for different flights we fit a multivariate model using a Gaussian copula and marginals given by the Poisson model. Dependence may exist between bookings for flights that depart at different times on the same day, because some customers might consider them as substitutes (i.e. when the time of flight is not a significant factor for a passenger). To date, only very few articles analyze the substitution patterns between flights in detail. One study to do so is Escobari (2017) who analyzes passenger choice behavior using a random coefficient logit model. However, this author found little evidence of significant cross-price elasticity at the departure time level, indicating limited substitution patterns between flights. In line with these results, estimates of the Gaussian copula model using our data suggest only low levels of dependence between bookings on different flights departing on the same day. Full details on the copula model and its application are given in Section 4 of the Web Appendix.

Last, we summarize here our main empirical findings. Correcting for price endogeneity in a mixture model framework has a substantial effect on the estimates of price elasticity, which is underestimated if the price is incorrectly treating as exogenous. Even though the consideration of price endogeneity is not novel to the literature, it is novel in a mixture model framework for latent segmentation. We identify a rich segmentation, with between five and seven latent classes for flights that depart on weekdays, but only two for weekend flights; although there is always at least one price insensitive and one highly price sensitive segment. The (i) day of the week on which bookings are made, (ii) number of days before departure, and (iii) time of the day at which the flight departs, are all strong nonlinear predictors of both the mixture component probabilities and baseline booking intensity. These three covariates all vary by flight and booking day, so that both the demand and price elasticity estimates from the model also vary by flight and booking day. Price sensitive customers tend to dominate up to 75 days before departure, and are replaced by price insensitive customers closer to the departure date. Interestingly, price elasticities are higher for customers who book on the weekend, compared to those who book their flights on a weekday. Thus, the date of booking (both the day type and the number of days before departure) reveals a great deal about the price elasticity of customers. Similarly, the time of departure of the flight itself is highly revealing, with morning and evening peak time flights having a higher proportion of price insensitive customers; presumably, because these flights are dominated by customers flying for business purposes. As all of the covariates used in our model are observable, our approach does not depend upon individual customer level data which is difficult to retain under data privacy provisions, such as the EU General Data Protection Regulation (GDPR). Hence, our segmentation model allows for ready forecasting of elasticity and demand for use in airlines' revenue management systems and therefore aid AirABC in effective variable pricing by flight and day of booking — an approach that it has adopted in practice.

The rest of the paper is organized as follows. Section 3.3 introduces the new dataset we employ, while Section 3.4 outlines the flexible Poisson model. The latter includes the mixture model, penalized spline smoothing, penalized maximum likelihood estimation and the approach to endogeniety correction. Section 3.5 contains the empirical analysis and Section 3.6 concludes our work. Extensive additional material is provided in the Web Appendix. This includes an in-depth literature review, additional empirical results, implementation details, and specification of the multivariate Gaussian copula model to account for additional dependence between bookings for different flights.

## 3.3 Data

### 3.3.1 Setting

The data are extracted from the booking system of AirABC, which provides a complete record of bookings. We analyze flight and matching retail booking data for a busy short-haul route over the two year period between April 1 2012 and 31 March 2014. The route is direct between two Western cities, which we do not name to ensure anonymity of AirABC, and for simplicity we only consider flights in one direction. Analysis of demand for this route is of particular interest because during this period only AirABC offered direct flights between these destinations, so that alternatives were limited either to indirect flights and other transportation modes. Both economy and business cabin classes were available, although our empirical analysis focuses on the economy cabin, which is the much larger of the two.

### 3.3.2 Flight Data

The route has up to 17 flights per day, and from these we exclude flights departing on public or school holidays, or correspond to major fairs, exhibitions and conferences at either the origin or destination cities. For these special day types, it is advisable to build separate models for passenger demand, which differs greatly from that on other departure days. If a flight is cancelled, then we retain all bookings over the days prior to cancellation, and do not consider any booking days afterwards. If a flight is rescheduled, we retain the original data on bookings prior to the date of reschedule and consider the initial flight cancelled afterwards. We then create a second flight with the departure details of the rescheduled flight, but with bookings possible only on days after the date of reschedule. With these exclusions and rules, our data includes a total of 9,602 flights scheduled to depart on a total of 730 days.

Flights are scheduled to depart every day of the week. There are also 61 distinct scheduled departure times recorded in our data, with the earliest departure at 06:00 and the latest at 21:55. The variable DDAY records the day of the week (Monday through Sunday) on which the flight departs, and the variable DTIME records the time of the day of the departure; both have a substantial impact on passenger demand.

### 3.3.3 Retail Booking Data

We only consider retail demand, based on bookings made within the published fare structure. Bookings made outside this fare structure, which includes those based on frequent flyer miles, corporate and private tariffs, or by airline staff, are omitted. Moreover, we only consider bookings that were also ticketed. This includes online transactions, where booking and ticketing are completed together. However, it excludes some bookings made by phone or via travel agents, where a booking can be made but is not ultimately ticketed due to non-payment. In addition, as discussed above, if a flight is rescheduled or cancelled by the airline, we retain the bookings in our data. We also retain a booking if the passenger cancels after ticketing, as this usually involves some monetary cost to the passenger.

Both return and single tickets are sold for this route. Purchasing the return ticket is always cheaper than two single tickets for the same two flights. Therefore, the motivation for purchasing each ticket type is likely to be different, so that we separate them. In our empirical work we only consider bookings made as part of a return ticket, both when the flight is the inbound or outbound section of a return ticket. We note that return tickets are more common than single ticket bookings for this route, at 93.1% of total bookings.

We construct booking specific variables as follows. We record the day of the week on which each booking was made (BDAY), along with the number of days prior to departure of the flight ($t$) and also the price paid (PRICE). Over 96.4% of total economy cabin class bookings were made within 120 days before the departure, and we only consider these bookings in our analysis. Bookings made on the day of departure have a value of $t = 0$, so that $0 \leq t \leq 120$. If all flights were open for booking during the 121 day period, there would be a total of $121 \times 9,602 = 1,161,842$ possible booking days. However, with flight cancellations and rescheduling as discussed above, the number of booking days in our data is slightly less at 1,109,559.

For historical reasons, airlines typically associate each ticket sold with a unique 'booking class', which should not be confused with the cabin class (i.e. economy or business). In our data there are 14 such booking classes which are ordered in terms of increasing price. During the two year period AirABC changed the fares associated with each booking class only once, which corresponded to an overall price increase. However, on any given day prior to departure, to change the price for a flight the airline simply opens or closes booking classes. This creates substantial variation in fares for each flight during the booking period. The majority of ticket purchases (94%) are at the lowest cost open booking class. The

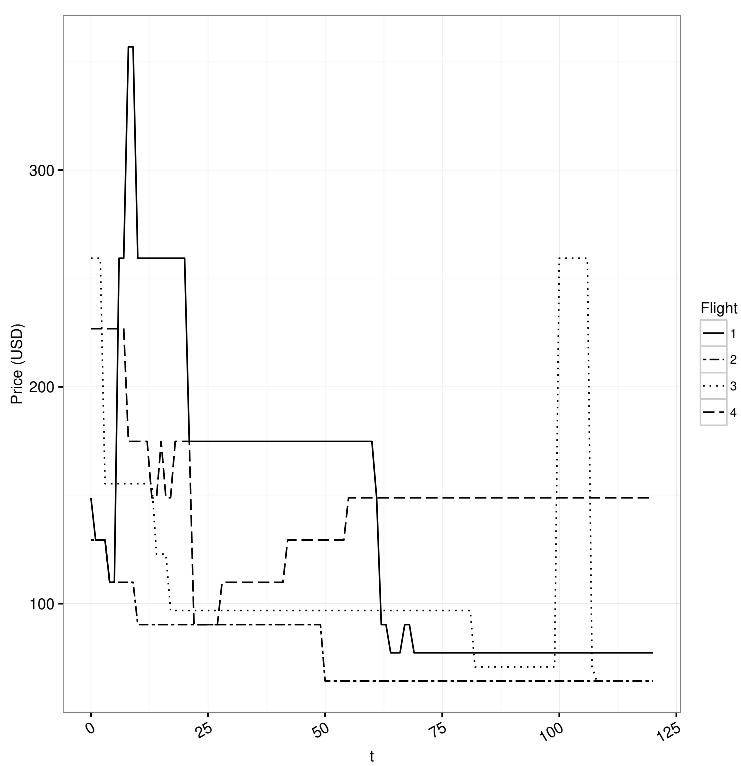



Figure 3.1: Prices of standard bookings (PRICE) for four flights against the time to departure ($t$), during the 120 day booking window. All three flights were open for booking throughout this window, and were scheduled to depart at 07:00.

remaining purchases are made at higher cost open booking classes, and are termed an 'upsell' by AirABC. In our data upsell bookings do not attract any meaningful additional customer benefits, and are likely due to complexities in the booking system. For simplicity, we exclude the small number of upsell bookings from our data, but note that our model can be readily applied to these bookings separately. Overall, there are 442,991 economy bookings recorded in our data for the 9,602 flights. To illustrate the level of variation in ticket prices for a flight, Figure 3.1 plots the prices (PRICE) of bookings for four typical flights over the 121 day booking period. Prices are quoted in U.S. dollars, although to help ensure anonymity of AirABC, we note that the tickets may, or may not, have been sold in this currency. The four flights were neither cancelled nor rescheduled during the

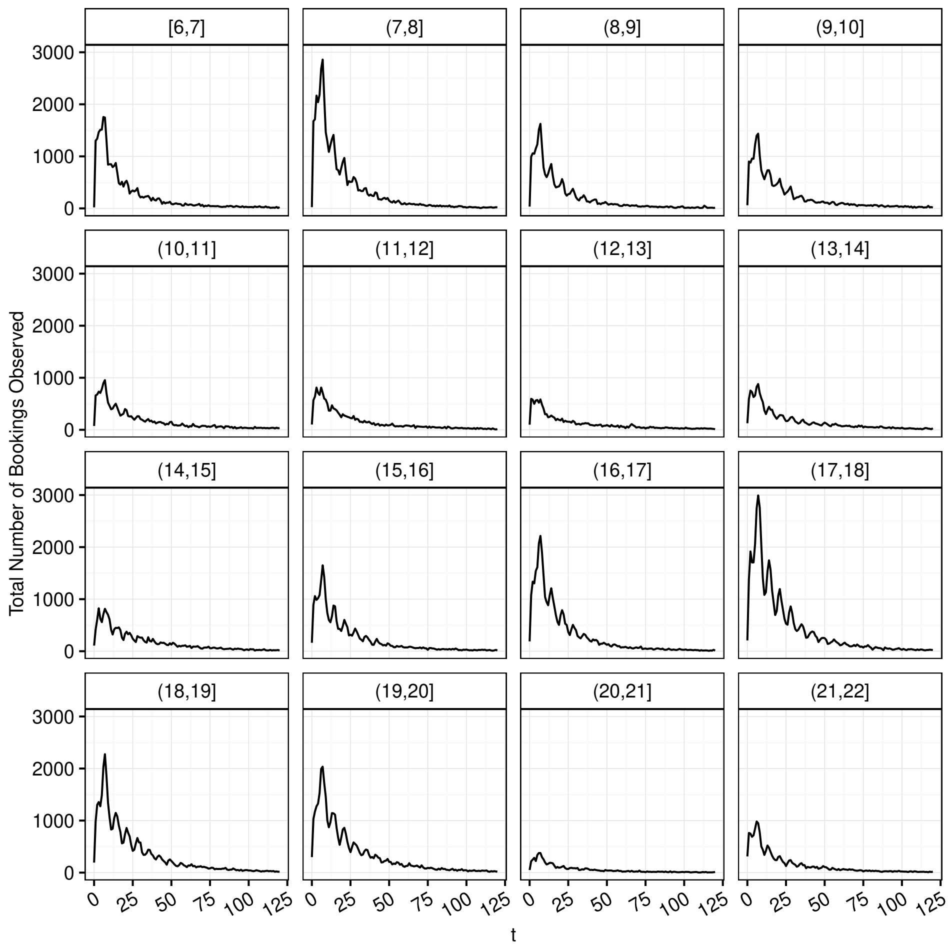



Figure 3.2: Total number of bookings in our data observed at each day prior to departure date. The bookings are further broken down into hourly intervals of flight departure times, with one panel for each hourly interval. For example, the top lefthand panel plots total bookings made up to 18 weeks prior to departure, only for flights departing between 06:00 and 07:00 (inclusive).

121 day booking period, and all depart at 07:00, which is during the daily peak period. The four price pathways reveal substantial price variation over the booking period, and

also across the three flights. This price variation is created by the process of opening and closing booking classes, as discussed above.

Figure 3.2 gives the total number of bookings in our data that were made in each seven day interval (i.e. week) prior to flight departure. The bookings are further broken down according to flight departure time, with each panel corresponding to flights leaving during different hourly intervals. Bookings are most heavily concentrated for flights departing between 06:00 – 08:00 and between 17:00 – 20:00. These are morning and evening travel peaks, and are typical of return ticket bookings for a short-haul flight. Regardless of the time of departure of a flight, booking intensity is strongest in the weeks immediately prior to departure; a feature that is again consistent with the short-haul nature of the flight. Last, we note that the day of the week on which the booking was made (BDAY) follows

| Variable | Mon | Tue | Wed | Thr | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| BDAY | 21.9 | 19.8 | 18.7 | 17.7 | 16.8 | 2.4 | 3.6 |
| DDAY | 14.5 | 16.8 | 19.9 | 21.1 | 17.3 | 4.9 | 5.5 |

Table 3.1: Relative frequency (in percent) of bookings made on different days of the week (BDAY), and also for the day of the week the flights depart (DDAY).

a different distribution than the day of the week on which flights depart (DDAY). To illustrate this, Table 3.1 provides the relative frequencies of both variables, from which we make three observations. First, bookings are almost exclusively made on weekdays for this route, with around 95% of bookings made between Monday and Friday. Second, while Monday and Tuesday are the most popular days on which to make a booking, Wednesday and Thursday are the most common departure days. Third, only 10% of bookings are for flights that depart on the weekend.

## 3.4 Model development

### 3.4.1 Semiparametric Mixed Poisson Regression for Bookings

Let $N_i(t)$ denote the total number of bookings for flight $i$ at $t$ days to departure, which is increasing for $t$ decreasing, so that $Y_i(t) = N_i(t) - N_i(t+1)$ is the number of passengers who book flight $i$ during day $t$. Flights departing on each day of the week are considered separately as different products, and DDAY is not incorporated into the notation to aid readability. Because we only consider bookings made up to 120 days prior to departure, we assume $N_i(121) = 0$, so that $N_i(0)$ is the total number of bookings for flight $i$ made during the 121 day window. The booking process $N_i(t)$ is modelled as a (time-reversed) nonhomogeneous Poisson process with intensity $\lambda_i(t) > 0$, which is factorized as

$$\lambda_i(t) = \lambda_{\text{BL}}(t) \left( \sum_{k=1}^{K} \pi_k(t) \delta_k \right) . \tag{3.1}$$

Here, $\lambda_{\text{BL}}(t) > 0$ is a time-varying baseline intensity, while the terms $\delta_1, \ldots, \delta_K$ are positive adjustments. These adjustments follow a latent finite mixture model with probabilities $\pi_1(t), \ldots, \pi_K(t)$, such that $0 \leq \pi_k(t) \leq 1$ and $\sum_{k=1}^{K} \pi_k(t) = 1$.

Eqn. (3.1) specifies a nonhomogeneous mixed Poisson model for booking activity (Karlis and Xekalaki, 2005), where the intensity follows a discrete mixing distribution with atoms at the points $\{\lambda_{\text{BL}}(t)\delta_1, \ldots, \lambda_{\text{BL}}(t)\delta_K\}$. The adoption of a mixture model is motivated by previous research which finds latent customer segments based on differing trip purposes and demographics of travelers; for example, see Teichert et al. (2008) and Wen and Lai (2010). To identify these segments we assume the intensity adjustment $\delta_k$ does not vary directly with day to departure, but we allow the probabilities $\pi_1(t), \ldots, \pi_K(t)$ to do so instead. However, the baseline intensity, adjustment values and associated probabilities are all functions of further flight and booking level covariates, as now discussed.

Table 3.1 illustrates that the booking intensity varies greatly with booking day (BDAY), while Figure 3.2 shows that it also varies substantially with departure time (DTIME) and day to departure ($t$). The logarithm of the baseline booking intensity is therefore modelled as an additive function of these variables, with

$$\log(\lambda_{\text{BL}}(t)) = \sum_{j=1}^{7} 1(\text{BDAY} = j)\beta_j^{(\lambda)} + s_0^{(\lambda)}(t) + s_1^{(\lambda)}(\text{DTIME})\,. \tag{3.2}$$

The term $1(A)$ is an indicator function equal to one if $A$ is true, and zero otherwise, so that $\boldsymbol{\beta}^{(\lambda)} = (\beta_1^{(\lambda)}, \ldots, \beta_7^{(\lambda)})$ is a vector of booking day intensity effects. Here, the superscript $\lambda$ distinguishes these baseline booking intensity effects from those for the segment probabilities $\pi_k(t)$ introduced later. The impact of $t$ and DTIME are modeled as unknown smooth functions $s_0^{(\lambda)}$ and $s_1^{(\lambda)}$ as discussed further below. To identify the level in equation (3.2), we follow Hastie and Tibshirani (1990) and set the integrals of these functions to zero over their domain.

Previous research (Hetrakul and Cirillo, 2014; Li et al., 2014; Vulcano et al., 2010) indicates there is strong customer heterogeneity in the price elasticity for passenger flights. Our objective in adopting the mixture model is to capture segment specific price elasticities parsimoniously. These are log-linear within each segment, with

$$\log(\delta_k) = \alpha_k \text{PRICE} \tag{3.3}$$

The overall price elasticity is therefore $E_\lambda = \sum_{k=1}^{K} \pi_k(t)\alpha_k$ which varies with $t$ and other covariates through the probabilities $\pi_1(t), \ldots, \pi_K(t)$. For modelling these segment probabilities a multinomial logistic regression (MNL) model is adopted. If segment $K$ is taken as reference category, then the log-odds are

$$\log\left(\frac{\pi_k(t)}{\pi_K(t)}\right) = \beta_{j,1} + \sum_{j=2}^{7} 1(\text{BDAY} = j)\beta_{j,k}^{(\pi)} + s_{0,k}^{(\pi)}(t) + s_{1,k}^{(\pi)}(\text{DTIME})\,, \tag{3.4}$$

for segments $k = 1, \ldots, K-1$. This is a semiparametric specification, because the effect of $t$ and DTIME are given by unknown smooth functions $s_{0,k}^{(\pi)}$ and $s_{1,k}^{(\pi)}$. As with the baseline intensity, the functions are constrained to integrate to zero to identify the level in Eqn. (3.4). The coefficients $\boldsymbol{\beta}_k^{(\pi)} = (\beta_{1,k}^{(\pi)}, \ldots, \beta_{7,k}^{(\pi)})$ capture the booking day type level effect for segment $k$, relative to the reference category.

### 3.4.2 Penalized Likelihood Estimation and Inference

The unknown functions $s_0^{(\lambda)}, s_1^{(\lambda)}$ for the intensity, and $\{s_{0,k}^{(\pi)}, s_{1,k}^{(\pi)}; j = 1, \ldots, K-1\}$ for the MNL model, are modelled using penalized splines. This is a popular approach to smooth function estimation; see Wood (2017) and Ruppert et al. (2009) for overviews and Smith and Kauermann (2011) for their use in transportation science. The advantage of using splines instead of flexible functional forms based on Fourier terms as in Wen and Chen (2017) and Lurkin et al. (2017), is that they allow for data-driven levels of smoothing (i.e. regularization). A penalized spline approximates an unknown function by the inner product of a vector of basis terms $\boldsymbol{w}(.)$ with a coefficient vector $\boldsymbol{\gamma}$, so that each function is $s(.) = \boldsymbol{w}(.)'\boldsymbol{\gamma}$. Smoothness is achieved by adopting a regularization penalty on $\boldsymbol{\gamma}$. For univariate functions, Eilers and Marx (1996) proposed for a B-spline basis an appropriate quadratic penalty $\rho\boldsymbol{\gamma}'D\boldsymbol{\gamma}$, where $\rho D$ is the precision matrix of a first order random walk in the elements of $\boldsymbol{\gamma}$. In this case, $D$ is a constant band one matrix, and $\rho$ is a scalar smoothing parameter. We adopt this basis and penalty here for each unknown function in our model, as discussed further in Section 5 of the Web Appendix, . Using the same super- and subscripts for the penalized spline coefficients as the unknown functions, the parameters of the model are therefore

$$\boldsymbol{\theta} = \left\{ (\boldsymbol{\beta}_1^{(\pi)}, \boldsymbol{\gamma}_{0,1}^{(\pi)}, \boldsymbol{\gamma}_{1,1}^{(\pi)}), \ldots, (\boldsymbol{\beta}_{K-1}^{(\pi)}, \boldsymbol{\gamma}_{0,K-1}^{(\pi)}, \boldsymbol{\gamma}_{1,K-1}^{(\pi)}), \boldsymbol{\beta}^{(\lambda)}, \boldsymbol{\gamma}_0^{(\lambda)}, \boldsymbol{\gamma}_1^{(\lambda)}, \alpha_1, \ldots, \alpha_K \right\}.$$

If $y_{i,t} \in \{0, 1, 2, \ldots\}$ is the number of bookings for flight $i$ made on $t$ days to departure, and the corresponding observation of the three covariates is

$$\boldsymbol{x}_{i,t} = (\mathrm{DTIME}_i, \mathrm{BDAY}_t, \mathrm{PRICE}_{i,t}),$$

then the (unpenalized) log-likelihood arising from Eqn. (3.1) is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{t=t_i^{\mathrm{close}}}^{t_i^{\mathrm{open}}} y_{i,t} \log(\lambda(\boldsymbol{x}_{i,t}, t; \boldsymbol{\theta})) - \lambda(\boldsymbol{x}_{i,t}, t; \boldsymbol{\theta}). \tag{3.5}$$

Here, the booking and flight specific intensity in Eqn. (3.1) is written as a function of the covariates and model parameters as $\lambda(\boldsymbol{x}_{i,t}, t; \boldsymbol{\theta})$. The outer summation is over the number of flights $n$ in the sample, as reported in Table 3.2, while $t_i^{\mathrm{open}}$ and $t_i^{\mathrm{close}}$ are the days to departure at the opening and closing of booking for flight $i$. For example, if flight $i$ is not canceled or rescheduled during the 121 day booking window, then these values are

$t_i^{\text{open}} = 120$ and $t_i^{\text{close}} = 0$. Whereas, if flight $i$ was canceled 100 days prior to departure, then $t_i^{\text{close}} = 100$.

In Eqn. (1), the covariates are observed on the same resolution as the booking variable, which is the daily level for each flight, which is also the resolution of the revenue management system used by AirABC. Both DTIME and BDAY are observed at this resolution, but the price of a ticket for a given flight can vary between multiple bookings made on the same day so that PRICE is not. In practice, the PRICE variable changes during the day whenever AirABC opens or closes booking classes for a flight mid-way through the day—for example, when a booking class quota is exhausted—and there are 13,988 booking day/flight combinations in our data where this occurs.

