Shuai Shao

# Applications of Statistical Methods in Airline Ancillary Pricing and Revenue Management

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

Eingereicht am 7. März 2019

Shuai Shao

# Applications of Statistical Methods in Airline Ancillary Pricing and Revenue Management

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

Eingereicht am 7. März 2019

Erster Berichterstatter: Prof. Dr. Göran Kauermann
Zweiter Berichterstatter: Prof. Dr. Helmut Küchenhoff
Dritter Berichterstatter: Prof. Dr. Ralf Wagner

Tag der Disputation: 23. Mai 2019

# Kurzfassung

Fluggesellschaften befinden sich in einem kompetitiven Markt, stark geprägt von andauernden Preiskämpfen. Gleichzeitig steigt der Kostendruck durch Abgaben wie die Luftverkehrssteuer und den Emissionshandel. Dadurch bedingt verringert sich die Marge, die Fluggesellschaften durch den klassischen Transport von Passagieren erwirtschaften können. Dieser Entwicklung versuchen die Fluggesellschaften Rechnung zu tragen, indem sie durch das Angebot zusätzlicher Leistungen (zum Beispiel eines vorzeitigen Sitzplatzreservierung) ihre Marge erhöhen. Dabei muss man die Kunden besser verstehen, um nachgefragte Produkte zu akzeptablen Preisen anbieten zu können.

Das Revenue Management ist eine relative junge wissenschaftliche Disziplin, die sich mit der erlösorientierten Gestaltung von Absatzprozessen beschäftigt. Das Revenue Management hat seinen Ursprung in der Luftfahrtindustrie und findet in den letzten Jahren zunehmenden Anklang in weiteren Branchen, wie zum Beispiel der Transport- und Touristikbranche. Eine der essentiellen Anwendungsvoraussetzungen für das Revenue Management ist ein heterogenes Nachfrageverhalten, welches sich abhängig von Zahlungsbereitschaft beziehungsweise Preiselastizität quantifizieren lässt. Für die Einbettung der zusätzlichen Leistungen in das bestehende Revenue Management, welches sich bisher nur der Kernleistung (Transport) widmet, bedarf es daher einer adäquaten Modellierung des Nachfrageverhaltens für Zusatzleistungen.

Im Rahmen der vorliegenden Dissertation werden statistische Verfahren vorgestellt und angewandt, die kundensegmentspezifische Präferenzen und Zahlungsbereitschaften für bestimmte Zusatzleistungen datengetrieben identifizieren. Die kumulative Arbeit beinhaltet zwei anwendungsorientierte Beiträge. Der erste Artikel beschäftigt sich mit einzelnen Zusatzleistungen, während der zweite Artikel mehrere gebündelte Zusatzleistungen betrachtet. Basierend auf den unterschiedlichen Datengrundlagen — von der granulären Buchungsebene bis zur aggregierten Marktebene — werden verschiedene statistische Modelle verwendet. Diese liefern insbesondere ökonomische Implikationen, zum Beispiel bezüglich potentieller Umsatzveränderung und die daraus resultierenden strategischen Handlungsempfehlungen für Preisgestaltung und Revenue Management. Daraus gewonnene Erkenntnisse lassen sich darüberhinaus direkt auf andere Bereiche des Transports und der Touristik übertragen.

# Abstract

In a competitive market, airlines are utterly influenced by ongoing price wars. At the same time, the cost pressure increases through duties such as aviation tax and emissions trading. As a result, the margin that airlines can generate through the classic transport of passengers is reduced. Taking this into account, airlines attempt to increase their profits by offering ancillary services; for example, advanced seat reservations. In doing so, one has to better understand the customer, in order to offer demanded products at acceptable prices.

Revenue Management is a relatively young scientific discipline that deals with the revenue-oriented organisation of sales processes. It has its origins in the airline industry and has recently become increasingly popular in other businesses, such as the transport and tourism industry. One of the essential requirements for the application of revenue management is a heterogeneous demand behaviour, which can be quantified depending on willingness to pay or price elasticity. For the embedding of the ancillary services in the existing Revenue Management, which to date has been dedicated only to the core service (transport), an adequate modelling of the demand behaviour for ancillary services is required.

In this thesis, data-driven statistical methods are introduced and applied to identify the preferences and willingness to pay for certain additional services on a customer-specific basis. The cumulative work includes two application-oriented contributions. The first article deals with one single ancillary service, while the second article considers the bundled ancillary services. Based on the different databases — from the granular booking level to the aggregated market level — various statistical models are used. They provide particular economic implications, for example, regarding the potential change in revenue and the resulting strategic recommendations for pricing and Revenue Management. Moreover, some of the findings gained thereof can be directly transferred to other areas of transport and tourism.

# Acknowledgements

This cumulative thesis has been a challenging yet rewarding journey, which would not have been so fruitful and joyful without the contributions of numerous persons. I want to express the depth of my gratitude to:

• Prof. Dr. Göran Kauermann for his excellent mentorship. I am utmost grateful for his guidance and advice for my study and research. He has provided me vital support, in both academical and personal aspects. The dedicated and patient supervision can by no means be taken for granted.

• Prof. Dr. Helmut Küchenhoff for serving as an additional reviewer and for bringing me to the Statistical Consulting Unit (StaBLab) in the very early stage of my studies, which provided me the perfect mix of research and applied statistics.

• Prof. Dr. Ralf Wagner for kindly agreeing to serve as an external reviewer, as well as Prof. Dr. Christian Heumann and Prof. Dr. Volker Schmid for their willingness to complete the examination committee.

• Prof. Michael Stanley Smith for sharing his comprehensive knowledge with me and for his persistent effort in proofreading our manuscript.

• Dr. Christopher Alder for initiating the cooperation between Lufthansa Group and the Department of Statistics, as well as his benevolence and trust in me. The professional and personal support of my research activities from all colleagues in project team PELE and Department of Statistics is also sincerely appreciated.

• Dr. Andreas Bender, Dr. David Rügamer and Almond Stöcker for their ongoing willingness to discuss both statistical and non-statistical issues during the last years. It has been an honour having them as study fellows and colleagues.

Last but not least, I would like to thank my family and friends for their continuous encouragement and unwavering support sustained me throughout. Without them, the completion of this thesis would not have been possible.

# Contents

# Chapter 1

# Introduction

## 1.1 Overview

In the era of big data, the data-driven mindset is reshaping all industries. The aviation industry has been an excellent paragon on collecting and using data to achieve better performance since its birth: for fuel efficiency, for plane maintenance and flight safety, as well as for sales and after-sales. From a statistical perspective, one can find enlightening data generating processes from pre-flight through in-flight to post-flight period. Indeed, it is not an exaggeration to say that airlines are not only pushing planes through the clouds but also terabytes of data.

The research questions in this cumulative thesis were initiated by a cooperation between the Department of Statistics at Ludwigs-Maximilian-Universität and Lufthansa Group. The scope of this work is to demonstrate the value of bespoke statistical methods which can make use for airline *Ancillary Pricing* and *Revenue Management*. As the main contributing parts, two related research articles can be read independently. These publications mainly focus on the applications of *statistical methods* that are tailored to provide additional value on accessing and quantifying the *Willingness To Pay* or rather *Price Elasticity* for *ancillary services*. In order to connect the various components from both statistical and economic discipline mentioned above, this chapter gives an elementary introduction to the different scientific fields: Providing the background of the research and area of the application, Section 1.2 introduces the Revenue Management as a decision support system in the perspective of an airline. Motivating the research goals, Section 1.3 reviews the increasing relevance and status quo of ancillary services. Moreover, Section 1.4 first presents the notation and general setup of basic statistical modelling; advanced models and inference procedures are demonstrated in further steps with more details. Last but not least, aspects of the data management are discussed in Section 1.5.