To manage intra-day price variation without losing information by averaging the PRICE variable (which could be employed with the likelihood function at Eqn. (1)) and ensure that the predictions are created on a daily level for each flight, we incorporate PRICE variation in the likelihood using differing aggregation levels. For example, if three bookings are observed on a single day, we assume an aggregation level of $\frac{1}{3}$ day. This leads to an offset mirroring the aggregation level as described, for instance, in Tutz (2012, Sec. 7.2). To specify this here, let

$$\boldsymbol{x}_{i,t,l} = \left(\text{DTIME}_i, \text{BDAY}_t, \text{PRICE}_{i,t,l}\right),$$

be the covariate vector for the $l$th booking made $t$ days to departure for flight $i$, where $l = 1, ..., \max(1, y_{i,t})$. On days without any bookings for flight $i$ (i.e. when $y_{i,t} = 0$), let $\boldsymbol{x}_{i,t,1}$ be the vector of covariate values, and set $y_{i,t,1} = 0$. Similarly, let $y_{i,t,l} = 1$ for $l = 1, ..., \max(1, y_{i,t})$ for days with observed bookings (that is, when $y_{i,t} \geq 1$). Then, the (unpenalized) log-likelihood with an aggregation offset is:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{t=t_i^{\text{close}}}^{t_i^{\text{open}}} \sum_{l=1}^{\max(1,y_{i,t})} y_{i,t,l} \log\Big(\lambda(\boldsymbol{x}_{i,t,l}, t; \boldsymbol{\theta})\Big) - \frac{\lambda(\boldsymbol{x}_{i,t,l}, t; \boldsymbol{\theta})}{\max(1, y_{i,t})}. \tag{3.6}$$

The multiple summation in Eqn. (3.6) is over all observed bookings, plus the booking days where no bookings were made for the $i$th flight (i.e. all instances where $y_{i,t} = 0$). These summations are over all flights $i$ that depart on each given day type. The bottom row of Table 3.2 reports the number of terms in the summation, and there are between $139,006$ and $228,026$ of these. Note that if there were no intraday variation in price, then Eqn. (3.6) and (1) would be the same.

Eqn. (3.6) is augmented with an additive penalty to account for smoothness in the functions. The first and second order derivatives are computed analytically (see Section 5.1 of the Web Appendix) enabling fast direct maximization of the penalized log-likelihood; even for the high sample sizes employed here. The optimal values of these smoothing parameters are selected by minimizing the Bayesian Information Criterion (BIC). The number of latent segments is also selected using BIC, where we fit models with increasing number of segments $K$ as long as this decreases the BIC as in Allenby and Rossi (1998). Bootstrap confidence intervals for the parameters and functions of a fitted model are computed using the 'leave out one individual' approach of Rice and Silverman (1991). The identification of the segment labels in the mixture model is achieved by ordering the segment specific price coefficients $\alpha_k$ in a monotone sequence. We refer to Section 6 of the Web Appendix for details.

We comment briefly on the suitability of selecting the number of latent segments using BIC. Whittaker and Miller (2021) explores the accuracy of enumerating the number of classes using different metrics in latent class analysis. They found strong evidence to suggest that sample size adjusted BIC (NBIC) was more accurate than a variety of alternatives, including cross-validation and BIC. However, the results also show similar enumeration accuracy for BIC and NBIC with an increasing sample size. Because our analysis is based on a large sample of size $n = 1,109,559$, BIC is an accurate metric for latent class enumeration.

### 3.4.3 Semiparametric Regression for Price

Treating price as an exogenous variable in a consumer demand model can lead to biased estimates of price elasticity; see discussions in Davidson and MacKinnon (1999, 1993), Wooldridge (2002), Petrin and Train (2010) and references therein. For example, Mumbower et al. (2014) show the importance of controlling for price endogeneity in a linear model for flight bookings using a two-stage least squares linear regression estimator, whereas Lurkin et al. (2017) do so for a choice model. For generalized nonlinear models, Marra and Radice (2011) suggest an extension of such two-stage estimators, similar to the control function approach of Petrin and Train (2010). We follow these authors and first build a nonlinear model for price based on an instrumental variable, and then include the price residual as a covariate in our model of passenger demand.

| | | | | DDAY | | | | |
|---|---|---|---|---|---|---|---|---|
| *Number of ...* | Mon | Tue | Wed | Thr | Fri | Sat | Sun | *Total* |
| Flights $n$ | 1,385 | 1,295 | 1,435 | 1,528 | 1,593 | 1,124 | 1,242 | 9,602 |
| Dep. Days $\|D\|$ | 105 | 104 | 104 | 104 | 104 | 104 | 105 | 730 |
| Booking Days | 157,932 | 147,518 | 164,323 | 173,406 | 179,898 | 131,000 | 137,482 | 1,091,559 |
| Bookings | 64,371 | 74,383 | 88,070 | 93,311 | 76,816 | 21,712 | 24,328 | 442,991 |
| Non-Bookings | 128,342 | 117,611 | 129,408 | 134,715 | 141,746 | 117,294 | 121,605 | 890,721 |
| Observations | 192,713 | 191,994 | 217,478 | 228,026 | 218,562 | 139,006 | 145,933 | 1,333,712 |

Table 3.2: Summary of data size, broken down by departure day type DDAY. The first three rows report the number of flights, departure days and possible booking days for these flights. The next two rows report the number of observed bookings, and booking days for each flight where no bookings were made. The final row gives the total of the the number of bookings and non-bookings observed, which is the number of terms in the likelihood at Eqn. (3.6).

To do so, we model the logarithm of prices at the daily and flight level as

$$\begin{aligned}\log(\text{PRICE}) &= \theta_0 + \theta_1 \text{IV} + \sum_{j=2}^{7} 1(\text{BDAY} = j)\theta_j + f_0(t) + f_1(\text{DTIME}) + U \\ &= \eta + U\,, \end{aligned} \tag{3.7}$$

where $U \sim N(0, \sigma^2)$. The effects of $t$ and DTIME are captured by unknown smooth functions $f_0$ and $f_1$ modelled by penalized splines, while IV is an instrumental variable.

Mumbower et al. (2014) discusses possible choices for IV and suitable candidates. Li et al. (2014) notes that many of these choices are invalid because both the IV and booking data need to be observed at the same level of aggregation to control effectively for price endogeneity. Supply shifters—for example, airport fees, transportation taxes and fuel costs—are constant over daily bookings. Hausman-style instruments at the firm level do not match to a model on the market level. Stern-type instruments that measure competition and market share do not vary on the booking level. Last, IVs that have an impact on marginal costs remain a feasible option, which is why we use (the logarithm of) a variable that is popular in the revenue management literature called the 'bid-price' (Talluri and van Ryzin, 2004, pp. 31). The bid-price is a measure of the (marginal) cost of offering a seat, taking into account that it cannot be sold again. Crucially, it varies between bookings because the airline updates its assessment frequently. The bid-price is available for all flights in the database and at all time points, as well as for predictive purposes, i.e. for flights that are yet to depart.

To ensure the validity of our choice the IV needs to fulfill the properties of relevance and exogeneity (Guevara, 2018). Whereas (strong) relevance can easily be demonstrated by the strong nonlinear dependence between the IV and the endogenous variable price, exogeneity needs to be addressed by a statistical (over-identification) test. Unfortunately, this test requires the availability of at least two instruments, so that exogeneity cannot be established definitively. From a qualitative perspective, the bid-price is a measurement of displacement cost, ensuring that revenue gain for the available airlines' network capacity is maximized. As pointed out by Li et al. (2014), the exogeneity (and hence the validity of the bid-price IV) means that a demand shock for flight $i$ at time to departure $t$ (i.e. $\varepsilon_i(t) = Y_i(t) - \lambda_i(t)$) is uncorrelated with the IV. Figure 1 describes two possible revenue management setups, where an airline only controls for displacement cost on route-level (left-hand side) or incorporates all possible demand-streams into the displacement cost calculation (right-hand side). As AirABC is a network carrier, it considers every demand stream when calculating the bid-price value. Therefore the bid-price defines the distribution of all network demand on the route. In our study, the share of transfer passengers, i.e., passengers not traveling solely between BBB and AAA, is approximately 50%. Thus, the bid-price value is largely determined by factors that are exogenous to the route under study. Hence, we conclude that the demand shock $\varepsilon_i(t)$ and the bid-price are uncorrelated.

We fit the model at Eqn. (6) using maximum likelihood, and then use this to estimate the error

$$\xi = \text{PRICE} - \mathbb{E}\left(\text{PRICE} \mid \text{IV,BDAY,t,DTIME}\right) = \text{PRICE} - \exp\left(\eta + \sigma^2/2\right)$$

for each flight and booking day combination. The resulting residuals values are observations on the covariate $\hat{\xi}$, which is included in the log-linear segment price adjustments, so that we replace Eqn. (3.3) by

$$\log(\delta_k) = \alpha_{1,k}\text{PRICE} + \alpha_{2,k}\hat{\xi} \tag{3.8}$$

We will subsequently refer to Model I if we ignore endogeneity and use Eqn. (3.3). Taking endogeneity into account and using Eqn. (3.8) is referred to as Model II. A more detailed motivation for this two-stage procedure using the bid-price as an instrumental variable is given in Section 7 of the Web Appendix.

## 3.5 Empirical analysis

We now discuss the estimates from our model. Because we fit it to bookings for flights departing on different day types—that is, different values of DDAY—separately, we give in detail the results arising from flights departing on Thursday. This is the departure day with the highest demand.

We fit the demand models with $K = 2, \dots, 7$ segments, both including and excluding the price model residuals $\hat{\xi}$ (the calculation of $\hat{\xi}$ is discussed in Section 2 of the Web Appendix). The inclusion of the residuals improves the fit of the demand models substantially—as measured using either AIC or BIC—in every case. A detailed discussion of the $K = 2$ segment

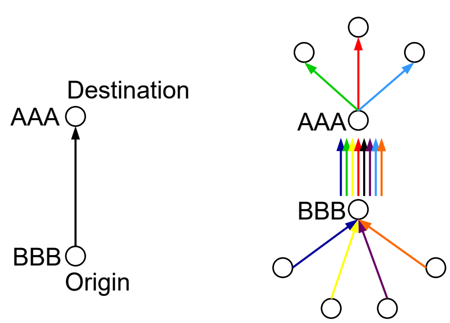



Figure 3.3: Description of two Airline-network scenarios. On the left-hand side, the airline controls for capacity constraints only taking passenger demand from the origin (BBB) to the destination (AAA) into account. Low-cost-carriers typically use this setup. On the right-hand side, the airline controls for the capacity constraint on the BBB to AAA route by taking all possible passenger demand streams coming from other origins than BBB (arrows going into BBB) to different destinations than AAA (arrows going out of AAA) into account. Network-carriers typically use this setup.

| Segment | BL | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ |
|---|---|---|---|---|---|---|
| Component | $\lambda_{\mathrm{BL}}(t)$ | | $\pi_k(t)$ | | | |
| | | | *Segment Adjustment Coefficients* | | | |
| PRICE | | 0.0019 | -0.0035 | -0.0287 | -0.0295 | -0.0301 |
| | | (0.0006) | (0.0022) | (0.0082) | (0.0025) | (0.0025) |
| $\widehat{\xi}$ | | -0.0030 | -0.0036 | -0.0126 | -0.0133 | -0.0135 |
| | | (0.0003) | (0.0004) | (0.0039) | (0.0017) | (0.0019) |
| | *Baseline Coefficients* | | *Log-odds Coefficients* | | | |
| Intercept | 1.4254 | -3.3756 | -0.5172 | 3.0461 | 0.6938 | – |
| | (0.3028) | (0.5777) | (1.0661) | (1.5426) | (1.3458) | – |
| BDAY = Mon | 0.3470 | 2.3653 | 0.1170 | -1.6610 | -2.0355 | – |
| | (0.3028) | (0.4349) | (0.5276) | (0.7180) | (0.4943) | – |
| BDAY = Tue | 0.2948 | 2.7756 | 0.4308 | -1.3427 | -1.2495 | – |
| | (0.1532) | (0.4976) | (0.4139) | (0.5904) | (0.5588) | – |
| BDAY = Wed | 0.3027 | 2.2079 | -0.0261 | -1.9197 | -2.1133 | – |
| | (0.1718) | (0.6987) | (0.2931) | (0.4627) | (0.8671) | – |
| BDAY = Thr | 0.3088 | 3.4974 | 1.0745 | -0.5302 | 0.5838 | – |
| | (0.1623) | (0.6321) | (0.5552) | (0.8554) | (0.9305) | – |
| BDAY = Fri | 0.2337 | 2.8789 | 0.2554 | -1.2382 | -0.4404 | – |
| | (0.1516) | (0.5650) | (0.3854) | (0.6682) | (1.8257) | – |
| BDAY = Sat | -0.4454 | -0.3175 | -0.0387 | -0.1650 | -0.2100 | – |
| | (0.0868) | (0.3989) | (0.4055) | (0.4170) | (0.5661) | – |

Table 3.3: Parameter estimates for Model II (i.e. the with the inclusion of the residuals $\hat{\xi}_i$) with $K = 5$ latent class segments, fitted to bookings on flights departing on Thursday. Bootstrap standard errors are given in parentheses.

model estimates, and the impact of controlling for endogeneity, is given in Section 3 of the Web Appendix. For all seven departure days (DDAY), Table 3.5 reports the BIC values for all fitted demand models that include the residuals $\hat{\xi}$ and different numbers of segments. For flights departing on thursday (DDAY = Thr), $K = 5$ segments are optimal with the minimum BIC value. Table 3.3 gives the estimates of the linear coefficients. Inclusion of the price residual has a substantial effect on the parameter estimates so that we subsequently only discuss the results with price endogeneity taken into account. The segment adjustment coefficients shows that the PRICE coefficient for segment 2 is insignificant and close to insignificant for segment 1. However, segments 3, 4, and 5 exhibit significant price sensitivities between $\hat{\alpha}_3 = -0.0287$ and $\hat{\alpha}_5 = -0.0301$.

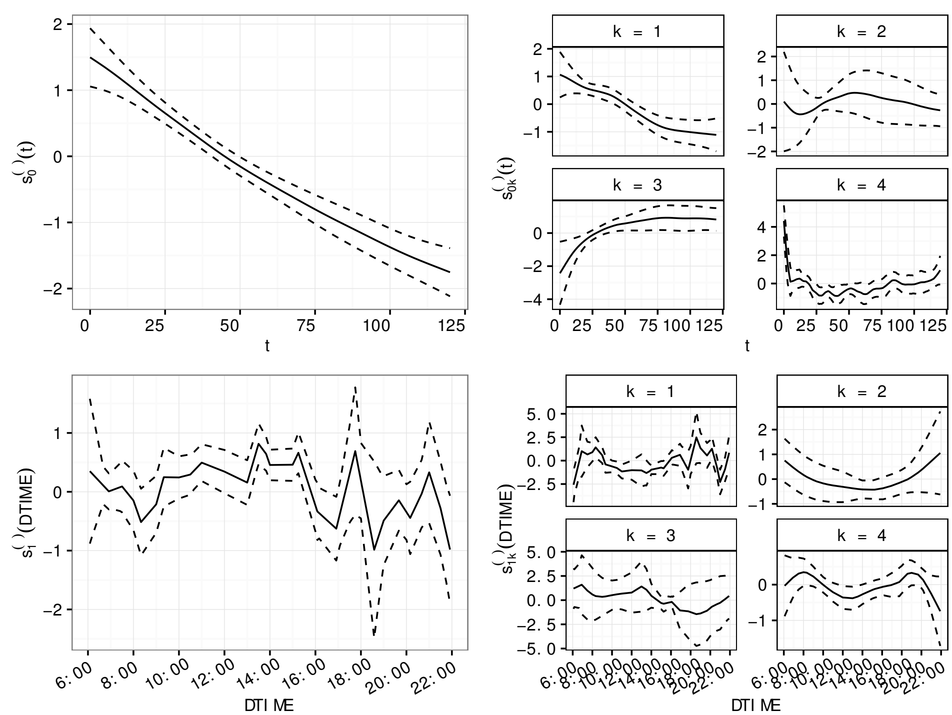



Figure 3.4: For $K = 5$ segments, the left-hand panels provide the function estimates for $s_0^{(\lambda)}(t)$ and $s_1^{(\lambda)}(\text{DTIME})$ in Eqn. (3.2) for bookings on flights that depart on Thursday. The right-hand side shows the estimates of $s_{0,k}^{(\pi)}(t)$ and $s_{1,k}^{(\pi)}(\text{DTIME}), k = 1, \ldots, 4$ in Eqn. (3.4). The first-stage residuals $\hat{\xi}$ are included (ie. Model II). The estimates are given by the solid line, while the dashed lines are 99% local confidence bands.

Figure 3.4 shows the fitted smooth terms of model component at Eqn. (3.2) (left panel) and Eqn. (3.4) (right panel). We see a general increase in demand closer to the day of departure (i.e. for lower values of $t$). Moreover, the size of segments 1 and 4 increase, and segment 3 decreases, closer to the day of departure. Segment 2 shows no significant time effect. DTIME has only a weak impact on demand, although this is not the case for customer segmentation which we discuss next.

To measure the composition of customers as a function of time to departure, we compute the ratio

$$q_k(t) = \frac{\pi_k(t)\delta_k}{\sum_{k'=1}^{K} \pi_{k'}(t)\delta_{k'}}\,, \tag{3.9}$$

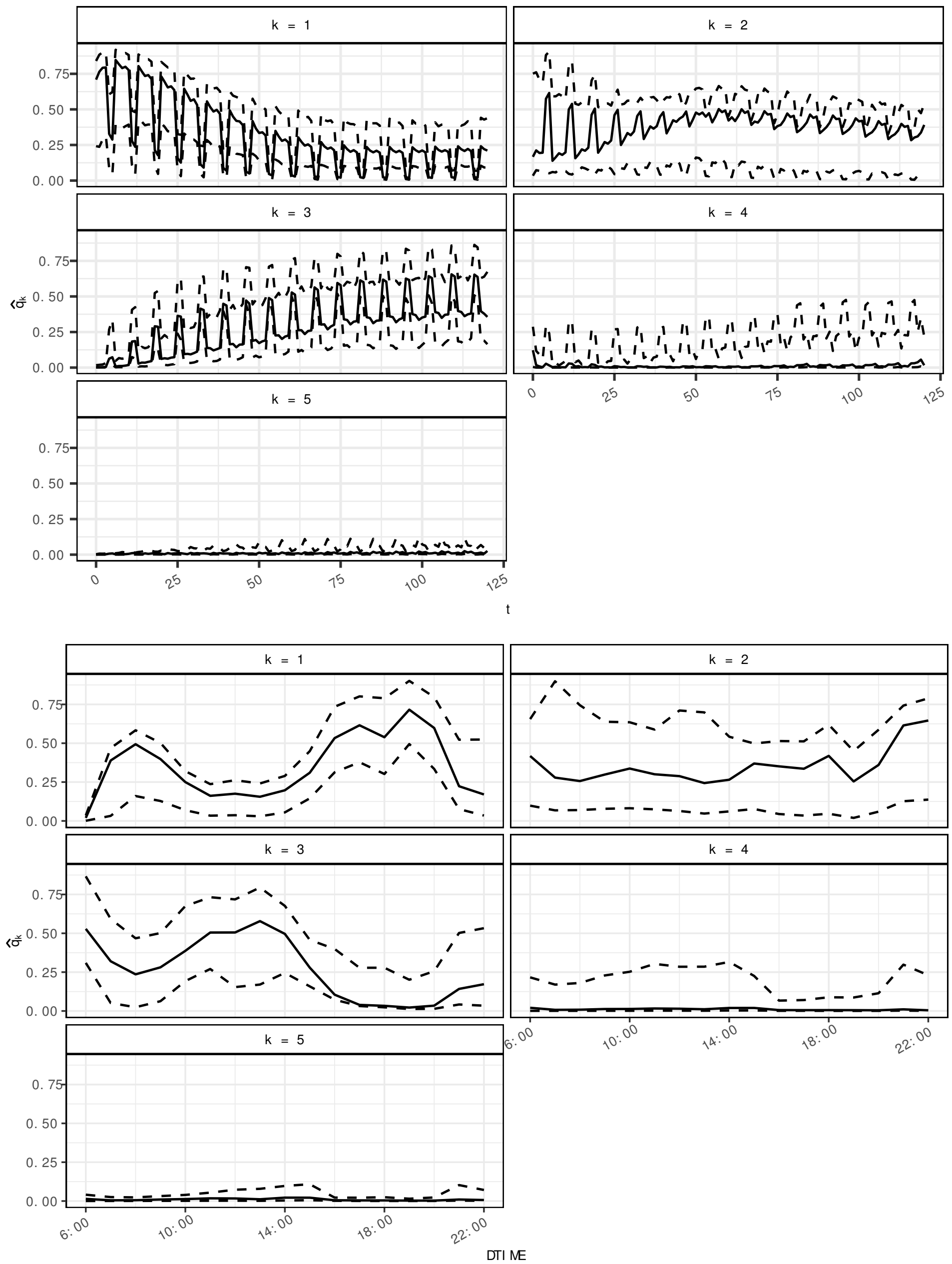



Figure 3.5: Plot of the average segment proportion computed from the model fitted to booking on flights departing on Thursday and $K = 5$ (solid line) with 99% local bootstrapped confidence bands (dashed lines). Top row: within each panel, $\bar{q}_k(t)$ is plotted against days to departure $t$. Bottom row: within each panel, $\bar{q}_k(\text{DTIME})$ is plotted against DTIME.

for $k = 1, 2, \ldots, K$. This ratio measures the proportion of customers in segment $k$. In our demand model the component $\pi_k(t)$ is a function of both flight and booking level covariates, so that we compute the mean $\bar{q}_k(t)$ by averaging $q_k(t)$ over all flights and bookings on a given day to departure $t$.

Figure 3.5 (top panel) plots $\bar{q}_k(t)$ for the five segments against days to departure. Only a very small proportion of bookings fall into the price sensitive segments 4 and 5. For segment 5 passengers arrive anytime, whereas segment 4 corresponds to a type of passenger that arrives shortly before departure. The vast bulk of bookings by price sensitive customers are in segment 3. This accounts for around 40-50% of all bookings made up to 75 days before departure, but gradually declines as the flight departure approaches, falling to almost none in the week prior to departure. Bookings made in this segment are also more likely to be made on the weekend (i.e. when BDAY is either Saturday or Sunday). The proportion of bookings that fall into the two price inelastic segments have quite different patterns. The probability of a booking in segment 1 is at most 20% until 75 days prior to departure, after which it increases rapidly until the day of departure, during which just over 80% of bookings arise from this segment. Bookings in segment 2 are common throughout the booking window, varying between around 20% to 60% of the total. Interestingly, bookings in this segment exhibit a strong booking day effect—with bookings much more likely on the weekend than weekdays—a stark difference with bookings in segment 1 which do not.

The probability $\pi_k$ of being in segment $k$ is also a function of DTIME through the MNL model at Eqn. (3.4). Thus, the diagnostic ratio can be also be computed as a function of DTIME, which we write as $\bar{q}_k(\text{DTIME})$. The bottom panel in Figure 3.5 plots this ratio against DTIME for each of the five segments. Of the two price insensitive latent classes, segment 1 accounts for around 50% of all bookings on fights departing during the morning peak, and a striking 70% of those during the evening peak. In contrast, segment 2 bookings exhibit a preference for the late evening. Bookings in the price sensitive segment 3 are largely for flights departing during off-peak periods, whereas segment 4 and 5 show no particular time preference.

Table 3.4 summarizes the main features of each latent segment, which we label as 'Rush Peak-time' (segment 1), 'Planned Evening Business' (segment 2), 'Planned Leisure' (segment 3), 'Bargain Catcher' (segment 4), and 'High Value Seeker' (segment 5). We also compute the overall elasticity estimate $E_\lambda$ that averages over the latent segments. Figure 3.6 plots $E_\lambda$ against the time to departure for select values of DTIME and BDAY. All panels show that the price elasticity decreases as the day of departure nears ($t = 0$). This effect is stronger for a weekday booking day, e.g. Monday, compared to a weekend booking day such as Sunday. In the weeks immediately prior to departure, tickets on morning and evening flights are much more price inelastic than tickets for midday flights. Overall, the results indicate that $K = 5$ passenger segments successfully identify customer heterogeneity in price elasticity broken down by time to departure ($t$) and departure time (DTIME), allowing for optimal variable pricing of tickets.

| | Segment | | | | |
|---|---|---|---|---|---|
| Booking Features & Preferences | 1 Rush Peak-time | 2 Planned Evening Business | 3 Planned Leisure | 4 Bargain Catcher | 5 High Value Seeker |
| Price Sensitive? | No | No | Yes | Yes | Yes |
| Relative Size | Large | Medium | Medium | Small | Tiny |
| Flight Time Pref. | Peak | Evening | Midday | No Pref. | No Pref. |
| Day of Booking | Weekday | Any day | Weekend | Thr | Baseline |
| Booking Day Relative to Flight Departure | Closer | Throughout | Earlier | Last Minute | Anytime |

Table 3.4: Summary of main booking features and flight preferences of bookings made in each of the four latent segments of the demand model (with $K = 5$ and price residual inclusion) fit to bookings made for flights departing on Thursday.

So far we have looked at Thursday departures only. We extend this now and fit the demand

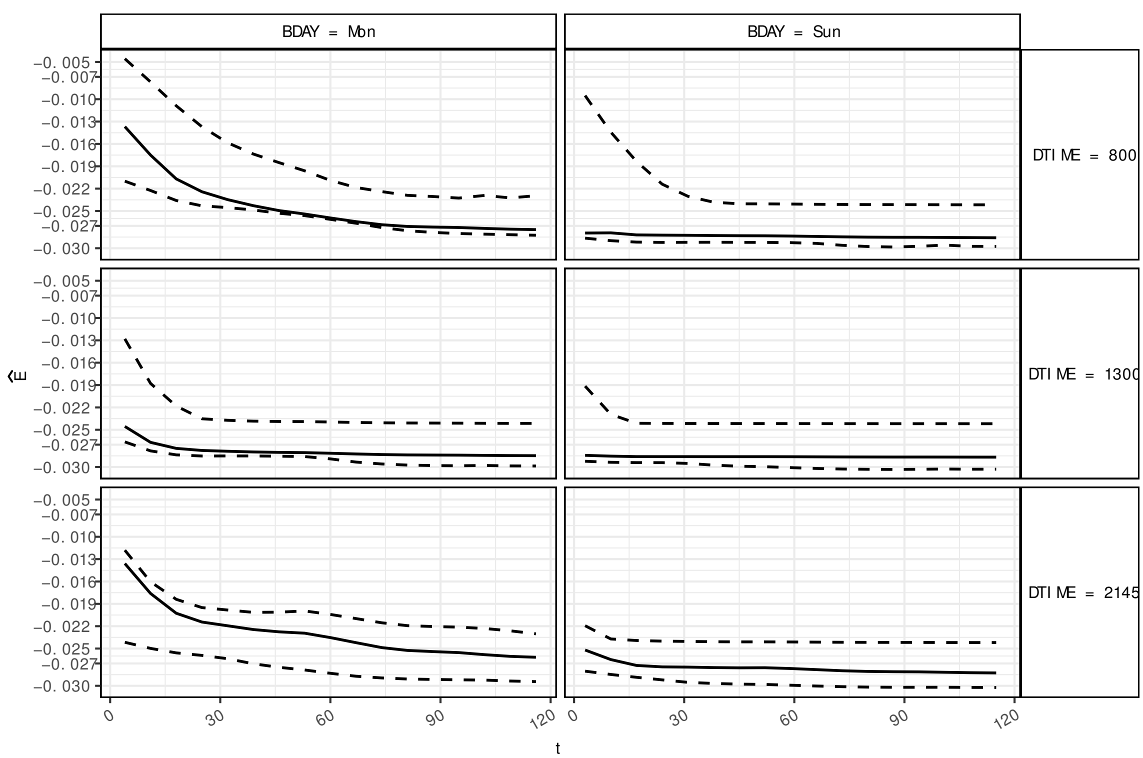



Figure 3.6: Estimated overall price elasticity $E_\lambda$ (solid line) for a mixture of $K = 5$ customer segments, estimated with endogeneity correction. Also plotted are 99% local bootstrapped confidence bands (dashed lines). Six combinations of booking day (BDAY) and departure time (DTIME) are considered, and days to departure ($t$) is on the horizontal axis.

model to bookings for flights on all departure days. The BIC values for $K = 1, \ldots, 7$ customer segments are shown in Table 3.5, while the corresponding estimated coefficients of PRICE for the optimal model based on the BIC are reported in Table 3.6. For weekday departures (except Monday) $K = 5$ is optimal throughout, and the segment specific price sensitivities are similar across departure day.

| | | No. Segments | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DDAY | | $K=1$ | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $K=6$ | $K=7$ |
| Mon | BIC | 191939.0 | 188487.0 | 188374.2 | 188259.2 | 188306.1 | 188273.0 | **188274.9** |
| | log-lik | -95503.1 | -93520.3 | -93285.5 | -93342.5 | -93310.0 | -93201.7 | **-93182.5** |
| | dist. | 0.0 | 3980.9 | 4198.3 | 4267.2 | 4243.5 | 4328.4 | **4337.1** |
| Tue | BIC | 190944.8 | 188571.7 | 188363.8 | 188181.1 | **188001.9** | 188086.4 | 188292.3 |
| | log-lik | -95005.5 | -93774.8 | -93354.7 | -93359.1 | **-93198.1** | -93194.1 | -93324.6 |
| | dist. | 0.0 | 2673.2 | 3063.7 | 3216.9 | **3453.6** | 3384.0 | 3140.2 |
| Wed | BIC | 223292.3 | 220656.5 | 220037.9 | 220285.9 | **219981.6** | 220096.5 | 220071.4 |
| | log-lik | -111173.9 | -109818.4 | -109284.7 | -109188.8 | **-109098.5** | -109170.9 | -109177.1 |
| | dist. | 0.0 | 2963.9 | 3763.0 | 3602.7 | **3907.5** | 3771.7 | 3789.7 |
| Thr | BIC | 251600.2 | 247740.6 | 246328.3 | 246035.5 | **245916.0** | 246052.4 | 246129.1 |
| | log-lik | -125325.9 | -123286.0 | -122505.7 | -122311.4 | **-122129.7** | -122110.6 | -122116.1 |
| | dist. | 0.0 | 4365.5 | 5978.9 | 6328.8 | **6521.2** | 6412.2 | 6343.2 |
| Fri | BIC | 259010.7 | 251940.9 | 250399.2 | 250427.2 | **250215.5** | 250202.4 | 250356.3 |
| | log-lik | -129033.3 | -125341.8 | -124400.1 | -124288.3 | **-124130.0** | -124182.0 | -124242.5 |
| | dist. | 0.0 | 7975.5 | 9778.7 | 9807.7 | **10069.6** | 10055.9 | 9891.9 |
| Sat | BIC | 117554.1 | **115858.9** | 116022.4 | 116088.2 | 116135.1 | 116156.7 | 116170.7 |
| | log-lik | -58551.9 | **-57554.7** | -57588.0 | -57582.2 | -57596.6 | -57584.5 | -57590.5 |
| | dist. | 0.0 | **1966.8** | 1809.8 | 1757.6 | 1710.6 | 1699.6 | 1684.7 |
| Sun | BIC | 124814.0 | **122542.7** | 122788.6 | 122872.7 | 122946.6 | 122988.3 | 122987.1 |
| | log-lik | -61990.1 | **-60968.8** | -61029.7 | -61016.3 | -61014.4 | -61024.5 | -61023.4 |
| | dist. | 0.0 | **2490.9** | 2241.5 | 2171.8 | 2106.9 | 2065.3 | 2066.9 |

Table 3.5: BIC- and log-likelihood-values for each DDAY and No. Segment combination. The distance-value reports the $L^2$-Norm of a model with No. Segments $K > 1$ to the model with $K = 1$. The numbers in bold indicates the model with the greatest distance (dist.) to the model with $K = 1$.

For example, there are two price insensitive segments, with the exception of Friday flights where there is only one. For flights departing during the weekend the optimal number of segments is $K = 2$, indicating less customer heterogeneity. For all seven departure days the individual segments exhibit significant differences in price elasticity, which can be exploited for variable pricing purposes.

In Section 4 of the Web Appendix, we validate the assumption of conditional independence of flight counts during a departure day. To do so, we extend the univariate model to a multivariate Poisson model to analyse possible dependencies between flights. No significant dependence between flights is found, and we conclude that the proposed mixture-of-experts model is unbiased by unobserved heterogeneity caused by additional dependence between demand for flights departing on the same day.

## 3.6 Conclusion

We propose a flexible nonhomogeneous Poisson model of demand for passenger flights and apply it to a large dataset constructed from the booking database of a major airline. The dataset contains daily booking counts for all flights on a single busy short-haul route, where the airline has no direct competition. In comparison to most previous studies, our data do not suffer from the exclusions typical of data constructed either using web crawlers or sourced from the Global Distribution System. Our empirical study reveals four substantive findings with managerial and marketing implications for airlines.

First, based on the BIC criteria (see, Table 3.5), our latent segmentation model suggests that there are typically between two and five consumer segments, which have very different levels of price elasticity. Using an MNL model, we show that the probability of segment membership varies substantially over the flight departure time, booking day type and number of days to departure at the time of booking in a nonlinear way, so that price elasticity does so also. Quantifying variable price elasticities, as a mixture of passenger segments, is essential for revenue management practices where the airlines try to maximize their revenue by optimally changing the price of a ticket. From a marketing perspective, the characterization of customer segments in Table 3.4 allows AirABC to better tailor its product and promotion activities.

Second, we consider a booking horizon of 120 days, which is longer than in most previous studies. During this period, as seen by the varying segment proportions of Figure 3.5, we

| | No. Segments | | | | | | |
|---|---|---|---|---|---|---|---|
| DDAY | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ |
| Mon | 0.0011<br>(0.0006) | 0.0003<br>(0.0014) | -0.0005<br>(0.0051) | -0.0039<br>(0.0088) | -0.0210<br>(0.0037) | -0.0277<br>(0.0021) | -0.0282<br>(0.0021) |
| Tue | -0.0011<br>(0.0005) | -0.0039<br>(0.0016) | -0.0214<br>(0.0012) | -0.0220<br>(0.0012) | -0.0227<br>(0.0012) | - | - |
| Wed | -0.0008<br>(0.0004) | -0.0077<br>(0.0047) | -0.0214<br>(0.0021) | -0.0225<br>(0.0019) | -0.0232<br>(0.0019) | - | - |
| Thr | 0.0019<br>(0.0006) | -0.0035<br>(0.0022) | -0.0287<br>(0.0082) | -0.0295<br>(0.0025) | -0.0301<br>(0.0025) | - | - |
| Fri | 0.0007<br>(0.0003) | -0.0160<br>(0.0018) | -0.0206<br>(0.0021) | -0.0354<br>(0.0015) | -0.0361<br>(0.0015) | - | - |
| Sat | -0.0037<br>(0.0007) | -0.0368<br>(0.0020) | - | - | - | - | - |
| Sun | -0.0027<br>(0.0007) | -0.0435<br>(0.0018) | - | - | - | - | - |

Table 3.6: Segment specific price-coefficients and bootsrapped standard errors for the optimal endogeneity corrected model seperated by DDAY.

find the determinants of demand (and elasticity) vary greatly, suggesting that continuous tailoring of price and marketing over the entire booking horizon is warranted.

Third, the covariates used in our model are all fully observable throughout the airline scheduling horizon of 365 days before departure and allow for forecasting of elasticity and demand for use in airlines' revenue management systems. In contrast, capturing consumer heterogeneity using individual customer level data that includes some customer characteristics would not allow for forecasting future demand and price elasticity because this data is typically unknown to the airline at the time of booking. Moreover, retention of individual-level customer data is likely to be increasingly difficult under data privacy provisions, such as the EU General Data Protection Regulation (GDPR).

Last, we highlight the importance of accounting for endogeneity when estimating price elasticity. While studies have shown this previously for aggregate data, we do so at a disaggregate level within a flexible mixture-of-experts framework with nonlinear effects captured using regularized splines. A control variate approach is used with the bid-price as an instrument, which is discussed in detail for two latent passenger segments in Section 3 of the Web Appendix. The advantage of using the bid-price is that it varies at the same resolution as our booking data—i.e. at the flight and daily level—and proves to be a strong instrument.

Our study uses data from customers purchasing published fares for the economy class cabin on a single route without any competition from other airlines. The advantage of focusing on this specific situation is that it can be seen as a controlled experiment. Nevertheless, the model developed is applicable more generally. It has been applied by AirABC to bookings on other routes with competitors and a varying share of passengers who buy published fares. To model and forecast demand in those scenarios, additional variables are simply added to describe the behaviour of competitors and passenger segments.

The extension of the model to a multivariate Poisson model using a Gaussian copula, as outlined in Section 4 of the Web Appendix, has strong potential. While we found little evidence of additional dependence between bookings on flights that depart on the same day, it can also be used to capture dependence between other bookings. For example, between bookings for (i) the same flight on adjacent days (which would be a type of longitudinal model) and (ii) different flights departing during the same hourly period but in adjacent days. Such analyses would enable a better understanding of how price variation at the flight and daily level affect demand for substitute flights and provide a step towards improved continuous pricing by airlines.

Our research was undertaken before the 2020 COVID-19 pandemic, which at the time of writing, has greatly affected flights around the world. However, as air travel resumes the insights listed above are likely to remain valid. This is because our statistical model has interpretable components, whereas black-box models (e.g. deep neural networks) are often difficult to extrapolate in the presence of a structural shock. We conclude by noting that prior to March 2020, insights from these results were incorporated into practice by AirABC.

Tickets on the considered route, as well as on comparable connections, were priced based on the proposed model. As air travel recovers post-COVID-19, AirABC will likely continue to price tickets using this model, while incorporating adjustments to key components (notably the baseline intensity) to reflect new demand realities as they emerge.

# Chapter 4

# Conclusion

In conclusion, this thesis has presented two novel modeling approaches for estimating price sensitivity in the context of airline revenue management. The first model, an augmented generalized additive model, offers a closed-form solution for continuous pricing and has demonstrated superior predictive performance compared to state-of-the-art alternatives. The second model, a finite mixture model, has uncovered latent customer segments and has provided interpretable estimates of demand and price elasticity for flights booked on different days before departure. It is essential to highlight that both models are ready to be used within the revenue management systems of airlines, where the augmented generalized additive model has demonstrated practical applicability and is, therefore, used by AirABC for a vast majority of their daily flight operations.

Despite these contributions, there remain important avenues for future research. One area is the estimation of price elasticity in the context of dynamic bundling and a-la-carte shopping of ancillary products such as advanced seat reservations, seating options, and meals, to name a few. With airlines increasingly offering personalized bundles of ancillary products and services, it is essential to understand how customers trade off different elements of the bundle and how pricing affects demand.

Another area of future research is the estimation of cross-price elasticity between the airline's and similar competitors' offers. Estimating cross-price elasticity enables airlines to understand the competitive landscape better and optimize their pricing strategies accordingly.

Finally, there is a need to identify anomalies within the data where natural disasters, political unrest, or pandemics influence price elasticity and demand. Developing methods to detect and account for such anomalies will be crucial for airlines to make accurate revenue management decisions in real-time.

Overall, this thesis has advanced the field of airline revenue management by developing innovative models for estimating price sensitivity and providing insights into customer segmentation and pricing strategies. These future research areas will further enhance our understanding of the airline industry's complex dynamics between pricing, demand, and customer behavior.

# Appendices

## A.1 Penalized Likelihood Estimation Approach

We approximate each of the unknown function $f_p(\cdot), f_t(\cdot)$, and $f_k(\cdot)$ by a weighted sum of local P-spline (penalised B-spline) basis functions (e.g. Eilers and Marx, 1996). For example, the representation of $f_k(\cdot)$ is represented by a B-spline basis function given by the column vectors $(\mathbf{b})_{k,j}(\cdot), j = 1, \dots m$. The $n \times m$ matrix of basis functions $(\mathbf{b})_k(\cdot) = \big((\mathbf{b})_{k,1}(\cdot), (\mathbf{b})_{k,2}(\cdot), \dots, (\mathbf{b})_{k,m}(\cdot)\big)$ is subsequently multiplied by a $m \times 1$ vector of weighting coefficients $(\gamma)_k$. Therefore, function $f_k(\cdot)$ is approximated by $(\mathbf{b})_k(\cdot)(\gamma)_k$. The column dimension $m$ depends on the number of knots and degree of the B-spline functions (see, e.g. Marsh and Marshall, 1999, Ch. 8, p. 187). The bivariate functions $f_{p,t}(\cdot,\cdot), f_{p,k\in I_2}(\cdot,\cdot)$, and $f_{t,k\in I_2}(\cdot,\cdot), f_{\substack{k_1<k_2\\ k_1,k_2\in I_2}}(\cdot,\cdot)$ are also replaced in this manner. For example, $f_{k_1,k_2}(\cdot,\cdot)$ is replaced by $(\mathbf{b})_{k_1,k_2}(\cdot,\cdot)(\gamma)_{k_1,k_2}$, where the matrix of basis functions $(\mathbf{b})_{k_1,k_2}(\cdot,\cdot)$ is built from their univariate marginal basis terms as $(\mathbf{b})_{k_1,k_2}(\cdot,\cdot) = (\mathbf{b})_{k_1}(\cdot) \otimes (\mathbf{b})_{k_2}(\cdot)$. The Kronecker product $\otimes$ is calculated row-wise fashion.

Estimating the parameters in equation (2.3), requires identifiability constraints on the spline representations of the functions. Following Hastie and Tibshirani (1987) and Wood (2017), we require that the univariate functions $f_p(\cdot), f_t(\cdot)$, and $f_k(\cdot)$ integrate out to zero. For the bivariate functions $f_{t,k\in I_2}(\cdot,\cdot)$ and $f_{\substack{k_1<k_2\\ k_1,k_2\in I_2}}(\cdot,\cdot)$, we follow Lee and Durbán (2011), by applying the mixed model framework for smoothing. Lee and Durbán (2011) prove that the imposed constraints are equal to a classical factorial design. For the functions $f_{p,t}(\cdot,\cdot)$ and $f_{p,k\in I_2}(\cdot,\cdot)$, we exclude the first column of each marginal B-spline basis to achieve identifiability. For example, $f_{p,t}(\cdot,\cdot)$ is replaced by $(\mathbf{b})_{p,t}(\cdot,\cdot)(\gamma)_{p,t}$, where the matrix of basis functions $(\mathbf{b})_{p,t}(\cdot,\cdot)$ is built from $(\mathbf{b})_{p,t}(\cdot,\cdot) = \big((\mathbf{b})_{p,2}(\cdot), \dots, (\mathbf{b})_{p,m_p}(\cdot)\big) \otimes \big((\mathbf{b})_{t,2}(\cdot), \dots, (\mathbf{b})_{t,m_t}(\cdot)\big)$.

Having achieved identifiability, we compute parameter values by penalised maximum likelihood estimation. The penalty balances model flexibility and parsimony. The parameters of model (2.3) are given by $(\theta) = ((\beta), (\gamma))^T$. Here, $(\beta) = (\beta_0, (\beta)_1, \dots, (\beta)_{p,I_1})^T$ concerns parametric covariates, and the coefficient vector for the unknown functions is $(\gamma) = ((\gamma)_p, (\gamma)_t, (\gamma)_1, \dots, (\gamma)_{I_2}, (\gamma)_{p,t}, (\gamma)_{p,1}, \dots, (\gamma)_{p,I_2}, (\gamma)_{t,1}, \dots, (\gamma)_{t,I_2}, (\gamma)_{1,2}, \dots, (\gamma)_{I_2-1,I_2})^T$. This creates a feasible semi-parametric model, where a high-dimensional basis is employed and smoothed by imposing a penalty on $(\gamma)$. The optimal values of the smoothing parameters are selected using the Bayesian Information Criterion (BIC) (see also Claeskens and Hjort, 2008, p. 100-102).

The (unpenalized) log-likelihood $\ell((\theta))$ arising from Equation (2.1) and (2.2) is

$$\ell((\theta)) = \sum_{i=1}^{M} \sum_{t=t_i^{\text{close}}}^{t_i^{\text{open}}} y_{i,t} \log \big(\lambda((\mathbf{b})_{i,t}, t; (\theta))\big) - \lambda((\mathbf{b})_{i,t}, t; (\theta)), \tag{1}$$

where the intensity $\lambda((\mathbf{x})_{i,t}, t; (\theta))$ in equation (2.3) is written as a function of the covariates and the model parameters.

We maximize equation (1) with an additive penalty to regulate the degree of smoothness for every function of (2.3). The structure of (1) allows us to calculate the first and second-order derivatives analytically. Thereby, the derivatives are quickly evaluated and maximizing the likelihood proves straightforward by quasi Newton-Methods, such as the Broyden-Fletcher-Goldfarb-Shanno algorithm (e.g. Broyden, 1970). The penalization is based on the ideas of Eilers and Marx (1996). We impose a penalty on the coefficients relating to all functional effects that define the demand model (2.3). We use linear B-splines for the marginals that concern $p_{i,t}$ and take quadratic B-splines otherwise. For the functions that have no shape constraint, i.e., $f_t, f_k, f_{t,k}, f_{k_1,k_2},\ k, k_1, k_2 \in I_2, k_1 < k_2$, we penalize neighbouring coefficients of second order.

For example, let $(\mathbf{b})_t(\mathrm{t})$ be the quadratic B-spline bases for the main effect $f_t(t)$ with column dimension $m_t$ and $(\gamma)_t$ as the vector of weights. By penalizing second order differences, i.e., $\Delta^2\gamma_{t,l} = \gamma_{t,l} - 2\gamma_{t,l-1} + \gamma_{t,l-2}, l = m_t, \ldots, 3$. With the $(m_t - 2) \times m_t$ matrix

$$(\mathbf{P})_t = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \tag{2}$$

the quadratic penalty for $f_t(t)$ is defined by $(\gamma)_t^T(\mathbf{S})_t(\gamma)_t$, where $(\mathbf{S})_t = (\mathbf{P})_t^T(\mathbf{P})_t$. Two penalty matrices $(\mathbf{S})$ exist for bivariate effects, one for each dimension.

For the constrained functions $f_p, f_{p,t}$ and $f_{p,k}, k \in I_2$, we follow Pya (2010), who discusses uni- and bi-variate as well as single and double constraint functions. However, building the bivariate functions $f_{p,t}$ and $f_{p,k}, k \in I_2$ without the intercept values requires some adjustments. The first concerns the function $f_{p,t}$, which has a double monotonicity constraint. Here, we remove the first row and column vector from the matrices $(\mathbf{\Sigma})_j, j = 1, 2$ (Pya, 2010, p. 58). For the functions $f_{p,k}, k \in I_2$ with a monotonicity constraint along the first dimension $p_{i,t}$, we remove the first row and column vector of $(\mathbf{\Sigma})_1$ and $(\mathbf{I})_2$ (Pya, 2010, p. 58). Secondly, for every bivariate function with monotonicity constraint, the penalty matrix $(\mathbf{S})_j = (\mathbf{P})_j^T(\mathbf{P})_j, j \in \{1, 2\}$ is built from $(\mathbf{P})_j$ without the first diagonal block element $(\mathbf{P})_{uj}$ (Pya, 2010, p. 60). The penalty matrix adjustments for $f_{k_1,k_2}$ are discussed by Lee and Durbán (2011).