## 1.2 Airline Revenue Management

As the core of every airline's economic model, ***R**evenue **M**anagement* (RM) has been a success story of transfer theory to practice since the deregulation of the U.S. airline market in 1978 (Morrison and Winston 1986). The earliest RM model was presented by Littlewood (1972). Considered as a seminal work in RM, it introduced the idea of maximising the revenue[1] instead of the number of passengers on a particular flight and established the foundations of many further models. Since then, various other industries adopted RM and contributed to it. The spectrum does not only include traditional tourism industries such as cruise lines (Ladany and Arbel 1991), car rentals (Geraghty and Johnson 1997) and hotels (Choi and Mattila 2004) but also other businesses in the tertiary sector such as entertaining (Huntington 1993), retailing (Vinod 2005) and advertising (Kimms and Müller-Bungart 2007). Due to many successful implementations and economical results of RM, even the manufacturing industries from the secondary sector – which traditionally has its main focus on supply chain management – are recently groping for research and innovation in the direction of RM (Gruß 2008; Ruhnau 2012).

Whereas supply chain management has its focus primarily on the optimisation of internal processes and hence the associated costs, RM spotlights the selling of products or services. *"[S]elling the right seats to the right customer at the right prices and the right time"*, this concise description of RM for the airline case in the annual report of American Airlines (1987) gives us an idea of the exact goals. To achieve these goals, many decisions need to be made. These decisions can be divided into three categories relating respectively to

- **structure**: which seats are right?
- **price**: what is the right price?
- **quantity**: who is the right customer (if and how many seats available for him)?

According to the laconic phrase of the Greek philosopher Heraclitus: "*everything flows*" (Beris and Giacomin 2014), answering these above questions depends on the *right time*. As illustrated in Table 1.1, structural decisions in RM are usually made on the strategic level and modified on the tactical level, i.e., they are aimed at the mid- or rather long-term effect and thus will not change frequently. In contrast, the timescale of the last two types of decisions depends on the context and can even vary across firms within an industry. For instance, most traditional airlines change their pricing structure infrequently and allocate the quantity to sell on an operative level; budget airlines on the other hand, mainly use price as their tactical variable.

[1]Low variable costs and high fixed costs in the airline business make maximising revenue approximately equal to maximising profits.

■ Revenue Management ■ Marketing and Production Management

| | Structure decisions | Price decisions | Quantity decisions |
|---|---|---|---|
| Strategical level | Program:<br>· Product/Service range<br>· ... | Market position:<br>· Luxury, premium or other<br>· ... | Resourcing:<br>· Fleet size and aircraft type<br>· ... |
| Tactical level | Design:<br>· Which selling format and channel<br>· How to bundle<br>· ... | Price differentiation:<br>· Segmentability<br>· How to price over time<br>· ... | Capacity adjustment:<br>· Number and location of hubs<br>· Fleet assignment<br>· ... |
| Operative level | | | Capacity steering:<br>· Whether to accept or reject a request<br>· Overbooking<br>· How to allocate capacity to segments<br>· ... |

Table 1.1: Combination of objects (horizontal) and levels (vertical) of the business decision with airline-specific examples.

The fundamental difference between quantity-based and price-based RM is that the former focuses on controlling capacity and not the price. Here, price changes are the consequences of the changing availability of each booking class, which have differently priced seat capacity. In contrast, price-based RM adjust the prices dynamically to a pre-set capacity to maximise revenue without booking classes. This antithesis comes from the different historical development of business models and their corresponding segmentability of the market. Whereas traditional airlines have a more comprehensive product and service range (hence, also often referred to as ***F****ull-****S****ervice* ***C****arriers*, FSC) and allocate their network's capacity through product and price differentiation, budget airlines focus on the price-sensitive customers (hence, also often referred to as ***L****ow-****C****ost* ***C****arriers*, LCC) and segment these through self selection along the booking horizon (Müller-Bungart 2007); Some LCC also use simple models without product and price differentiation, since their optimal pricing tactic is depending on the competing FSC and they must offer lower prices than their competitor (Marcus and Anderson 2008). More details on quantity- and price-based RM can be found in Talluri and van Ryzin (2004).

The crucial question of both price-based and quantity-based RM is, however, the same. It is the decision to accept or reject booking request under uncertain demand and capacity restriction. To sell or not to sell, that is the question. The RM system as decision support for this question can be seen as a weighing scale which quantifies the ***marginal revenue*** on the demand side and the ***opportunity cost*** on the supply side. If one requesting order will provide more marginal revenue than the opportunity cost of the requested goods, then it should be accepted.

As illustrated in Figure 1.1, these two factors are measured and optimised through two separate apparatuses, namely pricing and capacity steering.

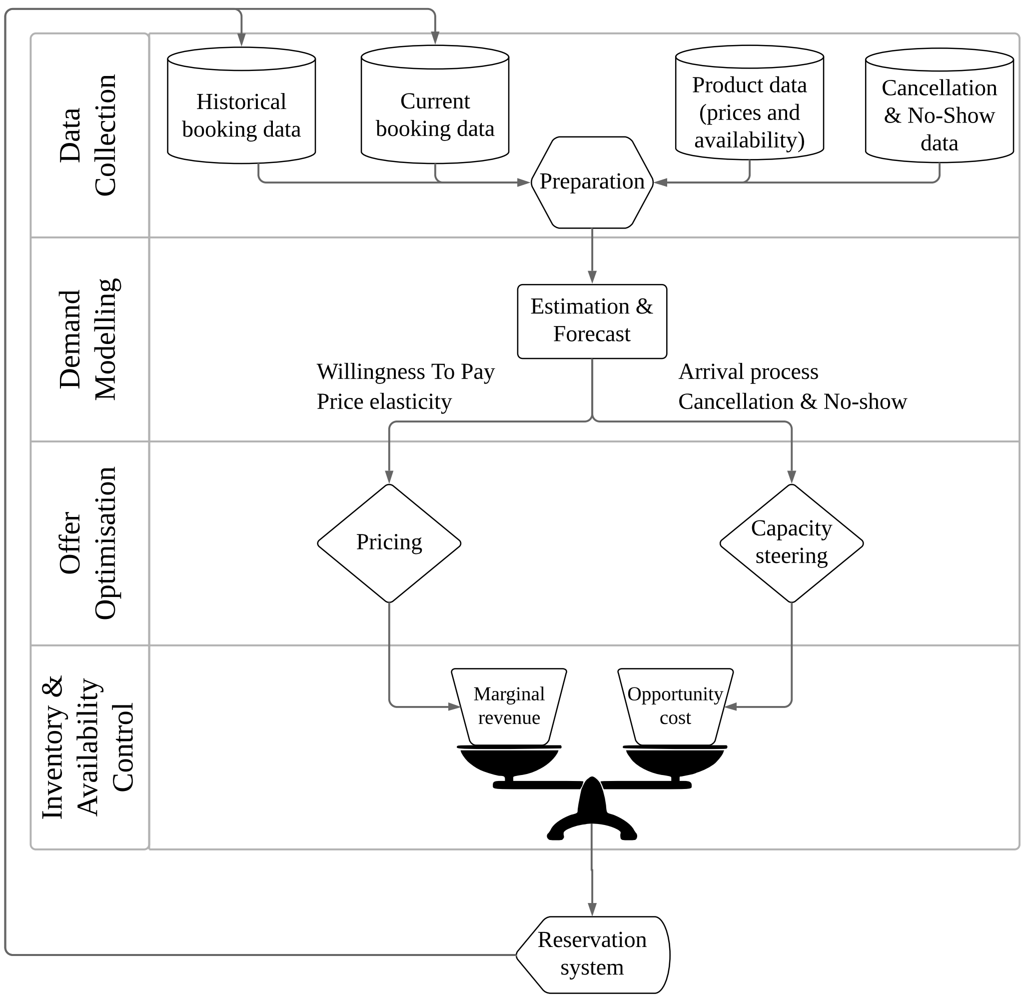


Figure 1.1: An RM System typically follows four steps: 1. Data collection; 2. Demand modelling; 3. Offer optimisation and 4. Inventory & Availability control. Quantity-based RM has more focus on the capacity steering, whereas price-based RM often only uses pricing to steer pre-set capacities.