The penalised likelihood $\ell_p(...)$ is defined by the unpenalised version (1) plus the sum of the weighted quadratic penalties. Thus, for every function of model (2.3), the penalty matrix $(\mathbf{S})$ is multiplied by a weighting factor $\rho$. For example, the weighted penalty term for $f_t(t)$ is defined by $\rho_t(\gamma)_t^T(\mathbf{S})_t(\gamma)_t$. Collecting all weighted penalty matrices for every function finally leads to the expression:

$$
\begin{aligned}
\ell_p((\theta),(\rho)) = \ell((\theta)) &+ \rho_p(\gamma)_p^T(\mathbf{S})_p(\gamma)_p + (\gamma)_{p,t}^T\Big(\rho_{p,t,1}(\mathbf{S})_{p,t,1} + \rho_{p,t,2}(\mathbf{S})_{p,t,2}\Big)(\gamma)_{p,t} \\
&+ \sum_{k\in I_2}(\gamma)_{p,k}^T\Big(\rho_{p,k,1}(\mathbf{S})_{p,k,1} + \rho_{p,k,2}(\mathbf{S})_{p,k,2}\Big)(\gamma)_{p,k} \\
&+ \rho_t(\gamma)_t^T(\mathbf{S})_t(\gamma)_t + \sum_{k\in I_2}\rho_k(\gamma)_k^T(\mathbf{S})_k(\gamma)_k \\
&+ \sum_{k\in I_2}(\gamma)_{t,k}^T\Big(\rho_{t,k,1}(\mathbf{S})_{t,k,1} + \rho_{t,k,2}(\mathbf{S})_{t,k,2}\Big)(\gamma)_{t,k} \\
&+ \sum_{\substack{k_1<k_2\\ k_1,k_2\in I_2}}(\gamma)_{k_1,k_2}^T\Big(\rho_{k_1,k_2,1}(\mathbf{S})_{k_1,k_2,1} + \rho_{k_1,k_2,2}(\mathbf{S})_{k_1,k_2,2}\Big)(\gamma)_{k_1,k_2},
\end{aligned} \tag{3}
$$

where $(\rho) = \left(\rho_p, (\rho)_{p,t}, (\rho)_{p,1}, \ldots, (\rho)_{p,I_2}, \rho_t, \rho_1, \ldots, \rho_{I_2}, (\rho)_{t,1}, \ldots, (\rho)_{t,I_2}, (\rho)_{1,2}, \ldots, (\rho)_{I_2-1,I_2}\right)^T$ refers to the vector of penalty parameters, weighting the quadratic penalties. Penalty parameters in bold correspond to column vectors. The first row gives the penalty of the first and the second row for the second dimension. For $(\rho) = 0$, one obtains unpenalized estimations.

The penalty parameters are selected using the Bayesian Information Criterion defined through

$$
BIC_\gamma((\rho)) = -2\ell_p((\theta),(\rho)) + \gamma\log(n)\mathrm{df}((\rho)), \tag{4}
$$

where $n$ is the number of observations ($\approx$ number of flights multiplied by the number of considered days to departure) and $\gamma$ inflates the influence of df to increase the smoothness of the fit. The model degree of freedom $\mathrm{df}((\rho))$ can be calculated through Fisher Matrices. Thus, let $F((\theta),(\rho))$ denote the penalized Fisher matrix, i.e.

$$
F((\theta),(\rho)) = E\left(-\frac{\partial\ell_p((\theta),(\rho))}{\partial(\theta)\partial(\theta)^T}\right). \tag{5}
$$

Then, the model degree can be approximated as

$$
\mathrm{df}((\rho)) = \mathrm{trace}\{F^{-1}(\widehat{(\theta)},(\rho))F(\widehat{(\theta)},(\rho)=0)\},
$$

see e.g. Krivobokova and Kauermann (2007a). To estimate (2.5), we first maximise (L.12.2) for $(\rho) = 0$. Secondly, given the estimate $\widehat{(\theta)}$, we estimate $(\rho)$ by minimising $BIC_\gamma((\rho))$. The corresponding estimate $\widehat{(\rho)}$ is subsequently used to maximize (L.12.2) once more. We alternate the maximisation of (L.12.2) and minimisation of $BIC_\gamma((\rho))$ until $\left\|\frac{\partial BIC_\gamma((\rho))}{\partial(\rho)}\right\|$, as calculated after the maximisation of (L.12.2), falls below a fixed threshold $\epsilon = 10^{-4}$.

## B.2 Two-Staged Estimation by Residual Inclusion

Treating price as an exogenous variable in a consumer demand model can lead to biased estimates of price elasticity; see discussions in Davidson and MacKinnon (1999, 1993), Wooldridge (2002), Petrin and Train (2010) and references therein. For example, Mumbower et al. (2014) show the importance of controlling for price endogeneity in a linear model for flight bookings using a two-stage least squares linear regression estimator, whereas Lurkin et al. (2017) do so for a choice model. For generalized nonlinear models, Marra and Radice (2011) suggest an extension of such two-stage estimators, similar to the control function approach of Petrin and Train (2010). We follow these authors and first build a nonlinear model for price based on an instrumental variable, and then include the price residual as a covariate in our model of passenger demand. To do so, we model the logarithm of prices at the daily and flight levels as

$$\begin{aligned} \log(p_{i,t}) &= \theta_0 + \theta_1 \mathrm{IV}_{i,t} + \sum_{k \in I_1} \sum_{j \in J_k} (\mathbf{1})_{\{x_{k,i,t}=j\}} \theta_{k,j} + s_t\left(t\right) + \sum_{k \in I_2} s_k\left(z_{k,i,t}\right) + u_{i,t} \\ &= \eta_{i,t} + u_{i,t}\,, \end{aligned} \tag{6}$$

where $u_{i,t} \sim N(0, \sigma^2)\ \forall\ i, t$. The effects of $t$ and $z_{k,i,t}, k \in I_2$ are captured by unknown smooth functions $s_t\left(\right)$ and $s_k\left(\right), k \in I_2$ modelled by penalized splines, while $\mathrm{IV}_{i,t}$ is an instrumental variable.

Mumbower et al. (2014) discusses possible choices for IV and suitable candidates. Li et al. (2014) describes that almost all of these choices are invalid as the researcher needs to observe both the IV and booking data at the same level of aggregation to control for price endogeneity effectively. Supply shifters—for example, airport fees, transportation taxes and fuel costs—are constant over daily bookings. Hausman-style instruments at the firm level do not match a model at the market level. Stern-type instruments that measure competition and market share do not vary on the booking level. Last, IVs that impact marginal costs remain a feasible option. Similar to (Meyer et al., 2022), we use (the logarithm of) a variable that is popular in the revenue management literature called the 'bidprice' (Talluri and van Ryzin, 2005, pp. 31). The bidprice is a measure of the (marginal) cost of offering a seat, taking into account that it cannot be sold again. Crucially, it varies between bookings because the airline updates its assessment frequently. The bidprice is available for all flights in the database and all time points and for prediction purposes for flights that have yet to depart.

To ensure the validity of our choice, the IV needs to fulfil the properties of relevance and exogeneity (Guevara, 2018). Whereas (strong) relevance can easily be demonstrated by the strong nonlinear dependence between the IV and the endogenous variable price, exogeneity needs to be addressed by a statistical (over-identification) test. Unfortunately, this test requires the availability of at least two instruments, so exogeneity can only be established definitely. From a qualitative perspective, the bidprice is a measurement of displacement cost, ensuring that revenue gain for the available airlines' network capacity

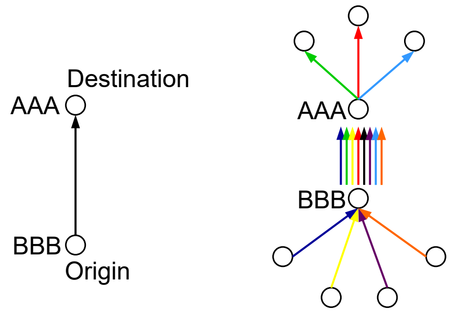



Figure 1: Description of two Airline-network scenarios. On the left-hand side, the airline controls for capacity constraints only taking passenger demand from the origin (BBB) to the destination (AAA) into account. Low-cost carriers typically use this setup. On the right-hand side, the airline controls for the capacity constraint on the BBB to AAA route by taking all possible passenger demand streams coming from other origins than BBB (arrows going into BBB) to different destinations than AAA (arrows going out of AAA) into account. Network carriers typically use this setup.

is maximized. As pointed out by Li et al. (2014), the exogeneity (and hence the validity of the bid-price IV) means that a demand shock for flight $i$ at the time to departure $t$ (i.e. $\varepsilon_i(t) = Y_i(t) - \lambda((\mathbf{x})_{i,t}, t)$, where $Y_i(t) = N_i(t) - N_i(t-1)$) is uncorrelated with the IV. Figure 1 describes two possible revenue management setups, where an airline only controls displacement cost on the route level (left-hand side) or incorporates all potential demand streams into the displacement cost calculation (right-hand side). AirABC is a network carrier that considers every demand stream when calculating the bidprice value. Therefore the bidprice defines the distribution of network demand on the route level. In our study, the share of transfer passengers, i.e., passengers not travelling solely between BBB and AAA, is approximately 50%. Thus, the bid-price value is largely determined by factors exogenous to the route under study. Hence, we conclude that the demand shock $\varepsilon_i(t)$ and the bidprice are uncorrelated.

After the parameters of the model (2.4) are estimated using maximum likelihood, the error

$$\xi_{i,t} = p_{i,t} - \mathbb{E}\left(p_{i,t} \mid \mathrm{IV}_{i,t}, z_{1,i,t}, \ldots, z_{I_2,i,t}\right) = p_{i,t} - \exp\left(\eta_{i,t} + \hat{\sigma}^2/2\right)$$

is estimated for each flight and booking day combination, where the squared residual standard error is calculated as

$$\hat{\sigma}^2 = \frac{1}{n - \mathrm{df}(\widehat{\theta})} \sum_{i=1}^{M} \sum_{t=t_i^{\mathrm{close}}}^{t_i^{\mathrm{open}}} \widehat{u}_{i,t}^2, \tag{7}$$

The resulting residuals values are observations on the covariate $\hat{\xi}_{i,t}$, which is included in the demand model (2.3) as an additional regressor.

## C.3 Model adjustments for price data on booking level

The ticket price $p_{i,t}$ can change during booking day $t$, but covariates BDAY, DTIME, YDAY, and $t$ are fixed. Therefore, the price is observed per booking, but BDAY, DTIME, YDAY, and $t$ are given per day. The likelihood of model (2.5) is augmented to incorporate different resolution levels for the observations, specifically price variation on booking level. The augmentation aims to maintain the price information $p_{i,t}$ per booking.

Suppose for model (2.5) with likelihood (1), three bookings are observed on a single day. We assume an aggregation level of $\frac{1}{3}$ day in that case. To specify this here, let

$$(\mathbf{x})_{i,t,l} = (\text{BDAY}_{i,t}, p_{i,t,l}, \text{DTIME}_{i,t}, \text{YDAY}_{i,t}, t)$$

be the covariate vector for the $l$th booking observed $t$ days to departure for flight $i$, where $l = 1, ..., \max(1, y_{i,t})$. On days without bookings of flight $i$ (i.e. when $y_{i,t} = 0$), let $(\mathbf{x})_{i,t,1}$ be the vector of covariate values, and set $y_{i,t,1} = 0$. Similarly, let $y_{i,t,l} = 1$ for $l = 1, ..., \max(1, y_{i,t})$ for days with observed bookings ($y_{i,t} \geq 1$). Then, the (unpenalized) log-likelihood is:

$$\ell((\theta)) = \sum_{i=1}^{M} \sum_{t=t_i^{\text{close}}}^{t_i^{\text{open}}} \sum_{l=1}^{\max(1,y_{i,t})} y_{i,t,l} \log\Big(\lambda((\mathbf{x})_{i,t,l}, t; (\theta))\Big) - \frac{\lambda((\mathbf{x})_{i,t,l}, t; (\theta))}{\max(1, y_{i,t})} . \tag{8}$$

The additional inner summation in Equation (8) runs over all observed bookings ($y_{i,t,l} = 1$) during one booking day $t$ and flight $i$. This summation drops out for days without bookings ($y_{i,t,l} = 0$).

## D.4 Benchmarking

We benchmark our approach against a heuristic, a parametric and a nonparametric model. As a representative heuristic, we select the model proposed by Weatherford and Pölt (2002) (WP). WP imputed the mean number of bookings as a demand estimate for days where the airline did not offer a fare. As a parametric approach, we select FCST of Fiig et al. (2014). This approach models demand depending on BDAY, DTIME, YDAY, t, and $p_{i,t}$. As a nonparametric approach, we select EM by Vulcano et al. (2012). EM requires price variation for all flights $i$ and values of $t$. However, for the analysed data set, this is only sometimes given. So instead of using the observed bookings with prices $p_{i,t}$, we calculate the average number of bookings for all observable prices. This logic imposes price variation for fixed values of $t$ by dropping the flight index $i$. Therefore, EM predicts the average demand level per flight without considering season or departure time effects. Hence, we expect EM to perform best when demand does not depend on season or departure time and worse otherwise.

| Label | Flexibility | Reference |
|---|---|---|
| WP | low | Weatherford and Pölt (2002) |
| FCST | medium | Fiig et al. (2014) |
| EM | high | Vulcano et al. (2012) |
| our model | high | Eq. (2.4), (2.5) |

Table 1: Properties of forecasting models for benchmarking.

Table 1 lists the benchmarked models per represented family. The second column assesses model flexibility: nonparametric models provide more flexibility than parametric models. WP is rated as less flexible than all alternatives, as it ignores the information contributed by the covariates BDAY, DTIME, YDAY, and t.

We measure the prediction error by K-fold cross-validation to quantify the forecasting accuracy. The smallest prediction error indicates the best demand estimate. We evaluate the prediction error per product, i.e., per flight $i$. To that end, we aggregate observed bookings $y_{i,t}$ and demand estimates $\widehat{\lambda}((\mathbf{x})_{i,t}, t) \equiv \widehat{\lambda}_{i,t}$ over $t$: $y_i = \sum_{t=t_i^{\text{close}}}^{t_i^{\text{open}}} y_{i,t}$ and $\widehat{\lambda}_i = \sum_{t=t_i^{\text{close}}}^{t_i^{\text{open}}} \widehat{\lambda}_{i,t}$. To create $K$ roughly equal-sized folds of data (indexed by $k \in \{1, \ldots, K\}$) from $M$ products ($K << M$), we randomly draw $m = \left\lfloor \frac{M}{K} \right\rfloor$ products, $K$-times without replacement. Finally, for each competing model, the cross-validation estimate of the prediction error $\text{CV}\left(\widehat{\lambda}\right)$ is

$$\text{CV}\left(\widehat{\lambda}\right) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i=1}^{M_k} \text{L}\left(y_i, \widehat{\lambda}_i^{-k(i)}\right), \tag{9}$$

where prediction $\widehat{\lambda}_i^{-k(i)}$ is created by excluding the data of fold $k$. The loss $\text{L}\left(y_i, \widehat{\lambda}^{-k(i)}\right)$ results by forecasting $\widehat{\lambda}_i^{-k(i)}$ and observing $y_i$.

As loss functions $L(\cdot)$, we consider a selection of absolute and relative measures. We measure absolute deviations by the root mean squared error (RMSE) and the mean absolute deviation (MAD). Relative deviations are evaluated by the root mean squared logarithmic error (RMSLE) and the symmetric mean absolute percentage error (SMAPE), which are feasible if the target attains a value of zero (if no demand is observed). The definitions for RMSE, MAD, RMSLE, and SMAPE are

$$\text{RMSE} = \sum_{i=1}^{M} \left(y_i - \widehat{\lambda}_i^{-k(i)}\right)^2 \tag{10}$$

$$\text{MAD} = \sum_{i=1}^{M} \left|y_i - \widehat{\lambda}_i^{-k(i)}\right| \tag{11}$$

$$\text{RMSLE} = \sum_{i=1}^{M} \log\left(\frac{\widehat{\lambda}_i^{-k(i)} + 1}{y_i + 1}\right)^2 \tag{12}$$

$$\text{SMAPE} = \frac{\sum_{i=1}^{M} \left| y_i - \widehat{\lambda}_i^{-k(i)} \right|}{\sum_{i=1}^{M} y_i + \widehat{\lambda}_i^{-k(i)}} \tag{13}$$

Figure 2 reports the resulting average cross-validation estimates $\overline{\text{CV}}(\widehat{\lambda})$ per benchmarked approach and sample size. Two P2P connections and seven departure days yield 14 combinations per OD. Thus, the average cross-validation estimate for prediction $\widehat{\lambda}$ is calculated as $\overline{\text{CV}}(\widehat{\lambda}) = \frac{1}{14}\sum_{j=1}^{14} \text{CV}_j$. Figure 2 shows that the two absolute measures tend to increase in the sample size, whereas the relative measures RMSLE and SMAPE decrease. Our approach ranges at the top independent of the sample size, even though FCST performs almost as well. The weak performance of EM originates from not considering seasonal or departure time dependencies but being dependent on aggregated data. The relative measures RMSLE and SMAPE highlight the superior performance of our model. As Bartke (2014) point out, small observations result if disaggregated booking data is used for demand estimation. Therefore, the final judgment should focus on relative forecasting performance as quantified by RMSLE and SMAPE.

## E.5 Proof of the discrete pricing problem

Given a discrete set of price points $\Omega_p = \{p_1, \ldots, p_J\}$, the optimal price $p_{i,t}^\star$ (2.8) defines the lower boundary point of the subset $\Omega_p' \subset \Omega_p$ of prices that are profitable to be offered. To show that $p_{i,t}^\star$ defines the boundary point of the set $\Omega_p'$, every price below ($p_k < p_{i,t}^\star$) has to have a marginal revenue contribution that is smaller than $\pi_{i,t}$ (bid-price) and a price above or equal ($p_j \geq p_{i,t}^\star$) has to have a marginal revenue contribution that is greater than $\pi$.

*Proof*: for simplicity all indices are dropped
For $p_k < p^\star$:

$$\begin{aligned} \lambda(p_k)r_k - \lambda(p_k)\pi &< \lambda(p^\star)r^\star - \lambda(p^\star)\pi \\ \iff \lambda(p_k)r_k - \lambda(p^\star)r^\star - \pi\left(\lambda(p_k) - \lambda(p^\star)\right) &< 0 \\ \iff \frac{\lambda(p_k)r_k - \lambda(p^\star)r^\star}{\lambda(p_k) - \lambda(p^\star)} &< \pi \end{aligned} \tag{14}$$

For $p_j > p^\star$:

$$\begin{aligned} \lambda(pj)r_j - \lambda(p_j)\pi &< \lambda(p^\star)r^\star - \lambda(p^\star)\pi \\ \iff \lambda(p_j)r_j - \lambda(p^\star)r^\star - \pi\left(\lambda(p_j) - \lambda(p^\star)\right) &< 0 \\ \iff \frac{\lambda(p^\star)r^\star - \lambda(p_j)r_j}{\lambda(p^\star) - \lambda(p_j)} &> \pi \end{aligned} \tag{15}$$

We conclude that prices below the optimal price ($p_k < p^\star$) have a marginal revenue contribution smaller than the bidprice and are therefore not included in the offer-set $\Omega_p'$ whereas

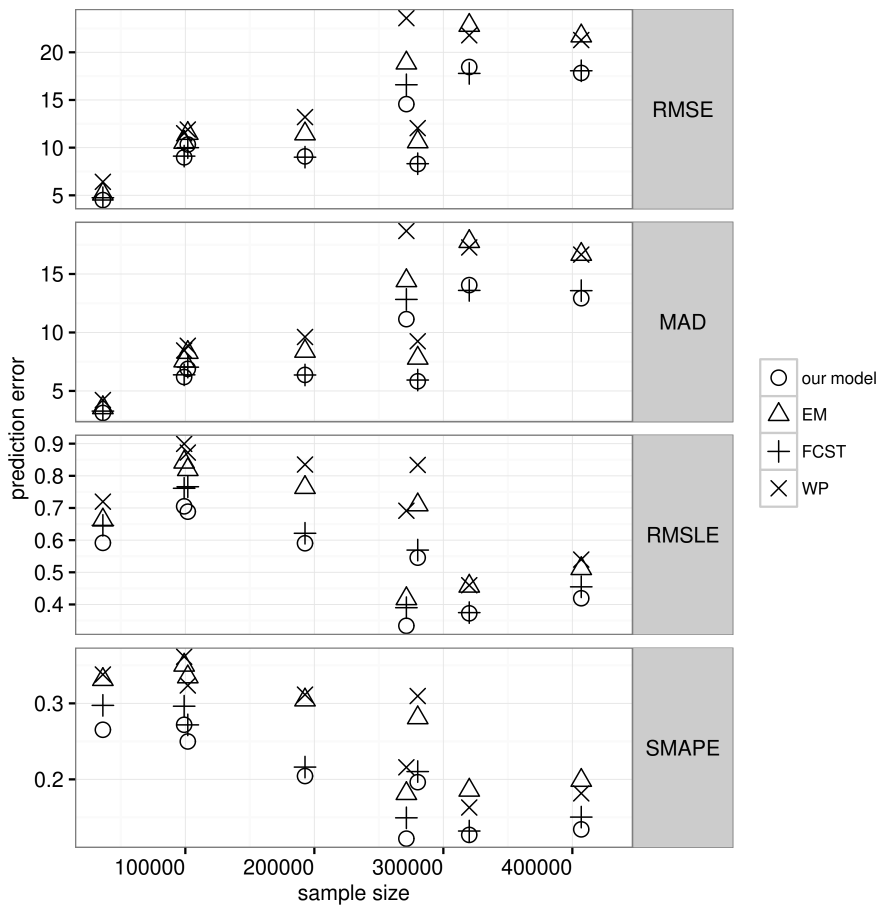



Figure 2: Estimates of the average prediction error $\overline{\mathrm{CV}}(\widehat{\lambda})$ versus the ODs sample size calculated by the criteria RMSE (1st row), MAD (2nd row), RMSLE (3rd row), and SMAPE (4th row). For every criterion, the model with the lowest prediction error among all other models (our model, EM, FCST, and WP) has the best forecasting performance.

prices equal or greater than the optimal price $p^\star$ have a marginal revenue contribution greater than the bidprice and are therefore included in $\Omega'_p$. Note that the first inequality of each case results as $p^\star$ is the optimal price that maximises (2.6).