The RM process typically involves cycling through four steps at repeated intervals. In the first step, relevant data will be prepared as input data for the following step of demand modelling. Diverse parameters of each model will be estimated in the second step and applied in the third step to find the optimal set of controlling rules until the next re-optimisation. In the last step, the availability or inventory of products will be controlled using the optimised controlling rules, which is done either through the airline's transaction-processing systems or shared distribution systems. The frequency with which each step is performed depending on many factors such as the volume of data, the speed that business conditions change, the modelling approach and the optimisation methods used. The methodological task of statistics in RM is hence the adequate modelling of demand in order to provide interpretable pricing or steering parameter. In following, the scientific contributions to RM will be briefly summarised with more focus on the statistical methods.

### Overbooking

Initially, airlines could mainly use overbooking to increase their revenues. Some passengers will cancel their ticket before departure; this is called *cancellation*. Some passengers, on the other hand, do not show up on the day of departure; these are called *no-shows*. Ticket cancellations and customer no-shows cause some of the seats to fly empty on the day of departure, even if the number of seats sold equals the flight capacity. To avoid this *spoilage* of capacity, airlines endeavour to utilise their resources efficiently and hence sell more seats than the flight's capacity. This excess booking above the capacity of the flight is called *overbooking*, and it is the oldest RM practice. The first contribution of statistical methods on overbooking dates back to Beckmann (1958); Rothstein (1985) discussed following efforts on characterising cancellation and no-show distributions. Other early developments are listed in McGill and van Ryzin (1999).

The downside of overbooking is the risk of *oversales*, where more passengers show up at check-in as the available seats. In favourable cases where there is still capacity in the next higher cabin, passengers are happy to be *upgraded*; in critical cases, however, they will suffer from so-called *denied boarding* which leads to negative customers experience and legal penalties for airlines. Thus, the benefits of overbooking have to be considered with costs from oversales. Siddappa et al. (2008) proposed to optimise profit through contrasting the revenue function estimated by *regression splines* and the cost function motivated by a binomial distribution of customers' show up. The binomial distribution imposes the strong assumption, that cancellation probabilities are memoryless and depend only on time to departure and not when the ticket was booked. Iliescu et al. (2008) relax this assumption by using a *discrete time model* to predict cancellation and also suggested further statistical methods such as competing risks.

### Arrival process

Motivated by the same reason as for overbooking, i.e., to avoid the spoilage of the seats, The ***B**ritish* ***O**verseas* ***A**irways* ***C**orporation* (BOAC, now British Airways) started in the early 70s to offer “early bird” bookings with discounted price (hence often referred to as low fare) to passengers who booked at least 21 days before departure. This innovation gave the airline new potential of gaining revenue from seats that would otherwise fly empty. On the other hand, the risk of *spill* full fare (also referred to as high fare) late booking customers occurs. Soon, Littlewood (1972) at BOAC came up with the new twist on the old game of demand and supply. He suggested to only accept bookings with the discounted fare, as long as their revenue value (i.e. marginal revenue) exceeded the expected revenue of future full fare bookings (i.e. opportunity cost). This simple inventory control rule (henceforth, *Littlewood’s rule*) marked the beginning of airline RM. After the Airline Deregulation Act of 1978, airlines linked certain services and restrictions to respectively booking classes, which creates product differentiation. The new main task of RM since this point has been the allocation of inventory/availability and booking classes.

Belobaba (1987) extended Littlewood’s rule to multiple fare classes and introduced ***E**xpected* ***M**arginal* ***S**eat* ***R**evenue* (EMSR) for the general approach. Although this heuristic method does not produce optimal booking limits except in the two-fare case, it became a widely used practice because it is easy to implement. An overview of extensive research since then on the allocation problem in quantity-based RM can be found in Chiang et al. (2007). From the perspective of supply side, development has progressed from optimising *single leg*, to *segment* and finally to *origin-destination* (Poelt 2016); to reflect the characteristics of the supply side, the term “*network RM*” is often used.

On the demand side, customer arrivals in RM are usually modelled using stochastic processes. It is assumed that the state is only influenced by the latest event, and the arrival of demand is regarded as not influenced by inventory/availability controls. Most of the early modelling approaches assume that customers are passive, in other words: the decision-making process of the customer was not considered, and they are merely governed by the demand profile specified at the outset (Shen and Su 2007). Later, a typical convention is to distinguish the *myopic customers* who make a one-time purchase decision at their arrival and the *strategic customers* who may postpone their purchase to a future time point. Different demand streams are often correlated; Stefanescu (2009) considered this pattern and proposed a class of multivariate demand models that capture both the time and the product dimension of demand correlation. Further developments in demand arrival modelling can be found in Cleophas et al. (2009).

### Customer behaviour

Correlation of product demands arises because of heterogeneous customer behaviour. Demand can be very erratic, and customers' preferences may differ, e.g., depending on the purpose of their travel. Hence, it is not a trivial task to match supply and demand. A common simplified distinction for demand in the airline industry is among between time-sensitive (e.g. business travellers) and price-sensitive (e.g. leisure travellers) types of customers. When these groups of customers are offered the same set of products, they will usually make different choices. For instance, if their preferred product is not available or priced over their ***W**illingness* ***T**o* ***P**ay* (WTP), they may show different substitution behaviour, e.g., switch to different products or not purchase at all.

The crucial task for airlines to understand customer behaviour and develop different marketing strategies to accommodate all types of customers is often referred to as "*choice-based RM*", see Vulcano et al. (2010). Conventional modelling approaches in this area are based on the framework of *discrete choice models*, which statistically relate the choice made by each customer to the attributes of the customer and the attributes of the available alternatives to the customer, see, e.g., Ben-Akiva and Lerman (1985) and Train (2009). These works track back to the theoretical basis developed by McFadden (1973, 1981, 1984), who was awarded the Nobel prize for it. The discrete choice models are usually derived under the assumption of utility-maximising behaviour by the decision-maker (Fishburn 1970), which can also be used to motivate *categorical regression models* (Tutz 2012). A detailed review of choice-based methods in RM can be found in Strauss et al. (2018), and technical details of regression models will be given in section 1.4.

The development of choice-based RM systems requires price information viewed by customers at the time of booking. The goal is to forecast demand as a function of price and maximise revenue by jointly determining what prices to offer in which market, as well as how many seats to sell at each price. In turn, airlines need to develop methods that take price fluctuations into account for estimating ***P**rice* ***E**lasticity* (PE), which is the per cent change in demand caused by a per cent change in price. This measurement of how customers respond to changes in price can conveniently find its counterpart in the *ceteris paribus* interpretation of the price variable in the context of statistical regression models in term of probability. In this thesis, PE is accessed and quantified in the regression framework. Further methods for calculating PE[2] can be found in Han and Li (2009).

[2]For interpretation note the distinction between short-term and long-term PE: the immediate response is radical but the permanent change could be less. Based on the aggregation level of data, the smoothed short-term effect and the limited long-term interpretation must be considered, see Simon (1989).

## 1.3 The motivation of research on Ancillary Pricing

Since the world economic crisis in 2008 reduced profits for many companies, they have carried out cost reduction policies which have a substantial impact on business travel; simultaneously, the demand on leisure travel has become more price sensitive, too. For the year 2009, a loss of 70 billion dollars (ca. 15.8%) on passenger revenue was observed by the International Air Transport Association (2018a). As a consequence, the fierce competition between low-cost and traditional carriers has been further intensified, and they are both forced to seek opportunities to generate additional revenues from other sources beyond the airfare: The former started to create more additional supplements to upgrade service coverage; the latter unbundle their products to provide more competitive price. Thus, one can say the most recent industry-wide development on the strategical level has been to derive revenue from ancillary sources. According to the global projection of IdeaWorks (2018), the percentage of ancillary revenue in total revenue has been more than doubled since 2010 (see also Figure 1.2). Indeed, the worldwide airline revenue in 2017 would be less than in 2015, if no ancillary revenue could be additionally generated. These figures show the impact that ancillary revenue has on the bottom lines of airlines. From the perspective of a passenger, ancillary service is also a critically important element of customer experience.

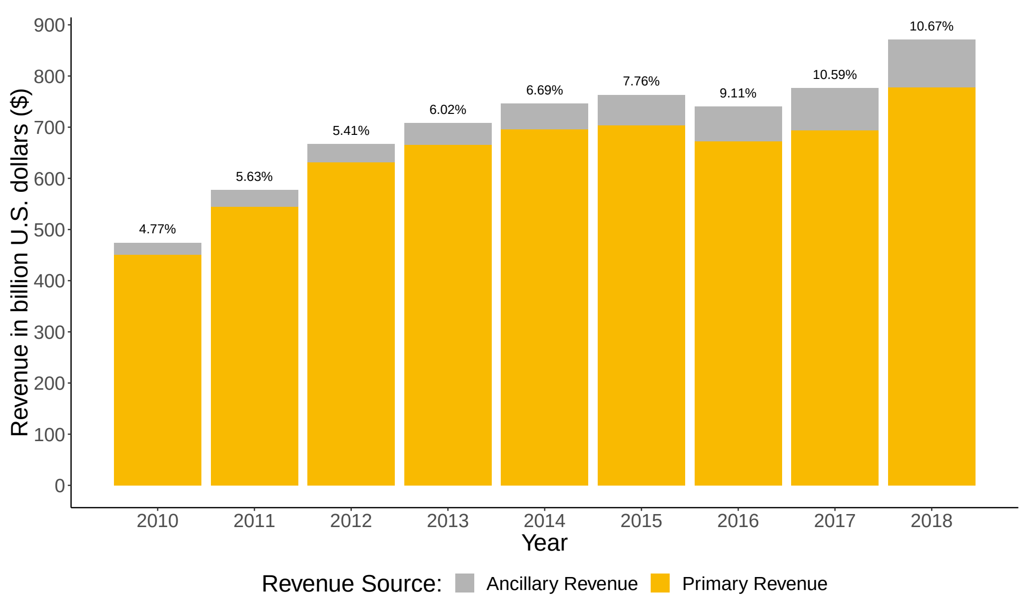


Figure 1.2: Global airline revenue development: The percentage an ancillary revenue in total revenue has been more than doubled since 2019. Data source: IdeaWorks (2018)

Taking the definition of ancillary in Oxford Dictionaries (2018) *"Providing necessary support to the primary activities or operation of an organisation, system, etc."*, a clear definition of primary activity/operation is necessary. As the core of the airline business, the primary service is to transport passengers from origin to destination safely and on time, suiting their schedule. Along the time horizon, Figure 1.3 illustrates various examples of ancillary services enveloping this primary core and these can be divided into four categories, where only the unbundled "a la carte" items are flight-related[3] and can thus be easier incorporated into airlines' existing RM system. Thus, for this thesis, the focus lies on these flight-related ancillaries. Bearing mind that other ancillary sources are also relevant as well to maximising revenue for airlines.

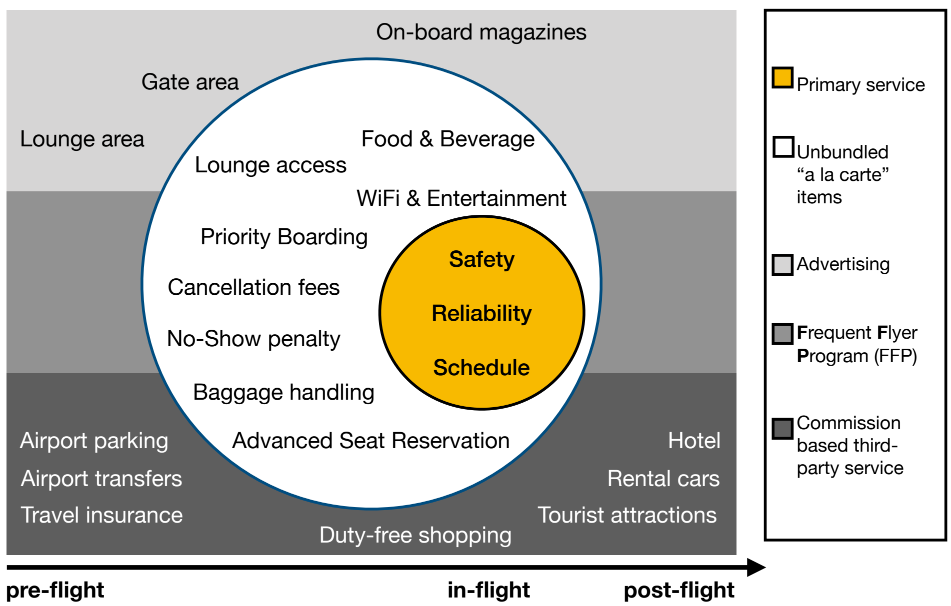


Figure 1.3: Airline primary and ancillary services on the time horizon: While primary service ranges from short pre-flight to in-flight phase, ancillary services accompany customers in a broader stretch.

The recent practice of unbundling has led to a mixed result. On the one hand, the customers can purchase the airfare at a competitive price and choose exactly additional services depending on their preference; this can be seen as a development towards personalisation of the tailor-made travel experience. On the other hand, passengers who were previously accustomed to services that were traditionally included within the fare complain that the airlines take every opportunity to nickel-and-dime their customers.

[3]In particular, only the data of ancillaries in this category are related to a (booking of a) flight, which can be managed within the current RM systems. From a pricing perspective, while advertising and commission based activities are priced on a *business-to-business*(B2B) level, FFP does also price on *business-to-customer*(B2C) level. However, the target of FFP is rather to manage customer relationship than revenue.

Garrow et al. (2012) reviewed the unbundling trends in the U.S. airline industry and anticipated that whereas LCCs will more broadly adopt ancillary fees of unbundled items, many traditional carriers will eliminate their ancillary fees due to negative impacts on customer perception. Wittmer and Rowley (2014) conducted a survey of economy passengers from a European FSC and found that they do perceive value in ancillary services and display a general intention to purchase, especially for services with attribute “hospitality”, e.g., lounge access and seat selection. Furthermore, the viability of (re)bundling the unbundled items was proven by their study, because most of the respondents chose such packages in the simulated purchase situation. Besides the geographical or rather cultural difference of preferences for ancillary service, there is also a different recommendation for (un)bundling of services depending on airlines’ structure. In a simplified combinational binary setting of high (e.g. business travellers) and low type (e.g. leisure travellers) customer, Cui et al. (2018) show that uniform-pricing (respectively, discriminatory-pricing) firm should unbundle the ancillary service if the fraction of high type consumers who value the ancillary service is large (respectively, small) enough. Despite the cultural and structural discussion, this thesis provides statistical modelling approaches for both unbundled “a la carte” item and ancillary bundle focusing on the pricing insights and possible application in RM. A short discussion on the research gaps and the goals of this thesis will be outlined in following subsections.