## F.6 Calculation of the optimal continuous price

$$
\begin{aligned}
\frac{\partial\Big(\lambda((\mathbf{x})_{i,t},t)(r_{i,t}-\pi_{i,t})\Big)}{\partial p_{i,t}} &\stackrel{!}{=} 0\\
\iff (\widehat{r}^\star_{i,t}-\pi_{i,t})\frac{\partial\lambda((\mathbf{x})_{i,t},t)}{\partial p_{i,t}} + \lambda((\mathbf{x})_{i,t},t)\frac{\partial(\widehat{r}^\star_{i,t}-\pi_{i,t})}{\partial p_{i,t}} &= 0\\
\iff \big(-\widehat{\alpha}_0+(1-\widehat{\alpha}_1)p^\star_{i,t}-\pi_{i,t}\big)\,\frac{\partial\lambda((\mathbf{x})_{i,t},t)}{\partial p_{i,t}} + \lambda((\mathbf{x})_{i,t},t)(1-\widehat{\alpha}_1) &= 0\\
\iff \big(-\widehat{\alpha}_0+(1-\widehat{\alpha}_1)p^\star_{i,t}-\pi_{i,t}\big)\,\lambda((\mathbf{x})_{i,t},t)\frac{\partial\log\left(\lambda((\mathbf{x})_{i,t},t)\right)}{\partial p_{i,t}} + \lambda((\mathbf{x})_{i,t},t)(1-\widehat{\alpha}_1) &= 0\\
\iff \big(-\widehat{\alpha}_0+(1-\widehat{\alpha}_1)p^\star_{i,t}-\pi_{i,t}\big)\left((\mathbf{1})_{2\mathrm{s}}\beta_{\widehat{\xi}}+f'_p+f'_{p,t}(t)+\sum_{k\in I_2}f'_{p,k}(z_{k,i,t})\right) &= -\left(1-\widehat{\alpha}_1\right)\\
\iff -\frac{1-\widehat{\alpha}_1}{(\mathbf{1})_{2\mathrm{s}}\beta_{\widehat{\xi}}+f'_p+f'_{p,t}(t)+\sum_{k\in I_2}f'_{p,k}(z_{k,i,t})}+\widehat{\alpha}_0+\pi_{i,t} &= (1-\widehat{\alpha}_1)p^\star_{i,t}\\
\iff -\frac{1}{(\mathbf{1})_{2\mathrm{s}}\beta_{\widehat{\xi}}+f'_p+f'_{p,t}(t)+\sum_{k\in I_2}f'_{p,k}(z_{k,i,t})}+\frac{\widehat{\alpha}_0}{1-\widehat{\alpha}_1}+\frac{\pi_{i,t}}{1-\widehat{\alpha}_1} &= p^\star_{i,t}
\end{aligned}
\tag{16}
$$

In line two, we use the fact that the estimated regression model (2.7) gives $\widehat{r}_{i,t} = -\widehat{\alpha}_0 + (1-\widehat{\alpha}_1)p_{i,t}$. Thus, for $p_{i,t} = p^\star_{i,t}$ we get the corresponding revenue gain $\widehat{r}^\star_{i,t}$. The model for $\log\left(\lambda((\mathbf{x})_{i,t},t)\right)$ is defined by equation (2.3). Specifically, equation (2.5) describes the model that is applied to airline data. In line 5, we used the structure of the airline model, where $f'$ corresponds to $\frac{\partial}{\partial p_{i,t}}f$.

## G.7 Supplementary plots

Figure 3 shows the offered price per departure time and for three values of bidprice $\pi$ for OD1-5 and OD7.

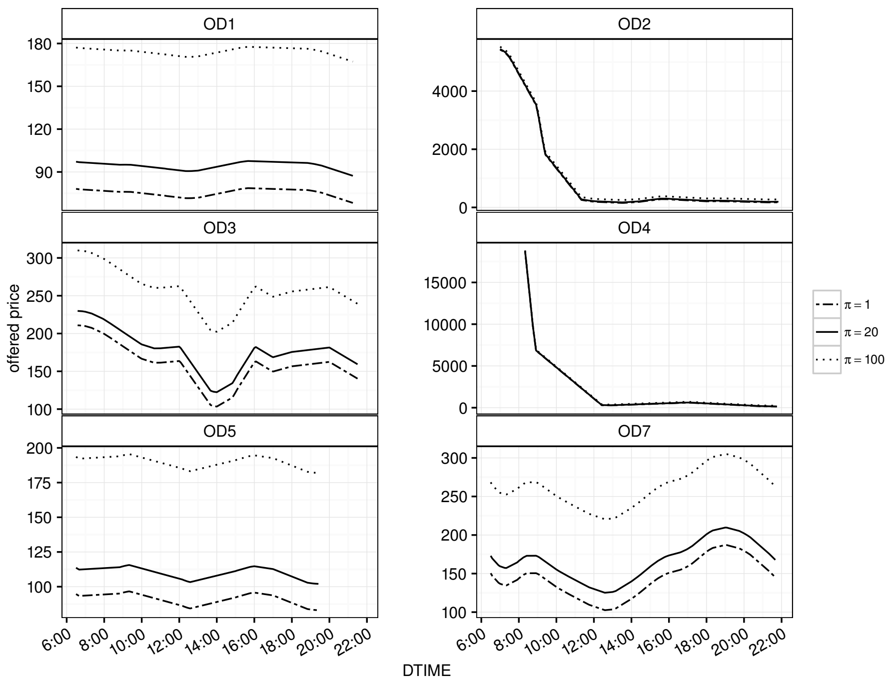
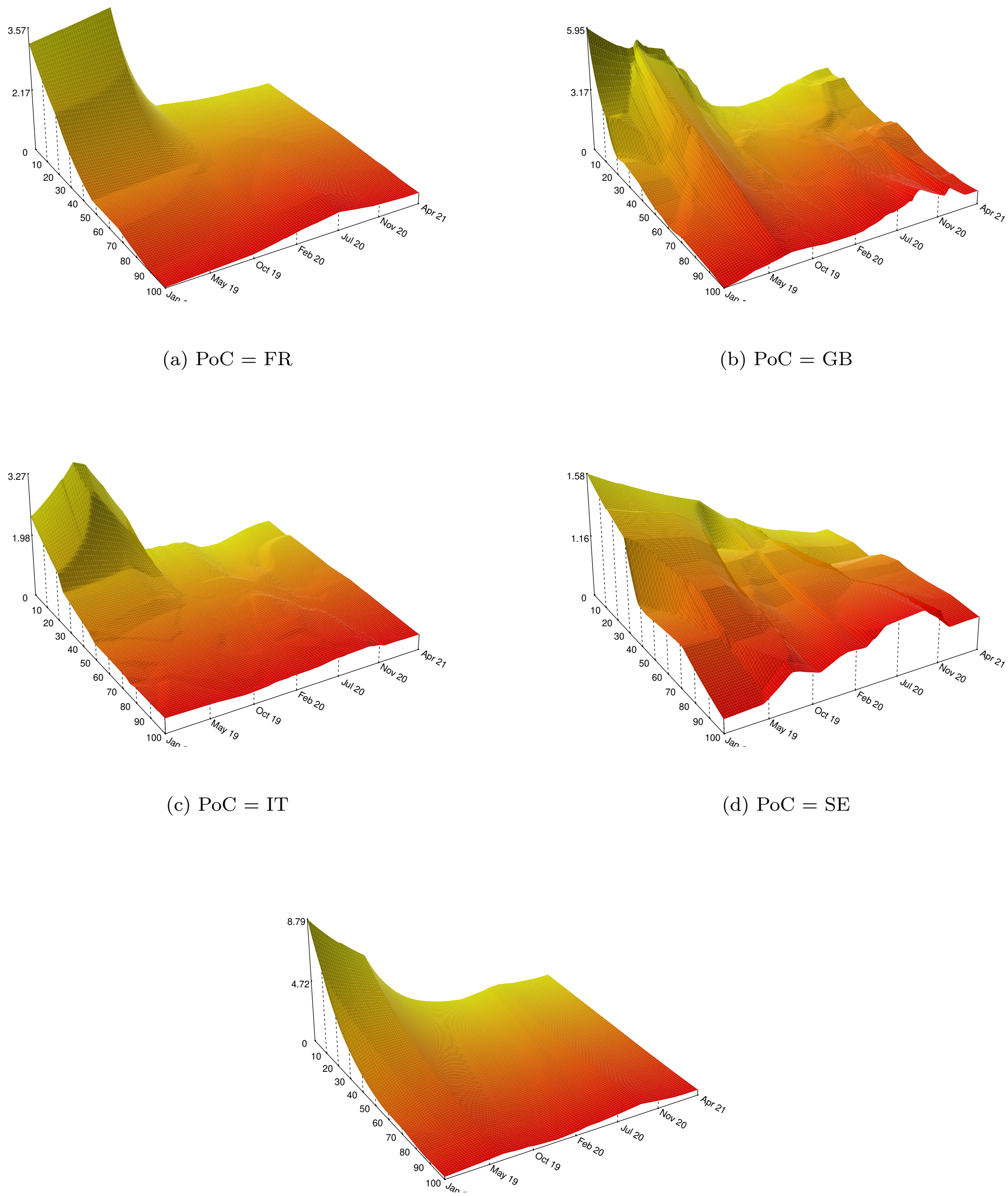



Figure 3: Optimal price values for OD1-5 and OD7 at DDAY=Thursday.

## H.8 Revenue Management Literature Review

Airlines' revenue management systems handle thousands of transactions per second, and sell decisions are due in milliseconds, so that no current revenue management system works in real time. Therefore, accurate estimates of demand and price elasticity are essential to precompute control values to maximise revenue from of its seat inventory (McGill and Van Ryzin, 1999). Early examples of price elasticity estimation include Jung and Fujii (1979), Oum et al. (1992), Brons et al. (2002), and Kremers et al. (2002), while Granados et al. (2012) study differences in price elasticities for flights due to distribution channel using a log-linear model. In the revenue management literature, multinomial discrete choice models of customer selection between available booking classes, cabin classes and/or flights times are popular; for examples, see Vulcano et al. (2010), Vulcano et al. (2012), and Dai et al. (2014). In contrast, we do not model individual customer choice, but model daily booking counts for a given flight and cabin class. Any remaining dependence in these counts due to customer choice between different flights departing on the same day is instead captured by the copula model. A number of authors also combine nonhomogeneous Poisson processes for the arrival of potential customers, with product choice models (Balaiyan et al., 2019). A 'no-buy' option is included in the choice set to accommodate potential customers who do not buy a ticket; see Vulcano et al. (2012), Besbes and Zeevi (2015), and Van Ryzin and Vulcano (2014). However, our model differs from this because we model the realized booking process — i.e. the bookings that are actually made — rather than the arrival of potential customers, including those who do not make a booking.




(a) PoC = FR

(b) PoC = GB

(c) PoC = IT

(d) PoC = SE

(e) PoC = ES

Figure 4: For the countries ES, FR, GB, IT, and SE, the graph shows how the factor $Q3^f$ changes for different days to departure, ranging from 100 days pre-departure to the day of departure and over the selling dates from January 2019 to April 2021.

Lo et al. (2015) and Li et al. (2014) also stress the importance of accounting for price endogeniety in models of demand in the airline industry, as do Mumbower et al. (2014) and Lurkin et al. (2017) who also make use of instrumental variables. Similarly, a number of other authors have also considered latent segmentation of customers when estimating demand and/or price elasticity using choice and other models; for example, see Teichert et al. (2008), Wen and Lai (2010), Martinez-Garcia and Royo-Vela (2010), Vij and Walker (2014), and Feldman and Topaloglu (2015). However, our study is the first of which we are aware that identifies such a rich latent segmentation of airline passenger bookings using a mixture-of-experts style model calibrated with a large disaggregate dataset. A similar modelling approach to our mixture-of-expert model is Li et al. (2014). The authors analyze strategic behaviour by a mixture of myopic and strategic customers, a special case of our model with two latent segments and no differentiation between the arrival time of strategic customers. Reviewing the statistical literature for two fundamental techniques accounting for unobserved heterogeneity, Sfeir et al. (2021) compare latent class with mixed logit models. Besides the advantages of latent class models, having fewer assumptions about the mixture distribution, being interpretable (as their mixture component typically depends on covariates — in our case, time to departure, departure time, and booking day of the week), and the correlation between the mixture component and the segment-specific variables and estimated elasticities are implicit in the model (mixed logit models need to assume a joint distribution for both components), the author mentions that latent class models may oversimplify the unobserved heterogeneity if the number of classes is small. To ensure that our model does not oversimplify the unobserved heterogeneity, up to 7 passenger segments are analyzed for each departure day.

Last, Wen and Chen (2017) account for the impact of the days to departure at booking on demand as a smooth nonlinear function, whereas Lurkin et al. (2017) do so for the flight departure time. Both papers employ parametric function bases constructed from low order Fourier terms. In contrast, following the statistical literature (Wood, 2017), we model both nonlinear effects using splines. These are more flexible and allow for data-driven levels of smoothing. Moreover, these two nonlinear effects are estimated for both the booking intensity and the mixture probabilities in the MNL.

Within Table 2, we summarizes the main features (sample size of data, usage of covariates, model-type, handling of endogeneity (Endo.), and usage of segmentation (Seg.)) of prior studies of passenger flight retail demand and price elasticity that are closest to ours.

Table 2: Comparison of relevant prior literature on modelling retail demand for passenger flights

| Author | Data | Covariates | Model | Endo. | Seg. |
|---|---|---|---|---|---|
| Dai et al. (2014) | Booking, ticketing, and availability data from 3 airlines between 2011 and 2012 (n = 748,076). | [1] Ticket price, [2] Departure Time, [3] Ticket change fee, [4] Milage gain, [5] Carrier, [6] Booking Time, [7] Booking Channel | MNL, Nested Logit, Mixed Logit | no | no |
| Mumbower et al. (2014) | JetBlue Webbot data for transcontinental flights between 2 and 22 September 2010 over a 28-day booking horizon (n = 7,522). | [1] Ticket price, [2] Departure day of week, [3] Departure Time, [4] Days to departure at booking, [5] Booking day of week, [6] Virgin America promotions, [7] Labor Day indicator | Ordinary Least Squares, Two Staged Least Squares | yes | no |
| Fiig et al. (2014) | Bookings at 22 selected traffic flows from Scandinavian Airlines (n = 7,780). | [1] Ticket price, [2] Departure day of week, [3] Departure Time, [4] Days to departure at booking, [5] Recurring special periods, [6] Departure Date | Nonlinear regression (multiplicative) | no | no |
| Vulcano et al. (2012) | Booking data from last 7 selling days for 11 Monday flights from January to March of 2004. | [1] 11 Products (fare-classes) with different fare-values, [2] 7 Booking Periods (each 24 hours), [3] 2 daily flights, [4] Market share | MNL | no | no |
| Teichert et al. (2008) | Stated preference survey data from frequent flyer passengers traveling on 11 European short-haul routes (n = 5,829). | [1] Product characteristics: compartment (business, economy), [2] Stated preferences: scheduled frequency, price, fare flexibility, punctuality, catering, ground service, [3] Behavioral and socio-demographic variables: gender, age, education level, profession, flying frequency | Latent Class | no | yes |

| | | | | | |
|---|---|---|---|---|---|
| Lurkin et al. (2017) | Ticket data purchased through travel agencies worldwide as collected by the Airlines Reporting Corporation (ARC) ($n = 10,034,935$). | [1] Itinerary Information (Departure Time, Travel Time, Equipment, Number of connections, Direct Flight Indicator), [2] Price (Average high yield fare, Average low yield fare), [3] Marketing relationships (Codeshare, Interline, Online), [4] Carrier preference | MNL | yes | no |
| Present Study | Retail bookings and flight data up to 120 days to departure (n = 1,333,712). | [1] Ticket price, [2] Departure day of week, [3] Departure Time, [4] Days to departure at booking, [5] Booking day of week | our model | yes | yes |

## I.9 Price Model

The initial step in our estimation is the construction of the residual $\hat{\xi}$ to accommodate the missing exogeneity of price. Table 4 provides estimates of the linear coefficients $\theta_0, \dots \theta_7$ of

| Component | Estimate ($\hat{\theta}_j$) | Std. Error | % Change |
|---|---|---|---|
| Intercept | 4.6204 | 0.0014 | – |
| log(IV) | 0.1052 | 0.0003 | 11.09% |
| BDAY = Mon | -0.0148 | 0.0017 | -1.47% |
| BDAY = Tue | -0.0170 | 0.0017 | -1.69% |
| BDAY = Wed | -0.0141 | 0.0017 | -1.40% |
| BDAY = Thr | -0.0127 | 0.0017 | -1.26% |
| BDAY = Fri | -0.0054 | 0.0017 | -0.54% |
| BDAY = Sat | 0.0045 | 0.0019 | 0.45% |

Table 4: Linear parameter estimates for the model for PRICE at Eqn. (6) fitted to bookings departing on Thursday. The point estimate and the standard error are reported, along with the effect on PRICE of increasing each covariate by 1 unit, which is given by $\exp(\hat{\theta}_j) - 1$.

model 6 and their impact on PRICE. The baseline for the BDAY dummy variable is Sunday, and the remaining weekday dummy variables have significant relationships with PRICE. There is a slight discount for tickets booked on weekdays of between 0.63% and 1.84%, compared to those booked on the weekends. The instrumental variable IV has a significant positive coefficient, with a z-value of $0.1052/0.0003 = 350.7$ for the null hypothesis that $\theta_1 = 0$; suggesting that the logarithm of bid-price is a strong instrument.

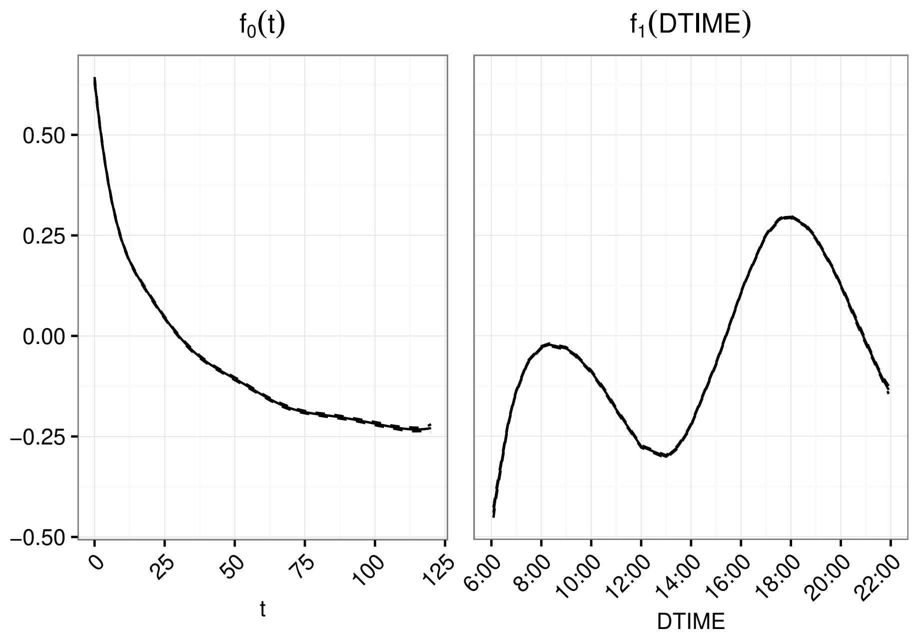



Figure 5: Estimates of $f_0$ (left panel) and $f_1$ (right panel) from fitting the price model at Eqn. (6) to bookings on flights that depart on Thursday. The dashed lines are 99% confidence bands, which are tight.

Figure 5 plots the estimates of the smooth functions $f_0$ and $f_1$, along with 99% confidence bands, constructed as in Marra and Wood (2012). The estimate of $f_0$ shows that ticket prices tend to increase closer to departure. Turning to the estimate of $f_1$, it can be seen that ticket prices tend to peak for flights departing at around 08:00 and 18:00. These are the morning and evening peak demand periods, and this increase is consistent with the demand profile for flights on a busy short-haul route.

## J.10 Demand Model with Two Segments

Table 5 gives the estimates of the linear coefficients, both excluding (Model I) and including (Model II) the residuals $\hat{\xi}$ from the price model. That is Model I ignores the missing exogeneity of price while Model II takes this into account through the above instrumental variable approach. We find a significant coefficients of $\hat{\xi}$, with z-statistics of $-0.0032/0.0004 = -8$ and $-0.0045/0.0010 = -4.5$ clearly highlights the importance of controlling for endogeneity here. Turning to the segments coefficients in Model II, the estimates are $\hat{\alpha}_{1,1} = 0.0008$ and $\hat{\alpha}_{1,2} = -0.0641$, suggesting that the first segment (which we label segment 1) consists of price inelastic customers, whereas the second consists of customers who are more price sensitive. Given the nature of the busy short-haul route, it is likely that segment 1 corresponds to a high proportion of customers travelling for business purposes, whereas the second segment includes a higher proportion of leisure travellers who are more budget conscious. Comparing the estimates of the coefficients of PRICE for Models I and II shows that controlling for endogeneity excentuates the price elasticity for

segment 2.
Recall that the reference category in the log-odds at Eqn. (3.4) is segment $K = 2$. Therefore, the estimates of $\beta^{(\pi)}_{2,1}, \ldots, \beta^{(\pi)}_{7,1}$ indicate the relative preference of customers in segment 1 for booking on different day types. The positive (and significant) coefficient values for the weekdays indicate that customers from segment 1 are more likely to make a booking on weekdays, rather than on Saturday or Sunday (where the latter is the baseline case for the BDAY dummies). This is consistent with the interpretation of customers in segment 1 booking flights for business purposes.
Figure 6 plots the estimates of the smooth components $s^{(\pi)}_{0,1}(t), s^{(\pi)}_{1,1}(\text{DTIME})$ for the log-odds equation as well as the estimates of the smooth components $s^{(\lambda)}_{0}(t), s^{(\lambda)}_{1}(\text{DTIME})$ of the baseline booking along with 99% confidence intervals for Model I. As the right-hand panels show, the probability of a booking increases with time getting closer to the departure date. Other than that there is only a little variation in time and DTIME which, as we will see, is also due to the fact that the model with $K = 2$ customer segments is too simplistic and does not appropriately describe customers' behavior. Comparing the results of Model I with the estimates of Model II, Figure 6 shows that the smooth components

| | Segment | BL | $k=1$ | $k=2$ |
|---|---|---|---|---|
| Model | Component | $\lambda_{\text{BL}}(t)$ | $\pi_k(t)$ | |
| | | | *Segment Adjustment Coefficients* | |
| I | PRICE | | **-0.0009** (0.0002) | **-0.0336** (0.0019) |
| | | *Baseline Coefficients* | *Log-odds Coefficients* | |
| I | Intercept | -1.1213 (0.0886) | -3.8781 (0.1450) | – |
| I | BDAY = Mon | 0.7753 (0.0743) | 3.9991 (0.1437) | – |
| I | BDAY = Tue | 0.9768 (0.0760) | 3.9786 (0.1540) | – |
| I | BDAY = Wed | 1.2515 (0.0806) | 3.4608 (0.1395) | – |
| I | BDAY = Thr | 1.0012 (0.0877) | 3.9168 (0.1482) | – |
| I | BDAY = Fri | 0.6714 (0.0848) | 4.3488 (0.1605) | – |
| I | BDAY = Sat | -0.3964 (0.0558) | 0.3479 (0.1175) | – |
| | | | *Segment Adjustment Coefficients* | |
| II | PRICE | | **0.0008** (0.0005) | **-0.0641** (0.0104) |
| II | $\widehat{\xi}$ | | -0.0032 (0.0004) | 0.0045 (0.0039) |
| | | *Baseline Coefficients* | *Log-odds Coefficients* | |
| II | Intercept | -0.9589 (0.5145) | -1.5856 (1.1263) | – |
| II | BDAY = Mon | 1.1680 (0.3021) | 2.5242 (0.4329) | – |
| II | BDAY = Tue | 1.1535 (0.3879) | 2.6881 (0.6621) | – |
| II | BDAY = Wed | 1.0423 (0.3924) | 3.2702 (0.5823) | – |
| II | BDAY = Thr | 1.0721 (0.3397) | 2.5534 (0.4931) | – |
| II | BDAY = Fri | 0.9795 (0.3058) | 2.5480 (0.4754) | – |
| II | BDAY = Sat | -0.4299 (0.0665) | -0.1150 (0.1905) | – |

Table 5: Parameter estimates and bootstrapped standard errors in parenthesize for $K = 2$ segments fitted to bookings on flights departing on Thursday. Results are given for models fit excluding (Model I) and including (Model II) the price model residuals $\hat{\xi}$.