### 1.3.1 “A la carte” pricing

As previously described, ancillary services as itself were historically less relevant to overall profitability or were not offered at all. Airlines are seeking methods to help them understand how passenger perceive ancillary services. In the economics literature, a la carte item is also called as an *add-on*, regardless if new created or unbundled from the previous full-service. Ellison (2005) showed that add-on pricing could be used as a price discrimination tool for demand segmentation. Price discrimination stems from the fundamentals of WTP. For a specific add-on on a flight, the WTP can differ from customer to customer, just as the WTP for the flight differs between them. To date, the calculation of marginal revenue has been based on the WTP for airfare only. With no doubt, there are inherent differences between airfare and ancillaries. One is a necessity, and the other is an option. As such, customers have different price expectations, sensitivities and motivations and thus different WTP to “must have” versus “nice to have”. As an illustrative example with an opportunity cost at 90€, the booking request of a passenger $A$ who is willing to pay 100€ for the airfare and nothing for ancillary services would be accepted in contrast to a passenger $B$ whose WTP for airfare is 80€ but for ancillaries is 40€ as the add-on.

In order to improve the existing practice and incorporate the ancillary revenue into pricing optimisation, airlines first need a reasonable estimate of the revenue potential of ancillary services. This estimate can then be used to adjust fares to account for both ticket and ancillary revenue when RM systems calculate availability. As will be described in more details in Chapter 2, most of the studies estimated WTP from ***S****tated* ***P****reference* (SP) via a questionnaire survey and not from ***R****evealed* ***P****reference* (RP) via sales data. SP information from surveys has been extensively used in literature to elicit WTP for ancillary services, see, e.g., Balcombe et al. (2009), Correia et al. (2012), and Menezes and Vieira (2008). Despite the progress in survey design and in results interpretation, it is difficult to avoid the intention-behaviour gap, i.e., the difference between "customer attitude" and "customer behaviour", or in other words the difference between what consumers claim they are "ready to pay" and what they "actually pay". Hence, one crucial task of the research article presented in Chapter 2 is to provide pricing insights based on RP through sales data and suggest a general modelling approach for unbundled ancillary items based on statistical methods.

### 1.3.2 Bundle pricing

Product bundling and unbundling by firms has been a focus of researchers in the industrial organisation ever since the seminal contribution of Adams and Yellen (1976). The literature identifies three bundling strategies. Under the pure components (or unbundling) strategy, the seller offers the products separately; under pure bundling, the seller offers the bundle alone; under mixed bundling, the seller offers the bundle as well as each single items. As a strategy, bundling is most suitable for high volume and high margin (i.e. low marginal cost) products, which makes it a perfect match for the business conditions of airline ancillary services. *Branded Fares* were introduced to the airline industry by Air New Zealand in 2004 as a mixed bundling strategy. Since then, numerous airlines have adopted similar schemes due to the double-edge of creating ancillary revenue sources, as discussed at the beginning of this section. Besides offering a baseline brand which only including the primary service transportation, selections of bundled ancillary services are provided as the so-called *up-sell* brands. These up-sell brands (usually 2-3 incremental bundles) add a variety of ancillary services to the baseline brand such as a combination of ***A****dvanced* ***S****eat* ***R****eservation* (ASR), baggage handling, onboard food & beverage or refund and rebook options for a discounted amount. Because of the discounted amount, bundling can be seen as a value pricing strategy. However, other strategic advantages of bundling should not be neglected. In particular, for better brand awareness and passenger segmentation (Fiig et al. 2012; Vinod and Moore 2009), as well as if certain (new) products require more publicity and need to be promoted.

To date, the majority in the airline industry is still using bundling primarily as a strategic tool to promote ancillary items. Hence, they follow a uniform-pricing approach for the up-sell steps, i.e., the discounted amount. This homogeneous pricing strategy may not justify the heterogeneous demand on ancillary items included in the bundle. While differential pricing for up-sell steps can be made based on the length of haul and the corresponding utilities for the customer, Chapter 3 proposes a statistical modelling approach to understand market-specific customer behaviour, which does not need to be solely dependent on the length of haul. Furthermore, the forecast for ancillary revenue can be achieved thereby.

### 1.3.3 The positioning of the contributing articles

The goal in RM is to maximise revenue: given a flight or a network of flights, allocate availability to passengers who are willing to pay the most. The decision, whom to prioritise availability to, depends on the forecast of the expected revenue contribution of different passengers. Traditionally, the expected revenue is based on posted flight ticket fare which mirrors the corresponding WTP. With the growing relevance of ancillary revenue, future RM systems must be capable of considering these ancillary revenues as well. The diverse data source and differences in price sensitivity create a complex task that today's RM solutions have yet to address.

In order to integrate the ancillary revenue into the RM system, airlines need to estimate the ancillary revenue potential of the passenger. Subject to the level of detail in the available data, these estimates can vary in granularity. Figure 1.4 illustrates a coordinate system based on the horizontal dimension of data granularity and vertical dimension of product selling format. The most extreme level on the data dimension is to estimate ancillary revenue potential by each individual. If differential pricing based on individual data is also practised to individuals, it is equivalent to the price differentiation on the first degree in the economic theory (Pigou 1932, Chap. XVII). As it is the case in most businesses[4], first-degree price differentiation is an (uncommon) ideal[5] and is difficult to practise in the airline[6] industry, too. The data collection of individual customers is often compromised as the collection of individual booking data, which can contain more than one individual and hence labelled as semi-individual data in this thesis. On the other hand, the collection and storage of more granular data also require more resources. Thus, the estimation based on aggregated booking class or market provides a more available — especially more economical — alternative level of detail.

[4]Auctions are the most practised differential pricing methods near to first-degree price differentiation.

[5]See, e.g., Executive Office of the President of the United States (2015) for further discussion on this topic.

[6]Although exact individual data are available by some customer (e.g. frequent flyer), they are not the majority affected by RM systems.

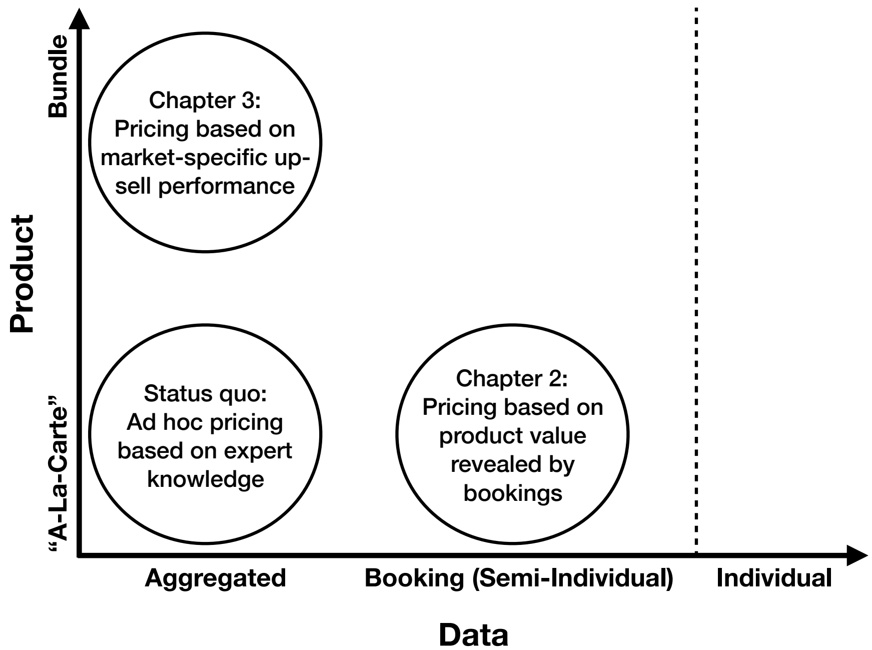


Figure 1.4: Positioning of the contributing articles: Considering the pricing problem of "a la carte" ancillary products, Chapter 2 applies statistical models to booking data. On the level of aggregated data, Chapter 3 provides a model-based market-specific pricing approach for ancillary bundles.

In the current industry practice, prices of single "a la carte" ancillary items are set by pricing experts in an *ad hoc* manner. The models presented in Chapter 2 can be used to discover the additional potential in product pricing on the one hand, and improve the expected revenue contribution through WTP (either fare or availability) adjustment on the other hand. For demonstration, Chapter 2 uses ASR as an example, the possible application is however not restricted to ASR only and can be extended to further flight-related ancillary items.

Given the need to (re)bundle the ancillary services and controversy situation on protecting data privacy, Chapter 3 takes a step back on the data dimension and engages in the area of bundle pricing. The estimated market-specific PE provides decision support for making and evaluating bundle pricing policy.

Economically, both applications in Chapter 2 and 3 can be seen as differential pricing tools on the second or third degree by providing their managerial implications and strategical recommendation for airlines based on statistical models. To better understand the employed models, the elementary foundation of statistical modelling is briefly sketched in the following section.

## 1.4 Statistical methods

As will be shown in Chapter 2 and 3, the probability of a customer purchasing an ancillary item or a bundle of ancillaries can be modelled in the regression context. Regression is the most commonly practised statistical methodology for analysing empirical problems in many scientific disciplines such as life sciences, social sciences and economics (McCullagh and Nelder 1989). Since the first regression analysis by Galton (1886), the methodology has been developed in many ways. In this section, the framework of regression is briefly summarised. The general setup will be first described and then extended to different model classes. Further details as well as practical examples can be found in Fahrmeir et al. (2013).

Following conventions of notation are used: random variables are denoted by upper case italic letters and their observed values by the corresponding lower case italic letters, e.g., the observations $y_1, y_2, ...y_n$ are regarded as realisations of the random variables $Y_1, Y_2, ..., Y_n$. Vectors are written in bold, i.e., $\boldsymbol{y}$ represents a vector of observations

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and $\boldsymbol{Y}$ is a vector of random variables

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

Note that a matrix is also written in bold upper case letters but not italic, e.g., $\mathbf{X}$. The superscript $(\cdot)^T$ is used for a matrix transpose or when a column vector is written as a row, e.g., $\boldsymbol{y} = (y_1, ..., y_n)^T$. Greek letters denote parameters, and the symbol $\hat{}$ is used for estimators, e.g., the parameter vector $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}$.

Both the probability density function of a continuous random variable and the probability mass function for a discrete random variable are referred to their distributions and denoted by $f_\theta(y)$, where $\theta$ represents the parameters of the distribution. The formulation $\hat{\mathbb{P}}(Y|X)$ is used for the predicted outcome probability, in particular with the focus on conditioning on the model input $X$ and not the probability distribution itself nor its (estimated) parameter. The expected value and the variance of a random variable $Y$ are denoted by $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$ respectively.

### 1.4.1 Regression models

In the setting of classical linear regression, it is assumed that the true expectation $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ of the *response* or *dependent variable* $\boldsymbol{Y} = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ with realisations $\boldsymbol{y} = (y_1, ..., y_n)^T$ is linked to a linear combination of $k$ *explanatory* or *independent variables* (or simply *covariates*) and unknown parameter $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$. This is given by

$$\boldsymbol{\mu} = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},\ \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n), \tag{1.1}$$

where the *design matrix* $\mathbf{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_k)$ contains $k$ columns of covariates and $n$ rows of observation. The linear *predictor* $\boldsymbol{\eta}$ results directly from the linear combination $\mathbf{X}\boldsymbol{\beta}$. $\mathbf{I}_n$ is an $n$-dimensional identity matrix and $\sigma^2 > 0$ is the variance of the independent and identically distributed (i.i.d.) errors, stemming from observing an erroneous version

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e},\ \boldsymbol{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n). \tag{1.2}$$

It is assumed that the errors are uncorrelated amongst each other.

The usual estimation procedure for the parameters $\boldsymbol{\beta}$ follows by minimising the squared errors

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}}\ (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{1.3}$$

and $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ hence called the ***O**rdinary **L**east **S**quares* (OLS) estimator.

This setting provides a simple way to describe the relationship between explanatory variables and a response variable. Together with the straightforward computation to obtaining parameter estimates due to a closed-form solution $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$, as well as the intuitive interpretation of the results, make the classical linear regression model an attractive and frequently used approach in the empirical analysis.

As the famous aphorism of the British statistician Box and Draper (1987) states, "*all models are wrong, but some are useful*", and models are more useful if they are tailored to the characteristics of the nature of the investigating object. The adequate formulation and interpretation of regression models require therefore explicit consideration of the different types of the response variables as well as the explanatory variables. Following this principle, questions from various scientific disciplines led to the need to extend the classical linear regression model, mainly for the following two types of more general situations:

Type I: The response variable follows a distribution different from a normal distribution, for instance, the response variable can be binary or categorical.

Type II: The relationship between the response and explanatory variables is not linear.

### Generalised Linear Model

Nelder and Wedderburn (1972) demonstrated the unity of many extended statistical methods which overcome the limitation of OLS for both above-described situations. They created the necessary concept with the introduction of ***G**eneralised **L**inear **M**odels* (GLM). The GLM is characterised by two assumptions:

The **distributional assumption** constitutes that, given covariates $\boldsymbol{x}_i, i = 1, ..., n$, the response $Y_1, ..., Y_n$ are conditionally independent and the conditional distribution of these responses belongs to a distribution family with densities of the form

$$f_{\phi,\theta}(y_i) = \exp\left[\frac{y_i\theta - a(\theta)}{\phi} + c(y_i, \phi)\right], \tag{1.4}$$

which presents the *exponential family* when with a fixed dispersion parameter $\phi$ scaling the variance. Note this scale parameter $\phi$ does not depend on the observation $i = 1, ..., n$.

Table 1.2 lists some well-known distributions in the exponential family. In this context, $\theta$ is the *natural* or *canonical* parameter, and both $a(\cdot)$ and $c(\cdot)$ are functions corresponding to the type of the distribution. The exponential family has many desirable properties that make it useful for statistical analysis. One particular is the direct access to a *sufficient statistics* for the parameter of interest (Barankin and Maitra 1963). For a collection of $n$ i.i.d. random variables sampled from the same exponential family distribution, the joint likelihood[7] of $\boldsymbol{\theta}$ can be obtained by taking the product

$$L_y(\boldsymbol{\theta}) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} f_\theta(y_i) = \left[\prod_{i=1}^{n} c(y_i)\right] \exp\left[\left\langle \boldsymbol{\theta}, \sum_{i=1}^{n} y_i \right\rangle - n \cdot a(\boldsymbol{\theta})\right] \tag{1.5}$$

with sufficient statistic $t(y) = \sum_{i=1}^{n} y_i$ summarising the data. Note that an exponential family distribution can have a multidimensional parameter, i.e., $\boldsymbol{\theta} \in \mathbb{R}^d, d \geq 1$.

[7] Here, the roles of the data $y$ and the parameter $\theta$ are interchanged. The likelihood is considered as a function of $\theta$ for fixed data $y$, in contrast to the distribution as a function of $y$ for fixed $\theta$. However, both are the same function of $y$ and $\theta$ jointly.

| | Normal $\mathcal{N}(\mu, \sigma^2)$ | Bernoulli $\mathcal{B}(1, \pi)$ | Binomial $\mathcal{B}(n, \pi)$ | Multinomial $\mathcal{M}(n, \boldsymbol{\pi})$ |
|---|---|---|---|---|
| $f_{\boldsymbol{\theta}}(y) =$ | $\frac{1}{\sqrt{2\pi\sigma}} \exp\big[-\frac{1}{2\sigma^2}(y-\mu)^2\big]$ | $\pi^y(1-\pi)^{(1-y)}$ | $\binom{n}{ny}\pi^{ny}(1-\pi)^{(n-ny)}$ | $\frac{n!}{y_1!\cdots y_m!}\exp(\sum_{m=1}^M y_m \log \pi_m)$ |
| $y_i \in$ | $\mathbb{R}$ | $\{0,1\}$ | $\{0,1\}$ | $\{1, ..., M\}$ |
| $\mu =$ | $\mu$ | $\pi$ | $\pi$ | $(\pi_1, ..., \pi_M)^T$ |
| $\theta(\mu) =$ | $\mu$ | $\log\big(\frac{\pi}{1-\pi}\big)$ | $\log\big(\frac{\pi}{1-\pi}\big)$ | $\Big[\log\big(\frac{\pi_1}{\pi_M}\big), ..., \log\big(\frac{\pi_{M-1}}{\pi_M}\big), 0\Big]^T$ |
| $a(\theta) =$ | $\frac{\theta^2}{2}$ | $\log(1+e^\theta)$ | $n\log(1+e^\theta)$ | $n\log(\sum_m^M e^{\theta_m})$ |

Table 1.2: Selection of well-known distributions belonging to the exponential family, for extensions see Wood (2006, p. 61), Tutz (2012, p. 61) and Fahrmeir et al. (2013, p. 303).