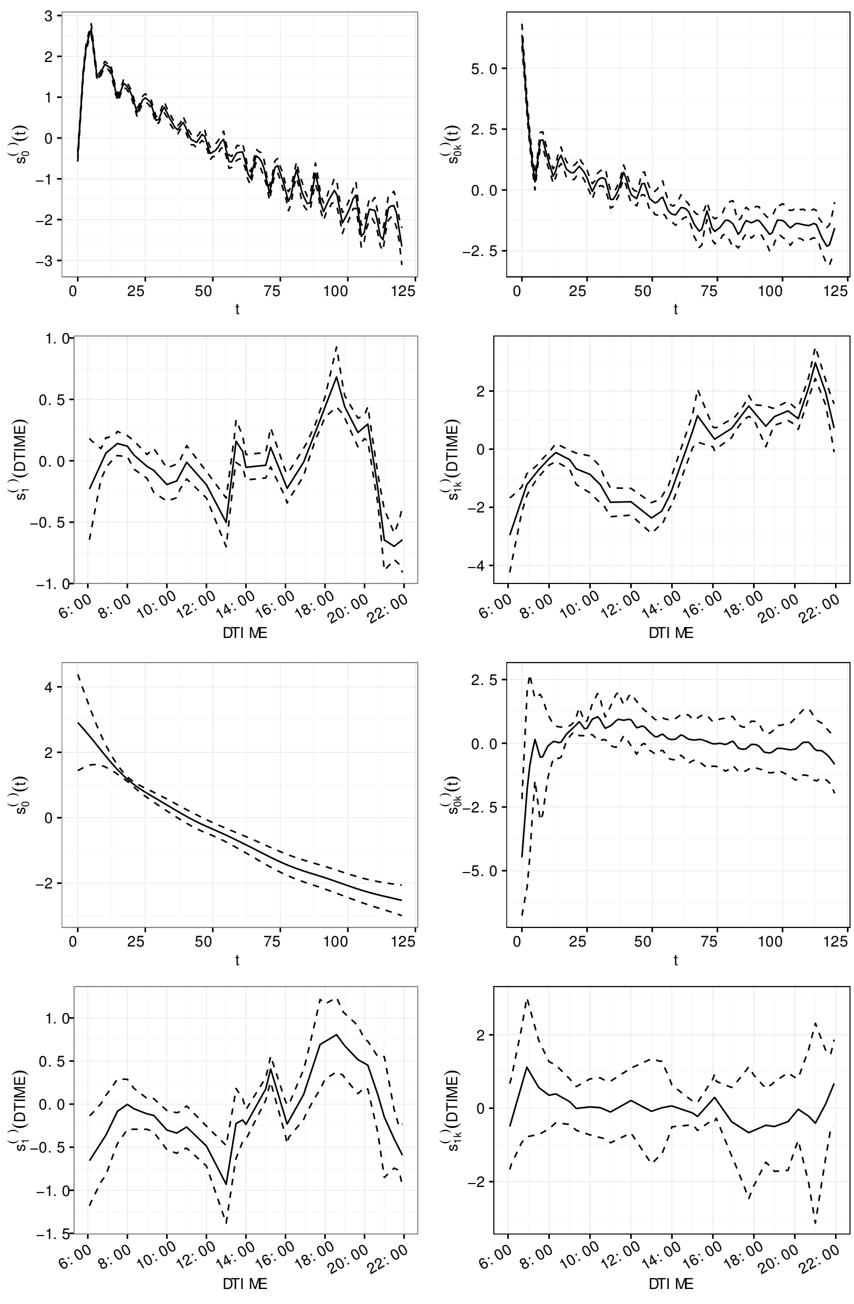



Figure 6: For $K = 2$ segments, function estimates are given for models fit excluding (Model I, first two rows) and including (Model II, row 3 and 4) the price model residuals $\hat{\xi}$. The left-hand panels provide the function estimates for $s_0^{(\lambda)}(t)$ and $s_1^{(\lambda)}(\text{DTIME})$ in Eqn. (3.2) for bookings on flights that depart on Thursday. The right-hand side shows the estimates of $s_{0,1}^{(\pi)}(t)$ and $s_{1,1}^{(\pi)}(\text{DTIME})$ in Eqn. (3.4). The estimates are given by the solid line, while the dashed lines are 99% local confidence bands.

are less erratic if controlling for price endogeneity. This is explained by the reduction in the models' degrees of freedom (134.42 for Model I and 94.73 for Model II). Though both models show similar results long before departure, i.e., overall booking intensity is relatively low and the mix of segments show increasing booking probability of the price-insensitive segment if going closer to departure, the most striking differences show during the week before departure. Here, $s_0^{(\lambda)}(t)$ no longer decreases whereas $s_{0,1}^{(\pi)}(t)$ indicates a decreasing booking probability of the price-insensitive segment. Additionally $s_{1,1}^{(\pi)}(\text{DTIME})$ is no longer significant proposing no segment specific booking probabilities with respect to departure time. As the interpretations of the smooth components from Model I and Model II make equal sense, i.e., Model I suggest that there is a general decline in booking intensity the week prior to departure where only the price-insensitive segment books whereas Model II depicts a steadily increase in booking intensity and a price-sensitive segment close to departure (last minute passengers only willing to travel if the price is cheap) this indecisiveness points towards the possibility of having at least an additional segment of price-sensitive passengers which Model I is not able to describe.

Figure 7 plots $\bar{q}_1(t)$ and $\bar{q}_2(t)$, see (3.9) for Model II. The upper row shows that the proportion of customers in segment 1 — customers with demand patterns consistent with business travel — increases as the departure day gets closer. A strong weekly pattern due to the booking day type is also apparent. The bottom row of Figure 7 $\bar{q}_1(\text{DTIME})$ and $\bar{q}_2(\text{DTIME})$ for Model II. We see that the proportion of customers in segment 1 increases during the peak periods during the morning and evening, which is also consistent with business travel.

Last, we estimate any over-dispersion in the Poisson model by computing the Pearson residuals

$$\varepsilon_{i,t} = \frac{Y_i(t) - E(Y_i(t))}{(\text{Var}(Y_i(t)))^{1/2}} = \frac{y_{i,t} - \lambda(\boldsymbol{x}_{i,t}, t; \boldsymbol{\theta})}{\lambda(\boldsymbol{x}_{i,t}, t; \boldsymbol{\theta})^{1/2}}.$$

The mean of the squared residuals is 1.77, indicating only moderate over-dispersion to the Poisson model. We also investigated whether the squared residuals are related to the covariates, and also to the intensity, and we find no indication of structured heterogeneity.

## K.11 Model Evaluation for intra-day dependence

So far we have treated bookings as independent, conditional on the covariates. However, dependence may exist between bookings made on the same day for flights departing on a given day, that is unaccounted for by the Poisson regression model. We call this 'intra-day dependence' in bookings, and to account for it we use a multivariate Gaussian copula model (Song, 2000) with the margins given by the Poisson regression models fitted above. Copulas models for discrete-valued responsesn have been used previously in the transportation sciences literature; for examples, see Bhat and Eluru (2009),Eluru et al. (2010), and Smith and Kauermann (2011). However, here our copula model needs to capture dependence between vectors that differ in length and composition for each observation, as

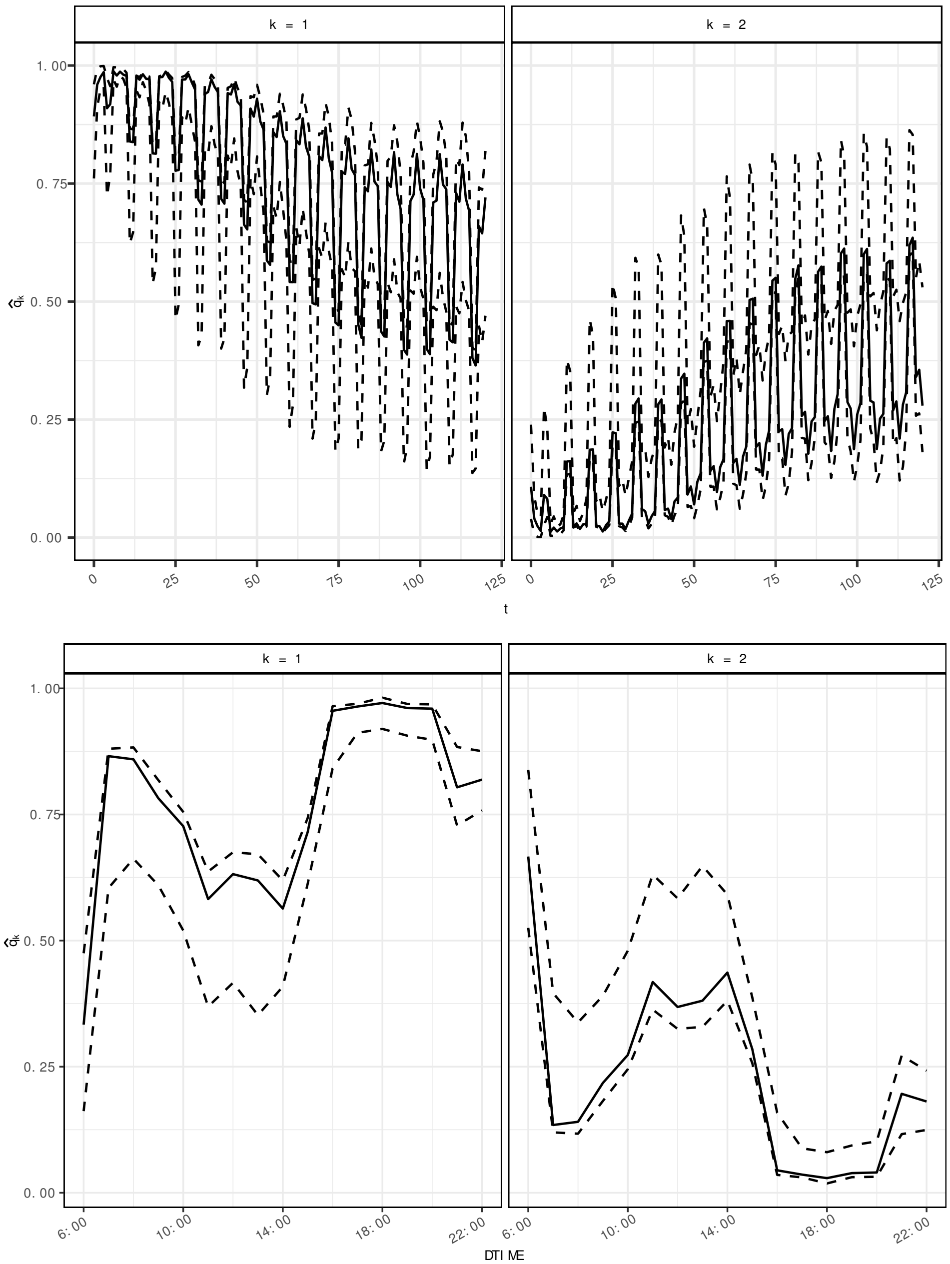



Figure 7: Plot of the average segment proportion computed from the model fitted to booking on flights departing on Thursday and $K = 2$ (solid line) with 99% local bootstrapped confidence bands (dashed lines). Top row: within each panel, $\bar{q}_k(t)$ is plotted against days to departure $t$. Bottom row: within each panel, $\bar{q}_k(\text{DTIME})$ is plotted against DTIME.

we now discuss.

### K.11.1 Gaussian Copula Model

In a copula model, dependence between the elements in a random vector of length $m$ is captured by its 'copula function'. In practice, only vine or elliptical copulas currently are suitable for problems where $m \geq 3$ and pairwise dependence can vary between elements. Vine copulas can be difficult to specify (Dissmann et al., 2013), so that we instead use the Gaussian copula, which is the most popular elliptical copula. It has copula function

$$C_{\text{Ga}}(u_1, \ldots, u_m; \Gamma) = \Phi_m(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m); \Gamma) = \Phi(y_1^\star, \ldots, y_m^\star; \Gamma),$$

where $\Phi_m(\cdot;\Gamma)$ is the distribution function of a $N(\mathbf{0}, \Gamma)$ density with $\Gamma$ as correlation matrix, and $\Phi$ is the standard normal distribution function. The $m \times m$ correlation matrix $\Gamma$ is the copula parameter that requires estimation. We define $y_l^\star = \Phi^{-1}(u_l)$ for $l = 1, \ldots, m$ so that $(y_1^\star, \ldots, y_m^\star) \sim N(0, \Gamma)$. As discussed in Danaher and Smith (2011), because the marginal distributions of the bookings are discrete-valued, we link the continuous Gaussian copula to the observed bookings through the constraint

$$\Phi^{-1}(\text{Po}(y_l - 1; \lambda_l)) < y_l^\star < \Phi^{-1}(\text{Po}(y_l; \lambda_l)),$$

with $\text{Po}(\cdot; \lambda_l)$ as the distribution function from the fitted Poisson regression model from above.

Because on our route flights depart at 61 distinct times, we consider capturing intra-day dependence at the hourly resolution, with flights departing in hourly intervals from 06:00 to 22:00 (except for flights departing between [06:00,08:00) which we consider as a single interval because only a few flights depart prior to 07:00). This requires estimation of a 15-dimensional copula function $C_{\text{Ga}}(u_1, \ldots, u_{15}; \Omega)$, where the parameter matrix $\Omega = \{\omega_{i,j}\}$ for $i = 1, \ldots, 15$ and $j = 1, \ldots, 15$. For example, element $\omega_{2,4}$ captures the dependence between flights departing in intervals [08:00,09:00) and [10:00,11:00). However, there are four complicating factors that make specifying the likelihood of such a copula model difficult for our data. For any given departure day $d$, (i) the number of flights scheduled to depart varies, (ii) the hours at which these flights depart varies, (iii) multiple flights can leave during the same hourly interval (particularly during peak periods), and (iv) a different set of flights can be open or closed for different booking days $t \in \{0, \ldots, 120\}$. Thus, the vector of booking counts for each departure day $d$ and day to departure $t$, given as pair $(d, t)$, can be considered to have 'ragged edges', because it differs both in size and composition of its elements. That is to say, in practice we do not have multiple observations on the 15-dimensional vector of bookings and hence direct application of a copula model to address intraday dependence is not possible.[1]

[1] As an illustrative example, if on day $d_1$ flights were scheduled to depart at 07:15, 07:30, 9:30 and 10:30, then the vector would be of length $K(d_1, t) = 4$. If 30 days prior to departure the 9:30 flight was cancelled, then the vector would of length $K(d_1, t) = 3$ for bookings with $t \leq 30$. And if on the next departure day $d_1 + 1$ there are additional flights also scheduled to depart at 11:00 and 11:30, then $K(d_1 + 1, t) = 6$ with no cancellations.

To account for these complications, we introduce the following notation. For each pair of values $(d,t)$, let $K(d,t)$ denote the number of flights for which booking is possible. Label the hours of departure of these flights as $H(d,t) = \{h_1, \ldots, h_{K(d,t)}\}$, with values from 1 to 15, and where it is possible that $h_i = h_j$ when two or more flights are scheduled to leave during the same hour.[2] Using this notation, all observed booking counts (including occurrences of zero bookings) made $t$ days before departure on flights departing on day $d$, can be stacked into a $K(d,t)$ vector of varying length $\boldsymbol{y}^{d,t} = (y_1^{d,t}, y_2^{d,t}, \ldots, y_{K(d,t)}^{d,t})$.[3] Its joint distribution function is also given by the copula decomposition. It is a property of the Gaussian copula that it is closed under marginalization, including for a subset of elements as here, so that the distribution function of $\boldsymbol{y}^{d,t}$ is

$$F(\boldsymbol{y}^{d,t};\Omega) = \Phi_{K(d,t)}(y_1^{\star,d,t}, y_2^{\star,d,t}, \ldots, y_{K(d,t)}^{\star,d,t}; \Omega^{d,t}) \,.$$

Here, $\Omega^{d,t} = \{\omega_{i,j}^{d,t}\}$ is a $K(d,t) \times K(d,t)$ matrix formed from $\Omega = \{\omega_{i,j}\}$, by setting element $\omega_{i,j}^{d,t} = \omega_{h_i,h_j}$ for $i = 1, \ldots, K(d,t)$ and $j = 1, \ldots, K(d,t)$. That is, $\Omega^{d,t}$ is function of $\Omega$ formed by simply 'pulling out' the relevant elements. The latent variables are distributed $\boldsymbol{y}^{\star,d,t} = (y_1^{\star,d,t}, y_2^{\star,d,t}, \ldots, y_{K(d,t)}^{\star,d,t}) \sim N(\mathbf{0}, \Omega^{d,t})$, constrained by the bounds as at Eqn. $\Phi^{-1}(\text{Po}(y_l - 1; \lambda_l)) < y_l^\star < \Phi^{-1}(\text{Po}(y_l; \lambda_l))$.
The likelihood of the proposed multivariate copula model is the product of the probability mass functions obtained from Eqn. $\Phi_{K(d,t)}(y_1^{\star,d,t}, y_2^{\star,d,t}, \ldots, y_{K(d,t)}^{\star,d,t}; \Omega^{d,t})$ over all pairs of $(d,t)$. However, direct evaluation of each of these individual mass functions is an $O(2^{K(d,t)})$ operation, which is computationally infeasible for the values of $K(d,t)$ in our data, which are typically greater than 15. Instead, we follow Pitt et al. (2006), Danaher and Smith (2011), and Smith and Khaled (2012), and estimate the copula model using Bayesian data augmentation, which generates the constrained latent variables $\boldsymbol{y}^{\star,d,t}$ observing $\Phi^{-1}(\text{Po}(y_l - 1; \lambda_l)) < y_l^\star < \Phi^{-1}(\text{Po}(y_l; \lambda_l))$ using Markov chain Monte Carlo (MCMC) methods. Details are discussed next.

### K.11.2 Copula Estimation

It is computationally infeasible to evaluate the likelihood of high dimensional copula models with discrete margins directly; for example, see the discussion in Smith and Khaled (2012). Therefore, we follow Pitt et al. (2006); Danaher and Smith (2011) and subsequent authors and estimate the copula parameters using Bayesian data augmentation. This provides estimates of the copula parameters—and associated Spearman correlations—from the Bayesian posterior distribution.
Because this is a Bayesian approach, a prior distribution for the copula parameters has to be adopted. For this, we follow Joe (2006); Daniels and Pourahmadi (2009) and parameterize $\Omega$ through its partial correlations. If $1 \leq j < i \leq 15$, these are given by

[2] To continue the illustrative example, $H(d_1,t) = (2,2,4,5)$ for $t > 30$, $H(d_1,t) = (2,2,5)$ for $t \leq 30$ and $H(d_1+1,t) = (2,2,4,5,6,6)$.

[3] To further continue the illustrative example, if there were 3 and 6 bookings on day $t > 30$ for the flights departing at 7:15 and 9:30 on day $d_1$, respectively, then $\boldsymbol{y}^{d_1,t} = (3,0,6,0)$.

$r_{i,j} = \text{Corr}(y_j^\star, y_i^\star | y_{j+1}^\star, \ldots, y_{i-1}^\star)$, where the correlation is defined to be unconditional when $j = i-1$. The set of all partials is therefore $\boldsymbol{r} = \{r_{i,j}; i = 1, \ldots, 15; j < i\}$. This parameterization is invariant with respect to the ordering of the elements of $\boldsymbol{y}^\star$, unlike the Cholesky decomposition of $\Omega$ used in Smith and Kauermann (2011); Danaher and Smith (2011) and others. Daniels and Pourahmadi (2009) give a one-to-one transformation between $\Omega$ and $\boldsymbol{r}$, that is widely attributed to Yule.
The approach generates the latent Gaussian variables $\boldsymbol{y}^\star = \{\boldsymbol{y}^{\star,d,t}; d \in D, t = 0, \ldots, 120\}$ as part of the Markov chain Monte Carlo (MCMC) scheme below. This greatly simplifies estimation, because the posterior of $\boldsymbol{r}$ conditional on $\boldsymbol{y}^\star$ is fast to compute.

**Sampling Scheme**

Step 1. For $d = 1, \ldots, D$, $t = 0, \ldots, 120$, generate from $f(y_i^{\star,d,t} | \{\boldsymbol{y}^\star \backslash y_i^{\star,d,t}\}, \boldsymbol{r}, \boldsymbol{y}) = f(y_i^{\star,d,t} | \{\boldsymbol{y}^{\star,d,t} \backslash y_i^{\star,d,t}\}, \Omega^{d,t}, y^{\star,d,t})$.
Step 2. Generate from $f(\boldsymbol{r}|\boldsymbol{y}^\star)$ element-by-element using (adaptive) random walk Metropolis-Hastings.
Step 3. Compute $\Omega$ from $\boldsymbol{r}$ using Yule's one-to-one transformation.

For Step 1, note that $\boldsymbol{y}^{\star,d,t} \sim N(0, \Omega^{d,t})$, from which the mean $\mu$ and variance $s^2$ of the conditional distribution of the element $y_i^{\star,d,t} | \boldsymbol{y}^{\star,d,t} \backslash y_i^{\star,d,t} \sim N(\mu, s^2)$ can be computed easily. To compute the required conditional posterior, this needs to be combined with the constraint $(L_i^{d,t} < y_i^{\star,d,t} < U_i^{d,t})$, where the lower bound $L_i^{d,t} = \Phi^{-1}\left(\text{Po}(y_i^{d,t} - 1; \lambda_i^{d,t})\right)$ and the upper bound $U_i^{d,t} = \Phi^{-1}\left(\text{Po}(y_i^{d,t}; \lambda_i^{d,t})\right)$. Here, $\Phi$ is the standard normal distribution function, and $\text{Po}(\cdot; \lambda_i^{d,t})$ is the distribution function of the Poisson regression model in Section 4 with intensity value $\lambda_i^{d,t}$ for booking count $y_i^{d,t}$. (Note that we define $\text{Po}(-1; \lambda) = -\infty$ here). The conditional posterior in Step 1 is therefore a $N(\mu, s^2)$ distribution constrained to the range $(L_i^{d,t}, U_i^{d,t}]$. The bounds are computed only once, based on the fitted Poisson regression model, so that it is fast to sample each element. Moreover, the elements can be sampled in parallel because the loops in $d$ and $t$ are not recursive.
To implement the random walk Metropolis-Hastings (MH) in Step 2, $f(\boldsymbol{r}|\boldsymbol{y}^\star) \propto f(\boldsymbol{y}^\star|\boldsymbol{r})f(\boldsymbol{r})$, where the prior $f(\boldsymbol{r})$ is flat on the partial correlations. The augmented likelihood is

$$f(\boldsymbol{y}^\star|\boldsymbol{r}) = \prod_d \prod_t \phi_{K(d,t)}(\boldsymbol{y}^{\star,d,t}; \boldsymbol{0}, \Omega^{d,t}) \,.$$

By first computing $\Omega$ from $\boldsymbol{r}$ using Yule's one-to-one transformation, the matrices $\Omega^{d,t}$ above can be formed by simply extracting their elements from $\Omega$. The density is then evaluated directly, which requires the Cholesky factorization of each matrix $\Omega^{d,t}$. When programmed in a low level language (Fortran 90) we found this is practical to implement on regular PCs with the sample sizes examined here. Moreover, the products can be readily computed in parallel, greatly speeding the evaluation. Note that all computations are undertaken on the logarithmic scale for numerical stability, as is usually the case when implementing a MH step. In general, we run our sampling scheme for a burnin of 40,000 iterates, and a collect a further 20,000 iterates from which to compute posterior inference, which takes around 4 hours on a standard desktop for our dataset.