The **structural assumption** determines the linear predictor $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ as in classical linear regression to the conditional expectation $\mathbb{E}(Y_i|\boldsymbol{x}_i) = \mu_i$ with a more general transformation

$$\mu_i = h(\eta_i) \Longleftrightarrow \eta_i = h^{-1}(\mu_i) = g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{1.6}$$

The *response function* $h(\cdot)$ is required to be bijective and two times continuously differentiable. The inverse $h^{-1}(\cdot) = g(\cdot)$ is called *link function.* Note that the natural parameter $\theta_i$ is also a function of the expectation $\mu_i$, i.e., $\theta_i = \theta(\mu_i)$. Furthermore, the mean is of the form

$$\mu_i = a'(\theta_i) = \frac{\partial a(\theta_i)}{\partial \theta_i} \tag{1.7}$$

and the variance $\mathbb{V}(Y_i|\boldsymbol{x}_i) = \phi\nu(\mu_i)$ results from the second derivative $\nu(\mu_i) = a''(\theta_i) = \frac{\partial^2 a(\theta_i)}{\partial \theta_i^2}$. If the natural parameter directly corresponds to the linear predictor, the link function is called *natural* or *canonical* link function and is given by $g(\mu_i) = \theta(\mu_i) = \eta_i$.

As shown in Table 1.2, both binary $y_i \in \{0,1\}$ and categorial $y_i \in \{1, ..., M\}$ outcomes can be modelled by GLM-type regression models. The binomial distribution is the sum of i.i.d distributed Bernoulli random variables and the multinomial distribution is a generalisation of the binomial distribution for multiple outcomes. A generic form of the (multinomial) logit model is given by

$$\mu_r = \mathbb{P}(Y_i = r|\boldsymbol{x}_i) = \text{logit}^{-1}(\eta_i) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_r)}{\sum_{m=1}^M \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_m)}, \tag{1.8}$$

where additional side constraints have to be specified ensuring the parameters $\boldsymbol{\beta}$ to be identifiable, e.g., $\sum_{m=1}^{M} \boldsymbol{\beta}_m = (0, ..., 0)$.

The parameter estimation in GLM is usually based on the ***M**aximum **L**ikelihood* (ML) principle

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \arg\max_{\boldsymbol{\beta}} \; L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_\theta(y_i), \tag{1.9}$$

where the likelihood $L(\boldsymbol{\beta})$ is given as the product of exponential family distributions due to stochastic independence of response variables $Y_1, ..., Y_n$. Following equation (1.7) and (1.6), the canonical parameter $\theta$ in the exponential family distribution $f_\theta(y_i)$ is determined by $\mu_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta})$ and hence ultimately by $\boldsymbol{\beta}$. Maximising the likelihood is equivalent to maximising the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \left[ f_\theta(y_i) \right] \tag{1.10}$$

due to the monotonicity of the log function. This estimating procedure is typically achieved through iterative numerical methods such as Newton-Raphson or Fisher-Scoring, since an explicit analytical solution in closed-form can only be found for some special cases. More details and examples can be found in Knight (2000, Sect. 5.7).

### Generalised Additive Model

Another extension of the class of regression models is the ***G**eneralised **A**dditive **M**odels* (GAM) proposed by Hastie and Tibshirani (1986). The idea is to combine the aspects of parametric and non-parametric regression models, i.e., some covariates are modelled in the predictor by linear combinations and others by the sum of $J$ unknown functions additionally:

$$\boldsymbol{\eta} = \boldsymbol{\eta}^{\text{lin}} + \boldsymbol{\eta}^{\text{add}}, \text{ with } \boldsymbol{\eta}^{\text{lin}} = \mathbf{X}\boldsymbol{\beta} \text{ and } \boldsymbol{\eta}^{\text{add}} = \sum_{j=1}^{J} f_j(\boldsymbol{z}_j). \tag{1.11}$$

These additive terms are particularly attractive to overcome the problems raised in general situations of type II as described at the beginning of this subsection; for instance, to capture the seasonality which plays a relevant role in the airline industry.

For a given continuous covariate $\boldsymbol{z} = (z_1, ..., z_n)^T$, a *smooth effect* $f$ is assumed and represented by suitable basis function expansions, e.g.,

$$f(\boldsymbol{z}) = \sum_{k=1}^{K} \gamma_k B_k(\boldsymbol{z}) \tag{1.12}$$

for a univariate and

$$f(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sum_{k=1}^{K} \sum_{l=1}^{L} \gamma_{k,l} B_k(\boldsymbol{z}_1) B_l(\boldsymbol{z}_2) \tag{1.13}$$

for a bivariate smooth representing a tensor product, where the (marginal) basis dimensions are given by $K$ and $L$ respectively. For the sake of easier illustration, the following examples are in the univariate case. The basis functions $B(\cdot)$ are evaluated at the observed values $\boldsymbol{z}$, resulting an $n \times K$-dimensional matrix $\mathbf{B} = [B_k(z_i)]_{i=1,...,n;k=1,...,K}$. The corresponding basis coefficient vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)^T$ can be estimated together with the parametric coefficients in $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^T$ by using a composed design matrix $(\mathbf{X}, \mathbf{B}_1, ..., \mathbf{B}_J)$. The resulting model is linear in the parameters and can be estimated within the GLM class via Newton-Raphson-type algorithm.

For the choice of suitable basis functions, Wood (2006, Sect. 4.1) provide a broad overview of different options. Among others, *B-splines* introduced by Schoenberg (1946a,b) is a commonly used basis representation. For sufficiently high basis dimension, resulting splines are continuous and differentiable functions, which can be evaluated efficiently and provide mathematically as well as numerically desirable properties (Boor 1972, 2001). The arbitrary choice of basis dimensions $K$ however, can lead to over-fitting if $K$ is too large and flexibility loss if $K$ is not large enough. Eilers and Marx (1996) proposed a penalised version of B-Splines (hence *P-Splines*) by estimating coefficients for a generous number of B-Spline basis functions with a quadratic penalty based on a *penalised log-likelihood*

$$l_p(\theta) = l(\theta) - \frac{1}{2} \sum_{k=1}^{K} \lambda \boldsymbol{\gamma}^T \mathbf{P} \boldsymbol{\gamma}, \tag{1.14}$$

where $\mathbf{P}$ represents a $K \times K$-dimensional penalty matrix and the smoothing parameter[8] $\lambda$ tuning the influence of the penalty and thus the smoothness of the resulting estimated function $\hat{f} = \mathbf{B}\hat{\boldsymbol{\gamma}}$. In particular for $\mathbf{P} = \mathbf{I}_K$, $\hat{f} \to 0$ when $\lambda \to \infty$ and $\hat{f}$ unpenalised when $\lambda = 0$.

[8] Using this to penalise non-parametric terms is hence called *semi-parametric* approach.