### K.11.3 Estimated Dependence

As in Section 3, we fit the model separately for different departure day types. We also further segment by the number of days prior to departure when the booking was made, and by the booking day type. To measure the overall level of dependence we compute the posterior estimates of the Spearman pairwise correlations between $y_i^\star$ and $y_j^\star$, which is $\rho^s_{i,j} = \frac{6}{\pi}\arcsin\omega_{i,j}$ for a Gaussian copula parameter matrix $\Omega = \{\omega_{i,j}\}$. Figure 8 plots the posterior mean of the matrix of Spearman pairwise correlations $R = \{\rho^s_{i,j}\}$ for bookings made on weekdays for flights departing Thursdays. The panels give estimates for bookings made between (a) $2 \leq t \leq 30$, (b) $30 < t \leq 60$ and (c) $t > 60$ days prior to departure. Blank cells show where the 99% posterior probability intervals for $\rho^s_{i,j}$ contain 0. For bookings made in the month prior to departure (panel (a)), there is positive dependence throughout. This is likely due to the omission of factors that drive demand for all flights at a daily level. A similar feature can be seen with bookings made between long before departure in panels (b,c), but mostly for flights that depart in the evening. In either case, the level of dependence is only mild, suggesting the proposed Poisson model accounts for the vast majority of dependence between bookings for flights that depart on the same day. While not reported here, very similar results were found for other segmentations of the bookings data.

## L.12 Penalized Maximum Likelihood Estimation

For simplicity of notation we write $\pi_{k,i,t}$ instead of $\pi_k(t)$ and define $\boldsymbol{\theta}_k^{(\pi)} = (\boldsymbol{\beta}_k^{(\pi)}, \boldsymbol{\gamma}_{0,k}^{(\pi)}, \boldsymbol{\gamma}_{1,k}^{(\pi)})$ as corresponding subvector of $\boldsymbol{\theta}$. The corresponding model design matrix for the $i$-th flight at $t$ days to departure is denoted as

$$\boldsymbol{w}_{i,t}^{(\pi)} = \Big(\mathcal{I}(BDAY_i = j), j = 1, ..., 7; \boldsymbol{w}_0^{(\pi)}(t), \boldsymbol{w}_1^{(\pi)}(DTIME_i)\Big)$$

where $\boldsymbol{w}_0^{(\pi)}(t)$ and $\boldsymbol{w}_1^{(\pi)}(DTIME_i)$ are B-spline basis functions in time and departure time, see also Appendix B. Analogously we define

$$\boldsymbol{w}_{i,t}^{(\lambda)} = \Big(\mathcal{I}(BDAY_i = j), j = 1, ..., 7; \boldsymbol{w}_0^{(\lambda)}(t), \boldsymbol{w}_1(\lambda)(DTIME_i)\Big)$$

and $\boldsymbol{\theta}^{(\lambda)} = (\boldsymbol{\beta}^{(\lambda)}, \boldsymbol{\gamma}_0^{(\lambda)}, \boldsymbol{\gamma}_1^{(\lambda)})^T$ to be the design matrix and corresponding parameter vector for modelling $\lambda_{\text{BL}}$. Finally for the group specific part $\delta_k$ we define the matrix as $\boldsymbol{v}_{i,t} = (\text{PRICE}_{i,t})$ or $\boldsymbol{v}_{i,t} = (\text{PRICE}_{i,t}, \widehat{\xi_{i,t}})$ depend on whether we fit the model without or with instrumental variable where $\text{PRICE}_{i,t}$ is the price for flight $i$ at $t$ days to departure and $\hat{\xi}_{i,t}$ the fitted residual of the OLS estimation of Eqn. (6). The matching vector of parameters is $\boldsymbol{\alpha}_k$. Then the first partial derivatives, defining the gradients, are:

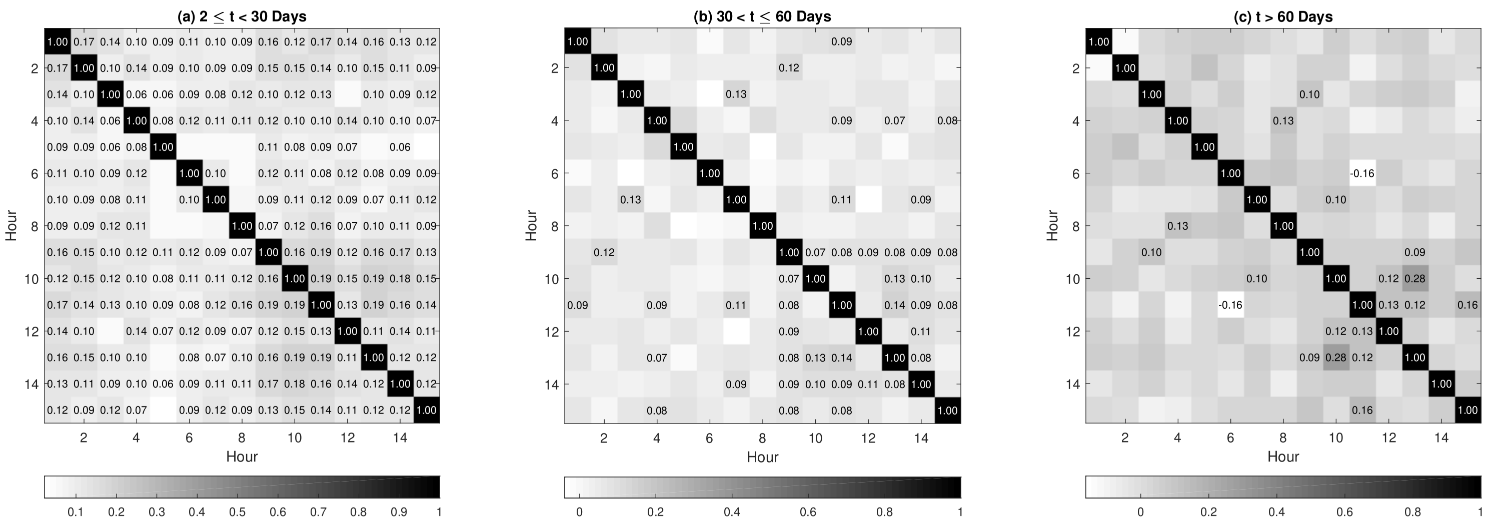



Figure 8: Estimates of pairwise Spearman correlations from the copula model fitted to bookings made on weekdays for flights that depart on Thursdays. Each panel plots the posterior mean of $R = \{\rho^S_{i,j}\}$, arranged as a matrix for hourly indices $i = 1, \ldots, 15$ and $j = 1 \ldots, 15$. Each cell contains the posterior mean of $\rho^S_{i,j}$, except when the 99% posterior probability interval for$\rho^S_{i,j}$ includes 0, in which case the cell is left blank. Results are given for bookings further segmented by (a) 2–30, (b) 31–60, (c) 60-120 days prior to departure.

## L.12.1 Derivatives

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k^{(\pi)}} = \sum_i \sum_t \boldsymbol{w}_{i,t}^{(\pi)^T} \Big(\frac{Y_{i,t}}{\lambda_{i,t}} - 1\Big)\lambda_{0,i,t}\Big(\pi_{k,i,t}(1-\pi_{k,i,t})(\delta_{k,i,t}-\delta_{K,i,t})\Big)$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(\lambda)}} = \sum_i \sum_t \boldsymbol{w}_{i,t}^{(\lambda)^T}(Y_{i,t}-\lambda_{i,t})$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \alpha_k} = \begin{cases} \sum_i \sum_t \boldsymbol{v}_{i,t}^T\Big(\frac{Y_{i,t}}{\lambda_{i,t}}-1\Big) & \text{if } k=1 \\ -c_k\Bigg(\sum_i \sum_t \boldsymbol{v}_{i,t}^T\Big(\frac{Y_{i,t}}{\lambda_{i,t}}-1\Big)\lambda_{0,i,t}\Big(\sum_k \pi_{k,i,t}\lambda_{k,i,t}\Big)\Bigg) & \text{if } k>1 \end{cases}$$

where $\lambda_{0,i,t} = \exp(\boldsymbol{w}_{i,t}^{(\lambda)}\boldsymbol{\theta}^{(\lambda)})$, $\delta_{k,i,t} = \exp(\boldsymbol{v}_{i,t}\boldsymbol{\gamma}_k)$ and $c_k = \exp(\alpha_k)$. The Fisher information results as

$$\mathbb{E}\Bigg(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k_1}^{(\pi)}\partial \boldsymbol{\theta}_{k_2}^{(\pi)^T}}\Bigg) = \begin{cases} \sum_i \sum_t \boldsymbol{w}_{i,t}^{(\pi)^T}\lambda_{0,i,t}^2\Bigg(\frac{\pi_{k,i,t}^2\big(1-\pi_{k,i,t}\big)^2\big(\lambda_{k,i,t}-\lambda_{K,i,t}\big)^2}{\lambda_{i,t}}\Bigg)\boldsymbol{w}_{i,t}^{(\pi)} & \text{if } k=k_1=k_2 \\ \sum_i \sum_t \boldsymbol{w}_{i,t}^{(\pi)^T}\lambda_{0,i,t}^2\Big(\prod_{k\in\{k_1,k_2\}}\pi_{k,i,t}\big(1-\pi_{k,i,t}\big)\Big)B_{k,i,t}\boldsymbol{w}_{i,t}^{(\pi)} & \text{if } k_1 \neq k_2 \end{cases}$$

$$\mathbb{E}\Bigg(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(\lambda)}\partial \boldsymbol{\theta}^{(\lambda)^T}}\Bigg) = \boldsymbol{w}_{i,t}^{(\lambda)^T}\lambda_{,i,t}\boldsymbol{w}_{i,t}^{(\lambda)}$$

$$\mathbb{E}\Bigg(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_{k_1}\partial \boldsymbol{\alpha}_{k_2}{}^T}\Bigg) = \begin{cases} \sum_i \sum_t \boldsymbol{v}_{i,t}^T\lambda_{it}\boldsymbol{v}_{i,t} & \text{if } k=k_1=k_2=1 \\ c_k\Bigg(\sum_i \sum_t \boldsymbol{v}_{i,t}^T\lambda_{0,i,t}^3\frac{A_{k,i,t}^2}{\lambda_{it}}\boldsymbol{v}_{i,t}\Bigg)c_k & \text{if } k=k_1=k_2>1 \\ -c_{k_2}\Big(\sum_i \sum_t \boldsymbol{v}_{i,t}^T\lambda_{0,i,t}A_{k_2,i,t}\boldsymbol{v}_{i,t}\Big) & \text{if } k_1=1, k_2>1 \\ c_{k_1}\Bigg(\sum_i \sum_t \boldsymbol{v}_{i,t}^T\lambda_{0,i,t}^3\frac{A_{k_1,i,t}A_{k_2,i,t}}{\lambda_{i,t}}\boldsymbol{v}_{i,t}\Bigg)c_{k_2} & \text{if } k_1,k_2>1, \quad k_1\neq k_2 \end{cases}$$

where $A_{k,i,t} = \Bigg(\sum_{j=k}^{K}\pi_{j,i,t}\lambda_{j,i,t}\Bigg)$ and $B_{k,i,t} = \frac{\lambda_{k,i,t}-\lambda_{K,i,t}}{\lambda_{i,t}}$

$$\mathbb{E}\Bigg(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k_1}^{(\pi)}\partial \boldsymbol{\alpha}_{k_2}{}^T}\Bigg) = \begin{cases} -\sum_i \sum_t \boldsymbol{w}_{i,t}^{(\pi)^T}\lambda_{0,i,t}\pi_{k_1,i,t}\big(1-\pi_{k_1,i,t}\big)B_{k_1,i,t}\boldsymbol{v}_{i,t} & \text{if } k_1\geq 1, k_2=1 \\ -c_{k_2}\Bigg(\sum_i \sum_t \boldsymbol{w}_{i,t}^{(\pi)^T}\lambda_{0,i,t}^2\pi_{k_1,i,t}\big(1-\pi_{k_1,i,t}\big)A_{k_2,i,t}B_{k_1,i,t}\boldsymbol{v}_{i,t}\Bigg) & \text{if } k_1\geq 1, k_2>1 \end{cases}$$

$$\mathbb{E}\Bigg(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k}^{(\pi)}\partial \boldsymbol{\theta}^{(\lambda)^T}}\Bigg) = \sum_i \sum_t \boldsymbol{w}_{i,t}^{(\pi)^T}\lambda_{\mathrm{BL}}\big(\lambda_{k,i,t}-\lambda_{K,i,t}\big)\pi_{k,i,t}\big(1-\pi_{k,i,t}\big)\boldsymbol{w}_{i,t}^{(\lambda)}$$

$$\mathbb{E}\Bigg(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k}^{(\lambda)}\partial \boldsymbol{\alpha}_{k}^T}\Bigg) = \begin{cases} \sum_i \sum_t \boldsymbol{w}_{i,t}^{(\lambda)^T}\lambda_{i,t}\boldsymbol{v}_{i,t} & \text{if } k=1 \\ -c_k\Bigg(\sum_i \sum_t \boldsymbol{w}_{i,t}^{(\lambda)^T}\lambda_{0,i,t}A_{k,i,t}\boldsymbol{v}_{i,t}\Bigg) & \text{if } k>1 \end{cases}$$

As (3.2) and (3.4) are typically not identifiable if the B-splines basis is used a mean centering constraint, see e.g. Wood (2017), is applied to each smooth component. For instance centering the component $\boldsymbol{w}_0^{(\pi)}(t)$ is achieved by finding the matrix $\boldsymbol{Z}_0$ which solves $\mathbf{1}^T\boldsymbol{w}_0^{(\pi)}(t)\boldsymbol{Z}_0 = \mathbf{0}$ where $\boldsymbol{Z}_0$ has one column less then the original design-matrix $\boldsymbol{w}_0^{(\pi)}(t)$. By the use of the re-parameterized parameter-vector $\boldsymbol{\gamma}_{0,k,c}^{(\pi)} = \boldsymbol{Z}_0\boldsymbol{\gamma}_{0,k}^{(\pi)}$ for estimation, the centering constraint is automatically satisfied.

## L.12.2 Penalization Setting

Based on the ideas of Eilers and Marx (1996) and Ruppert et al. (2003). We impose a penalty on the coefficients relating to the functional effects $s_{0k}^{(\pi)}$(t), $s_{1k}^{(\pi)}$(DTIME),$s_0^{(\lambda)}$(t) and $s_1^{(\lambda)}$(DTIME), respectively. We make use of linear B-splines and penalize neighboring coefficients. To be specific we set $\boldsymbol{w}_0^{(\pi)}$(t) as linear B-spline bases with 12 knots located at equidistantly between $-11$ and 133. We therefore penalize first order differences of the components of $\gamma_{0k}^{(\pi)}$, i.e., $\gamma_{0,k,l}^{(\pi)} - \gamma_{0,k,l-1}^{(\pi)}, l = 10, \ldots, 2$. Analogously we specify the remaining spline base matrices. The penalties can be written as quadratic form leading to the penalized likelihood $\ell_p(,)$

$$\ell_p(\boldsymbol{\theta}, \boldsymbol{\rho}) = \ell(\boldsymbol{\theta}) + \sum_{j=0}^{1}\sum_{k=1}^{K-1} \rho_{jk}^{(\pi)} \boldsymbol{\gamma}_{jk}^{(\pi)^T} D_{jk}^{(\pi)} \boldsymbol{\gamma}_{jk}^{(\pi)} + \sum_{j=0}^{1} \rho_j^{(\lambda)} \boldsymbol{\gamma}_j^{(\lambda)^T} D_j^{(\lambda)} \boldsymbol{\gamma}_j^{(\lambda)}$$

where $\boldsymbol{\rho} = (\rho_0^{(\lambda)}, \rho_1^{(\lambda)}, \rho_{0k}^{(\lambda)}, \rho_{1k}^{(\lambda)}, k = 1, ..., K)$ are the penalty parameters to be specified later. Apparently, if $\boldsymbol{\rho} = 0$ one obtains unpenalized estimation. The smoothing matrices $D$ result from taking differences of neighboring coefficients and exactly follows the convention of Eilers and Marx (1996). This means, for instance, that the difference of spline coefficients is penalized so that neighbouring spline coefficients are forced to be of similar size.
The penalty parameters need to be selected, data driven and we here use the Bayesian Information Criterion (BIC) defined through

$$BIC(\boldsymbol{\rho}) = -2\ell(\hat{\boldsymbol{\theta}}) + \log(n)\mathrm{df}(\boldsymbol{\rho})$$

where $\mathrm{df}(\boldsymbol{\rho})$ is the degree of the model and $n$ is the number of observations ($\approx$ number of flights multiplied by the number of considered days to departure). The degree of the model can be approximated through Fisher matrices as follows. Let $F(\boldsymbol{\theta}, \boldsymbol{\rho})$ denote the penalized Fisher matrix, i.e.

$$F(\boldsymbol{\theta}, \boldsymbol{\rho}) = E\left(-\frac{\partial \ell_p(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right).$$

Then, the degree of the model is approximated through

$$\mathrm{df}(\boldsymbol{\rho}) = \mathrm{trace}\{F^{-1}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho})F(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho} = \mathbf{0})\}$$

where $\hat{\boldsymbol{\theta}}$ is the penalized parameter estimate. For a justification of (A.6) see Ruppert et al. (2003) or Krivobokova and Kauermann (2007b).
We maximize (L.12.2) and apply for simplicity the same degree of smoothing for $t$ and DTIME. That is we set $\rho_{jk}^{(\pi)} = \rho_j^{(\lambda)}$ $\forall$ $j = 0, 1$. If $K = 2$ within the optimization procedure the control parameters $\rho_{01}^{(\pi)} = \rho_0^{(\lambda)}$ and $\rho_{11}^{(\pi)} = \rho_1^{(\lambda)}$ will be fixed to some value. These values are selected based on a grid search by minimizing the BIC.

## M.13 Bootstrapping for Mixture Models

We compute bootstrap confidence intervals using the 'leave out one individual' approach of Rice and Silverman (1991). Re-sampling is undertaken on the flight level to accommodate dependence between bookings on the same flight, and consistent with the likelihood at Eqn. (3.6). For each flight $i$, booking counts and associated covariates are re-sampled (with replacement) for the entire window of booking days between $t_i^{\mathrm{close}}$ and $t_i^{\mathrm{open}}$.
To control for label switching we choose to order the segment specific price coefficients $\alpha_k$ in a monotone sequence. The label switching problem occurs for such random samples from its population whenever at least two group labels $\delta_k, k = 1, \ldots, K$ from Eqn. (3.1) change their positions. If the superscript $l$ within $\delta_k^l$ denotes the group label for a random sample, than there exists at least two group indices $k$ for which $\delta_k^1 \neq \delta_k^2$ if a single label switch between two groups occurs.
To avoid label switching and its negative impact on confidence intervals that result from resampling techniques such as boostraping, we propose a frequentist control approach.
To identify the segment labels $\delta_k = \exp(\alpha_{1,k}\mathrm{PRICE} + \alpha_{2,k}\hat{\xi})$ in 3.1 we impose ordering constraints on the elements of the coefficient vector $\boldsymbol{\alpha}_j^T = (\alpha_{j,1}, \ldots, \alpha_{j,K})$. The identification of the group labels is finally achieved by constraints the elements of $\boldsymbol{\alpha}_j$ such that $\alpha_{j,k} < \alpha_{j,k+1}$ or that $\alpha_{j,k} > \alpha_{j,k+1}$. Allowing for every possible ordering $(<, >)$ between two neighboring coefficients $\alpha_{j,k}, \alpha_{j,k+1}$, a total of $K^J$ possible ordering combinations result. For Eqn. 3.1, we have $J = 2$ and if $K = 2$, the number of possible ordering combinations is 4. This number reduces to $K^{(J-1)}$ if it is acknowledged that the same ordering of $\alpha_{j,k} < \alpha_{j,k+1}$ is achieved by $\alpha_{j,k} > \alpha_{j,k+1} \forall j = 1, \ldots, J, k = 1, \ldots, K$ if the grouping index $k$ no longer runs from the lowest to the highest index but rather from the highest to the lowest, i.e., $\alpha_{j,k} > \alpha_{j,k+1}$ transforms into $\alpha_{j,k+1} > \alpha_{j,k}$. Abbreviating $\alpha_{j,k} < \alpha_{j,k+1}$ by $\mathrm{decr}_j$ and $\alpha_{j,k} > \alpha_{j,k+1}$ by $\mathrm{incr}_j$, for Eqn. (3.1) , with $J = 2$ and $K = 2$, every possible ordering constraints belongs to the set $\{\{\mathrm{decr}_1, \mathrm{decr}_2\}, \{\mathrm{decr}_1, \mathrm{incr}_2\}\}$. For a fixed value of $K$, a separate estimation of (3.6) for each ordering constraint is performed and the model with the lowest BIC values among all candidate models is finally chosen and the corresponding ordering constraint is used to derive bootstrap confidence bands. As the number of possible ordering constraints gets large for small values of $K$, we use the ordering constraints of the

$K = 2$ case for $K > 2$. A consequence of this restriction is that only monotone decreasing or increasing ordering constraints between group parameters are allowed, even though the true ordering between group parameters $\alpha_{j,k}$ is possibly different. Therefore, applying the minimum BIC rule to models with $K > 2$ to select the optimal number of groups only results in an upper bound and is therefore not exact. Exemplary let $K = 3$ be the true group size and $\alpha^*_{j,1} < \alpha^*_{j,k+1} > \alpha^*_{j,3}$ the optimal ordering constraint. As we only allow for $\alpha_{j,1} < \alpha_{j,2} < \alpha_{j,3}$ or $\alpha_{j,1} > \alpha_{j2} > \alpha_{j,3}$, the BIC rule selects a model with $K = 4$ groups as a convex combination of the parameters with ordering constraint $\alpha_{j,1} < \alpha_{j,2} < \alpha_{j,3} < \alpha_{j,4}$ is able to express the relation of $\alpha^*_{j,1} < \alpha^*_{j,2} > \alpha^*_{j,3}$ by setting $\alpha^*_{j,1} = \alpha_{j,1}, \alpha^*_{j,2} = \alpha_{j,3}$, and $\alpha^*_{j,3} = \pi_2(.)\alpha_{j,2} + \pi_4(.)\alpha_{j,4}$.

## N.14 Two-step Estimator by residual inclusion

We consider the two-step estimator of Marra and Radice (2011) to account for price-endogeneity. As the technical discussions of Marra and Radice (2011) concerns the omitted variable bias problem as a characteristic of endogeneity, some minor adjustments are necessary to provide a similar statement for the case of simultaneity.
Given that the number of arriving passengers are specified by $Y(t) = \lambda(t) + u_\lambda$, the systematic component $\lambda(t)$ characterizes through the segment-specific Eqn. $\log(\delta_k)$ (3.1) how demand depends on PRICE as:

$$Y(t) = \lambda_{\text{BL}}(t)\left(\sum_{k=1}^{K} \pi_k(t) \underbrace{\exp\left(\alpha_{1,k}\text{PRICE}\right)}_{\delta_k}\right) + u_\lambda$$

If PRICE is endogenous the assumption of $\mathbb{E}\left(u_\lambda \mid \text{PRICE}, \text{BDAY}, \text{DTIME}, t\right) = 0$ is violated and the estimation of the demand-equation results in biased estimates. For $\eta = \eta\big(\text{IV,BDAY,t,DTIME}\big)$ we plug the expression $\text{PRICE} = \exp\left(\eta + \sigma^2/2\right) + \xi$ into the demand-equation which results in

$$Y(t) = \lambda_{\text{BL}}(t)\left(\sum_{k=1}^{K} \pi_k(t) \exp\left(\alpha_{1,k}\Big\{\exp\left(\eta + \sigma^2/2\right) + \xi\Big\}\right)\right) + u_\lambda$$

As the unobservable error $\xi$ enters the segment-specific equations, biased estimates result if $\mathbb{E}\left(u_\lambda \mid \xi\right) \neq 0$. For the additive separation of $\xi$ into two parts

$$\xi = \xi_1 + \xi_2.$$

we assume that $\mathbb{E}\left(u_\lambda \mid \xi_1\right) = 0$ but $\mathbb{E}\left(u_\lambda \mid \xi_2\right) \neq 0$. If $\xi_2$ would be observable we could include this variable to (6) as an additional regressor. Thus, the new predictor changes to

$$\eta_{new} = \theta_0 + \theta_1 IV + \sum_{j=2}^{7} \mathcal{I}(\text{BDAY} = j)\theta_j + f_0(t) + f_1(\text{DTIME}) + \theta_8\xi_2$$

Therefore, the update on the PRICE-equation is

$$\text{PRICE} = \exp\left(\eta_{new} + \frac{\sigma_{new}^2}{2}\right) + v$$

Taking the Taylor approximation of $\exp\left(\eta_{new} + \frac{\sigma_{new}^2}{2}\right)$ around $\theta_8\xi_2 = 0$ results in

$$\text{PRICE} = \exp(\eta) + \underbrace{\frac{\partial \exp(\eta_{new})}{\partial \theta_8 \xi_2}\theta_8\xi_2}_{:=\zeta} + v$$

Thus estimation of the reduced form of price with Eqn. (6) with instrument IV gives an estimate of $\zeta$ that contains information about the unobservable variable $\xi_2$. Therefore, the inclusion of the estimate $\widehat{\xi}$ within the segment-specific equations controls for the endogeneity of price.

## O.15 Source-Code and Data-Files

The associated R- and Fortran-code for the estimation algorithm used, as well as the data for flights departing on a Thursday, can be downloaded at
https://github.com/JFMeyer2k/SMIJ.

# Bibliography

G. M. Allenby and P. E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1-2):57–78, 1998.

E. Arandia. *Spatial-temporal statistical modeling of treated drinking water usage*. PhD thesis, University of Cincinnati, 6 2013. URL `http://etd.ohiolink.edu/rws_etd/document/get/ucin1377870978/inline`.

A. Avramidis. Learning in revenue management: Exploiting estimation of arrival rate and price response. 2013. URL `http://www.personal.soton.ac.ukaa1w07/v35.pdf`.

K. Balaiyan, R. K. Amit, A. K. Malik, X. Luo, and A. A. Joint forecasting for airline pricing and revenue management. *J. Revenue Pricing Management*, 18(6):465–482, December 2019. doi: 10.1057/s41272-019-00188-. URL `https://ideas.repec.org/a/pal/jorapm/v18y2019i6d10.1057_s41272-019-00188-4.html`.

P. Bartke. *Demand Estimation in Airline Revenue Management*. PhD thesis, FU Berlin, 4 2014. URL `http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000096263?lang=en`.

P. Belobaba, A. Odoni, and C. Barnhart. *The Global Airline Industry*. Aerospace Series. Wiley, 2015. ISBN 9781118881170. URL `https://books.google.de/books?id=HwBhBgAAQBAJ`.

O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.*, 57(6):1407–1420, 2009. doi: 10.1287/opre.1080.0640. URL `http://dx.doi.org/10.1287/opre.1080.0640`.

O. Besbes and A. Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Sci.*, 61(4):723–739, 2015. doi: 10.1287/mnsc.2014.2031. Available at http://dx.doi.org/10.1287/mnsc.2014.2031.

C. R. Bhat and N. Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Res. B*, 43(7):749–765, 2009.

S. Bingemer. Back to the future with iata ndc? critical turning points in the history of airline distribution. *Journal of Tourism Futures*, 4(3):205–217, 2018.

R. Blundell, J. Horowitz, and M. Parey. Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quant. Econom.*, 3(1):29–51, 3 2012. ISSN 1759-7331. doi: 10.3982/QE91. URL `http://dx.doi.org/10.3982/QE91`.

G. Bonciolini. Automatic adjustment of dynamic pricing under turbulent market conditions. https://publications.agifors.org/documents/1-0609 2022. AGIFORS.

E. A. Boyd and I. C. Bilegan. Revenue management and e-commerce. *Management Sci.*, 49 (10):1363–1386, October 2003. doi: 10.1287/mnsc.49.10.1363.1. URL `https://ideas.repec.org/a/inm/ormnsc/v49y2003i10p1363-1386.html`.

A. Brezger and W. Steiner. Monotonic regression based on bayesian p-splines: An application to estimating price response functions from store-level scanner data. *J. Business Economic Sci.*, 26(1):90–104, 2008. ISSN 07350015.

M. Brons, R. Pels, P. Nijkamp, and P. Rietveld. Price elasticities of demand for passenger air travel: a meta-analysis. *J. Air Transp. Management*, 8(3):165–175, 2002. doi: http://dx.doi.org/10.1016/S0969-6997(01)00050-3. Available at http://www.sciencedirect.com/science/article/pii/S0969699701000503.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *SIAM J. Appl. Math.*, 6(1):76–90, 1970. doi: 10.1093/imamat/6.1.76. URL `http://imamat.oxfordjournals.org/content/6/1/76.abstract`.

Q. Chen, J. Stefanus, and I. Duenyas. Adaptive parametric and nonparametric multi-product pricing via self-adjusting controls. *Ross School of Business Paper*, (1258), 2014. URL `http://dx.doi.org/10.2139/ssrn.2533468`.

G. Claeskens and N. Hjort. *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press, 2008.

J. Dai, W. Ding, A. Kleywegt, X. Wang, and Y. Zhang. Choice based revenue management for parallel flights. *Available at SSRN 2404193*, 2014. Available at http://ssrn.com/abstract=2404193.

P. Danaher and M. Smith. Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Sci.*, 30(1):4–21, 2011. Available at http://EconPapers.repec.org/RePEc:inm:ormksc:v:30:y:2011:i:1:p:4-21.

M. J. Daniels and M. Pourahmadi. Modeling Covariance Matrices via Partial Autocorrelations. *Journal of Multivariate Analysis*, 100(10):2352–2363, 2009.

R. Davidson and J. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1st edition, 1993.

R. Davidson and J. MacKinnon. *Econometric Theory and Methods*. Oxford University Press, 1st edition, 1999.

C. de Boor. *A Practical Guide to Splines*. Springer, 1978.

A. den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in oper. res. and management sci.*, 20(1):1–18, 2015.

E. Diaconescu. The use of NARX neural networks to predict chaotic time series. *Wseas Transactions on computer research*, 3(3):182–191, 2008.

J. Dissmann, E. C. Brechmann, C. Czado, and D. Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69, 2013.

G. Doreswamy, A. S. Kothari, and S. Tirumalachetty. Simulating the flavors of revenue management for airlines. *J. Revenue Pricing Management*, 14:421–432, 2015.

P. Eilers and B. Marx. Flexible smoothing with B-splines and penalties. *Stat. Sci.*, 11: 89–121, 1996.

N. Eluru, R. Paleti, R. Pendyala, and C. Bhat. Modeling injury severity of multiple occupants of vehicles: Copula-based multivariate approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2165(1):1–11, 2010.

D. Escobari. Airport, airline and departure time choice and substitution patterns: An empirical analysis. *Transportation Res. A*, 103:198–210, 09 2017.

V. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Management Sci.*, 59(2):305–322, 2 2013. doi: 10.1287/mnsc.1120.1610. URL `http://dx.doi.org/10.1287/mnsc.1120.1610`.

J. Feldman and H. Topaloglu. Bounding optimal expected revenues for assortment optimization under mixtures of multinomial logits. *Prod. Oper. Management*, 24(10):1598–1620, 2015. Available at http://EconPapers.repec.org/RePEc:bla:popmgt:v:24:y:2015:i:10:p:1598-1620.

T. Fiig, K. Isler, C. Hopperstad, and P. Belobaba. Optimization of mixed fare structures: Theory and applications. *J. Revenue Pricing Management*, 9:152–170, 6 2010. Available at https://doi.org/10.1057/rpm.2009.18.

T. Fiig, R. Hardling, S. Pölt, and C. Hopperstad. Demand forecasting and measuring forecast accuracy. *J. Revenue Pricing Management*, 13(6):413–439, 12 2014. Available at http://dx.doi.org/10.1057/rpm.2014.29.

T. Fiig, U. Cholak, M. Gauchet, and B. Cany. What is the role of distribution in revenue management? - past and future. *J. Revenue Pricing Management*, 14(2):127–133, 2015.

G. Gatti Pinheiro, T. Fiig, M. D. Wittman, M. Defoin-Platel, and R. D. Jadanza. Demand change detection in airline revenue management. *Journal of Revenue and*

*Pricing Management*, 21(6):581–595, 2022. doi: 10.1057/s41272-022-00385-8. URL https://doi.org/10.1057/s41272-022-00385-8.

N. Granados, R. Kauffman, H. Lai, and H.-C. Lin. À la carte pricing and price elasticity of demand in air travel. *Decis. Support Syst.*, 53(2):381–394, 5 2012.

C. Gu. *Smoothing Spline ANOVA Models*. IMA Volumes in Mathematics and Its Applications. Springer, 2002.

C. A. Guevara. Overidentification tests for the exogeneity of instruments in discrete choice models. *Transportation Res. B*, 114(C):241–253, 2018.

T. Hastie and R. Tibshirani. Generalized additive models: Some applications. *J. of the American Statistical Association*, 82(398):371–386, 1987.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990.

P. Hetrakul and C. Cirillo. A latent class choice based model system for railway optimal pricing and seat allocation. *Transportation Res. E*, 61(C):68–83, 2014.

IATA. IATA Dynamic Offer Creation, a.

IATA. IATA Industry Facts & Statistics, b.

IATA. IATA Industry Facts & Statistics, c.

H. Joe. Generating Random Correlation Matrices based on Partial Correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189, 2006.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

J. M. Jung and E. T. Fujii. The prize elasticity of demand for air travel. *J. Transp., Econ., and Policy*, 10:257–262, 1979.

D. Karlis and E. Xekalaki. Mixed poisson distributions. *Int. Stat. Rev.*, 73(1):35–58, 04 2005.

G. Kauermann, T. Krivobokova, and T. Fahrmeir. Some asymptotic results on generalized penalized spline smoothing. *J. of the Royal Statistical Society, Series B*, 71(2):487–503, 2009.

J. Ke, H. Zheng, H. Yang, and X. M. Chen. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Res. C*, 85:591–608, 2017.

N. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.*, 62(5):1142–1167, 10 2014. doi: 10.1287/opre.2014.1294. URL http://dx.doi.org/10.1287/opre.2014.1294.

H. Kremers, P. Nijkamp, and P. Rietveld. A meta-analysis of price elasticities of transport demand in a general equilibrium framework. *Economic Modelling*, 19:463–485, 2002.

T. Krivobokova and G. Kauermann. A note on penalized spline smoothing with correlated errors. *J. of the American Statistical Association*, 102:1328–1337, 2007a.

T. Krivobokova and G. Kauermann. A note on penalized spline smoothing with correlated errors. *J. of the American Statistical Association*, 102(480):1328–1337, 2007b. doi: 10.1198/016214507000000978.

D.-J. Lee and M. Durbán. P-spline anova-type interaction models for spatio-temporal smoothing. *Stat. Model.*, 11(1):49–69, 2011. doi: 10.1177/1471082X1001100104. URL `http://smj.sagepub.com/content/11/1/49.abstract`.

L. Leemis. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous poisson process. *Management Sci.*, 37(7):886–900, 1991.

J. Li, N. Granados, and S. Netessine. Are consumers strategic? structural estimation from the air-travel industry. *Management Sci.*, 60(9):2114–2137, 2014.

L. Lin, Z. He, and S. Peeta. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Res. C*, 97:258–276, 2018.

W. Lo, Y. Wan, and A. Zhang. Empirical estimation of price and income elasticities of air cargo demand: The case of hong kong. *Transportation Res. A*, 78:309–324, 2015. ISSN 0965-8564. doi: http://dx.doi.org/10.1016/j.tra.2015.05.014. Available at http://www.sciencedirect.com/science/article/pii/S0965856415001421.

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

G. Marra and R. Radice. A flexible instrumental variable approach. *Stat. Model.*, 11(6): 581–603, 2011.

G. Marra and S. N. Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian J. Statist.*, 39(1):53–74, 2012. Available at https://EconPapers.repec.org/RePEc:bla:scjsta:v:39:y:2012:i:1:p:53-74.

D. Marsh and D. Marshall. *Applied Geometry for Computer Graphics*. Springer, London, UK, 1st edition, 1999. ISBN 1852330805.

E. Martinez-Garcia and M. Royo-Vela. Segmentation of low-cost flights users at secondary airports. *J. Air Transp. Management*, 16:234–237, 2010.

B. Marx, M. Durban, and P. Eilers. Twenty years of p-splines. *Stat. Oper. Res. Trans.*, 39 (2):149–186, 2016. URL `http://statweb.lsu.edu/faculty/marx/SORTTwentyYears.pdf`.

J. McGill and G. Van Ryzin. Revenue management: Research overview and prospects. *Transportation Sci.*, 33(2):233–256, 1999.

J. F. Meyer, G. Kauermann, and M. S. Smith. Interpretable modelling of retail demand and price elasticity for passenger flights using booking data. *Statistical Modelling*, page 1471082X221083343, 2023/01/30 2022. doi: 10.1177/1471082X221083343. URL `https://doi.org/10.1177/1471082X221083343`.

S. Mumbower, L. Garrow, and M. Higgins. Estimating flight-level price elasticities using online airline data: A first step toward integrating pricing, demand, and revenue optimization. *Transportation Res. A*, 66:196 – 212, 2014.

J. Newman, L. Garrow, M. Ferguson, T. Jacobs, and H. Purnomo. Estimation of choice-based models using sales data from a single firm. *Manufacturing Service Oper. Management*, 16(2):184–197, 2014.

T. H. Oum, W. G. Waters, and J. S. Yong. Concepts of price elasticities of transport demand and recent empirical estimates: An interpretative survey. *J. Transp., Econ., and Policy*, 26:139–154, 1992.

OWID. Coronavirus source data. https://ourworldindata.org/coronavirus-source-data. Accessed: 2021-04-12s.

Z. Pang, O. Berman, and M. Hu. Up then down: Bid-price trends in revenue management. *Production and Operations Management*, 24:1135–1147, 2015.

O. Peterson and A. Thankom. Spillover of covid-19: Impact on the global economy. *MPRA Paper*, (99317), 2020. URL `https://ideas.repec.org/p/pra/mprapa/99317.html`.

A. Petrin and K. Train. A control function approach to endogeneity in consumer choice models. *J. Marketing Res.*, 47:3–13, 2010.

M. Pitt, D. Chan, and R. Kohn. Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006. Available at http://www.jstor.org/stable/20441306.

S. Pölt. History of revenue management – from leg to o&d. *J. Revenue Pricing Management*, 15, 07 2016. doi: 10.1057/rpm.2016.6.

S. Pölt, J. Rauch, and K. Isler. Disentangling capacity control from price optimization. *J. Revenue Pricing Management*, 17:48–62, 2018.

N. Pya. *Additive models with shape constraints*. PhD thesis, University of Bath, 7 2010. URL `http://opus.bath.ac.uk/27546/`.

N. Pya and S. Wood. Shape constrained additive models. *Stat. Comput.*, 25(3):543–559, 2014. doi: 10.1007/s11222-013-9448-7. URL `http://dx.doi.org/10.1007/s11222-013-9448-7`.

S. Qiang and M. Bayati. Dynamic pricing with demand covariates. 2016. URL http://dx.doi.org/10.2139/ssrn.2765257.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. of the Royal Statistical Society, Series B*, 53(1): 233–243, 1991.

D. Ruppert, M. Wand, and R. Carroll. *Semiparametric Regression.* Cambridge University Press, 2003.

D. Ruppert, M. Wand, and R. Carroll. Semiparametric regression during 2003 - 2007. *Electron. J. Stat.*, 3:1193 – 1256, 2009.

H. B. Sankaranarayanan and J. Lalchandani. Smart omnichannel architecture for air travel applications using big data techniques. In S. Smys, J. I.-Z. Bestak, Robertand Chen, and I. Kotuliak, editors, *International Conference on Computer Networks and Communication Technologies*, pages 661–669, Singapore, 2019. Springer Singapore. ISBN 978-981-10-8681-6.

G. Sfeir, M. Abou-Zeid, F. Rodrigues, F. Camara Pereira, and I. Kaysi. Latent class choice model with a flexible class membership component: A mixture model approach. *Journal of Choice Modelling*, 41:100320, 2021. ISSN 1755-5345. doi: https://doi.org/10.1016/j.jocm.2021.100320. URL https://www.sciencedirect.com/science/article/pii/S1755534521000531.

M. Smith and M. Khaled. Estimation of copula models with discrete margins via bayesian data augmentation. *J. of the American Statistical Association*, 107(497):290–303, 2012.

M. S. Smith and G. Kauermann. Bicycle commuting in melbourne during the 2000s energy crisis: A semiparametric analysis of intraday volumes. *Transportation Res. B*, 45(10): 1846–1862, 2011.

P.-K. Song. Multivariate dispersion models generated from gaussian copula. *Scandinavian J. Statist.*, 27(2):305–320, 2000. Available at http://EconPapers.repec.org/RePEc:bla:scjsta:v:27:y:2000:i:2:p:305-320.

K. Talluri and G. van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Sci.*, 50(1):15–33, 2004.

K. Talluri and G. van Ryzin. *The Theory and Practice of Revenue Management.* International Series in Operations Research & Management Science. Springer US, 2005.

T. Teichert, E. Shehu, and I. von Wartburg. Customer segmentation revisited: The case of the airline industry. *Transportation Res. A*, 42(1):227–242, 2008.

G. Tutz. *Regression for Categorical Data.* Cambridge University Press, 2012.

G. Tutz and F. Leitenstorfer. Generalized monotonic regression based on b-splines with an application to air pollution data. *Biostatistics*, 8(3):654–673, 2007. doi: 10.1093/

biostatistics/kxl036. URL http://biostatistics.oxfordjournals.org/content/8/3/654.short.

G. Van Ryzin and G. Vulcano. A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Sci.*, 61(2):281–300, 2014.

A. Vij and J. L. Walker. Preference endogeneity in discrete choice models. *Transportation Res. B*, 64:90–105, 2014.

B. Vinod. Evolution of yield management in travel. *J. Revenue Pricing Management*, 15, 07 2016. doi: 10.1057/rpm.2016.15.

B. Vinod. An approach to adaptive robust revenue management with continuous demand management in a COVID-19 era. *J. Revenue Pricing Management*, 20(1):10–14, February 2021. doi: 10.1057/s41272-020-00269-. URL https://ideas.repec.org/a/pal/jorapm/v20y2021i1d10.1057_s41272-020-00269-9.html.

G. Vulcano, G. van Ryzin, and W. Chaar. Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing Service Oper. Management*, 12(3):371–392, 7 2010.

G. Vulcano, G. van Ryzin, and R. Ratliff. Estimating primary demand for substitutable products from sales transaction data. *Oper. Res.*, 60(2):313–334, 3 2012.

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

M. Wedel and W. A. Kamakura. *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media, 2012.

C.-H. Wen and P.-H. Chen. Passenger booking timing for low-cost airlines: A continuous logit approach. *J. Air Transp. Management*, 64:91–99, 2017.

C.-H. Wen and S.-C. Lai. Latent class models of international air carrier choice. *Transportation Res. E*, 46(2):211–221, 2010.

T. A. Whittaker and J. E. Miller. Exploring the enumeration accuracy of cross-validation indices in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(3):376–390, 2021.

M. D. Wittman and P. P. Belobaba. Personalization in airline revenue management – heuristics for real-time adjustment of availability and fares. *J. Revenue Pricing Management*, 16(4):376–396, Aug 2017. ISSN 1477-657X. doi: 10.1057/s41272-016-0002-z. URL https://doi.org/10.1057/s41272-016-0002-z.

M. D. Wittman and P. P. Belobaba. Dynamic pricing mechanisms for the airline industry: a definitional framework. *Journal of Revenue and Pricing Management*, 18(2):100–106, Apr 2019. ISSN 1477-657X. doi: 10.1057/s41272-018-00162-6. URL https://doi.org/10.1057/s41272-018-00162-6.

S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall, CRC texts in statistical science. CRC Press, 2017.

J. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western, 2nd edition, 2002.

S. Wu and A. Akbarov. Forecasting warranty claims for recently launched products. *Rel. Eng. Sys. Safety*, 106:160–164, 2012.

X. Xie, R. Verma, and C. Anderson. Demand growth in services: A discrete choice analysis of customer preferences and online selling. *Decision Sci.*, 47(3):473–491, 2016. ISSN 1540-5915. doi: 10.1111/deci.12177. URL http://dx.doi.org/10.1111/deci.12177.

C. Xu, J. Ji, and P. Liu. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Res. C*, 95:47–60, 2018.

I. Yeoman. Q. Can we manage demand in COVID-19 world? A. I don't know. *J. Revenue Pricing Management*, 20(1):1–2, February 2021. doi: 10.1057/s41272-021-00280-. URL https://ideas.repec.org/a/pal/jorapm/v20y2021i1d10.1057_s41272-021-00280-8.html.

I. Yeoman. Ukraine, price and inflation. *Journal of Revenue and Pricing Management*, 21(3):253–254, 2022. doi: 10.1057/s41272-022-00378-7. URL https://doi.org/10.1057/s41272-022-00378-7.

I. Yeoman and U. McMahon-Beattie, editors. *Revenue Management A Practical Pricing Perspective*. Palgrave Macmillan, 2010. ISBN 978-0-230-24141-1.

A. Zeileis, T. Hothorn, and K. Hornik. Model-based recursive partitioning. *J. of Computational and Graphical Statistics*, 17(2):492–514, 2008.

T. Zhang and S. C. Kou. Nonparametric inference of doubly stochastic poisson process data via the kernel method. *Ann. Appl. Stat.*, 4(4):1913–1941, 12 2010. doi: 10.1214/10-AOAS352. Available at http://dx.doi.org/10.1214/10-AOAS352.

# Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, §8 Abs. 2 Pkt. 5.)
Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

| Jan Felix Meyer | 23/12/2023 |
| --- | --- |
| Jan Felix Meyer | Winterthur |