In contrast to the log-likelihood in the case of GLM, there are two kinds of the parameter to be optimised in the case of GAM: the regression coefficient vector $\boldsymbol{\gamma}$ and the smoothing parameter $\lambda$. For fixed $\lambda$, the estimation of $\boldsymbol{\gamma}$ can be performed using ***P**enalised **I**teratively **R**e-weighted **L**east **S**quares* (P-IRLS), see Wood (2006, Sect. 3.4) for details. For finding optimal $\lambda$, different approaches are discussed in Wood (2006, Sect. 4.5). Among other common practices such as the (***G**eneralised*) ***C**ross-**V**alidation* (GCV) or the ***A**kaike **I**nformation **C**riterion* (AIC), restricting the likelihood by decoupling the smoothing parameter $\lambda$ (as variance parameter) and the regression coefficients serves as another alternative. Technically, the variance parameter is thereby fitted by the (scaled) average of the likelihood over all possible values of regression coefficients. This approach can also be derived from the ***M**ixed **M**odel* (MM) perspective, which will be briefly introduced in following. Ruppert et al. (2003) provides fuller coverage on the connection between GAM and MM.

### Mixed Model

Both GLM and GAM can be extended to ***G**eneralised **L**inear **M**ixed **M**odels* (GLMM) and ***G**eneralised **A**dditive **M**ixed **M**odels* (GAMM) with *random effects* $\boldsymbol{b}$, which, in contrast to the fixed coefficients $\boldsymbol{\beta}$, are assumed to be random variables following a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ with covariance matrix $\mathbf{R}$. The (parametric) predictor results to

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{b}, \tag{1.15}$$

and the marginal likelihood of $\boldsymbol{\theta}$ for all observations can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{w} \int \prod_{j=1}^{n_i} f(y_{i,j}|\boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{b}_i|\mathbf{R}) d\boldsymbol{b}_i, \tag{1.16}$$

where the random effects $\boldsymbol{b}_i$ are shared by the (repeated) measurements $y_{i,j}$ within a cluster $i = 1, ..., w$. This structure can be motivated by the heterogeneity of the clustered responses. The clusters can be, e.g., different customer groups or diverse products. Specifying the covariance matrix $\mathbf{R}$ account for correlations among measurements. Moreover, only the variance parameter of the random effects need to be estimated, in comparison to estimating models with fixed effects for each cluster.

For maximising equation (1.16), the model presented in Chapter 2 employs the Laplace approximation (Breslow and Clayton 1993) as a special case in approximating the integral (Pinheiro and Chao 2006). Approximating the data with quasi-likelihood (Wolfinger and O'Connell 1993) would be another option, which however does not allow inferential statements.

### 1.4.2 Summary statistics

The main task in network RM is to find the revenue-optimal allocation of (future) booking request on the total flight network. Hence, the dimension of the data used in the RM system can be massive for airlines with a large network. In practice, analysing massive data is often challenging due to either memory, storage or computation limitation. Confronting such limitations, numerous researchers have been contributed to the application of statistical models based on distributed data, see Caragea et al. (2004), Chu et al. (2013), Lee et al. (2017). The techniques of data management in this area have been developed from dividing the entire dataset into vertical fragments (reducing covariates/coefficients dimension $k$) or horizontal fragments (reducing observations dimension $n$).

Taking the classical linear regression as in equation 1.2 for an illustrative example, directly computing the "all data" estimate $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ from all observations may not be feasible for large scale data, because the design matrix $\mathbf{X}$ with dimension $n \times k$ cannot be constructed within memory as a whole. In order to solve this problem, the large scale data can be partitioned into subsets $s \in \mathcal{S}$, and the corresponding design matrices $\mathbf{X}_s$ with smaller dimensions $n_s \times k$ are created.

One strategy is to estimate each of these horizontal fragments and obtain $\hat{\boldsymbol{\beta}}_s = (\mathbf{X}_s^T\mathbf{X}_s)^{-1}\mathbf{X}_s^T\boldsymbol{y}_s$, then weight the estimates to obtain $\sum_s \mathbf{W}_s\hat{\boldsymbol{\beta}}_s / \sum_s \mathbf{W}_s$ with weighting matrix $\mathbf{W}_s = \mathbf{X}_s^T\mathbf{X}_s$. The use of this weighting matrix rewards data fragments with lower variability and vice versa. Note that the estimation procedure must be performed $|\mathcal{S}|$ times.

A tempting alternative is to compress each data fragment into summary statistics $t_{s,1} = \mathbf{X}_s^T\mathbf{X}_s$ and $t_{s,2} = \mathbf{X}_s^T\boldsymbol{y}_s$, with respective dimensions $k \times k$ and $k \times 1$. Combining the compressed summary statistics via $\sum_s t_{s,1}^{-1} \sum_s t_{s,2} = (\sum_s \mathbf{X}_s^T\mathbf{X}_s)^{-1} \sum_s \mathbf{X}_s^T\boldsymbol{y}_s = \hat{\boldsymbol{\beta}}$ is precisely equivalent to the "all data" estimate due to the matrix properties and the estimating procedure only requires to be run for only one time. For extended regression models such as GL(M)M and GA(M)M, this convenient equivalence does not generally hold due to their non-linear transformation and estimating equations. As a remedy, Xi et al. (2009) and Lin and Xi (2011) proposed to linearise the likelihood with Taylor's expansion and if the derivatives only depend on sufficient statistics $t_s$ and $\hat{\boldsymbol{\beta}}_s$, the aggregated estimate approaches the "all data" estimate asymptotically. Note this estimating strategy with summary statistics avoids accessing raw data from each of the fragments to a central location. In some applications, this is an additional advantageous property to preserve data privacy. Further aspects of data privacy will be discussed in the next section.

## 1.5 Data management

The intense competition among airlines limits the amount of data shared within the industry as well as to academia. Combined with inherent difficulties in linking purchase and pricing data, most academic publications are either based on SP information as described in Section 1.3 or simulations studies, e.g., Bockelie and Belobaba (2017). Taking sales data as RP information, certain technicalities of data management need to be clarified.

First of all, the collection and linkage of different data sources require a standardised procedure. Airlines usually treat ancillary services with the same business rules as flight tickets (TKT) on a supporting ***E**lectronic **M**iscellaneous **D**ocuments* (EMD) and require association between the flight tickets and EMD. For each specific ancillary service (e.g. ASR), further data sets which contain detailed features, e.g., seat row and column, need to be acquired. Note that most airlines only maintain TKT and EMD data but not specific ancillary data in the past (Ratliff and Gallego 2013). The linkage between these data sets can be achieved through ***P**assenger **N**ame **R**ecord* (PNR), which is also commonly known as *Booking Reference* to the customer. Table 1.3 illustrates the structure of the matched data.

| Data set: | PNR | | | ... | TKT | | | | ... | EMD | | | ... | ASR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable: | Booking reference | ... | Number of passengers | ... | Flight number | Orig. & Dest. | ... | Ticket price | ... | Ancillary item | ... | Ancillary price | ... | Seat |
| | ABC123 | ... | 1 | ... | 811 | A-B | ... | 666 | ... | NA | ... | NA | ... | NA |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | XYZ789 | ... | 2 | ... | 111 | A-C | ... | 999 | ... | ASR | ... | 30 | ... | 1A |
| | XYZ789 | ... | 2 | ... | 111 | A-C | ... | 999 | ... | ASR | ... | 30 | ... | 1B |

Table 1.3: Structure of the matched booking data: PNR is the most practised matching variable.

PNR was originally used to compare flight passenger data (e.g. identity and flight destination) against information held by security and border agencies. Note the information contained in PNR are not only for the ordinary course of airline business (i.e. enabling reservations and carrying out the check-in process) but also they can be sensitive personal information helping to fight terrorism and other crimes (The European Parliament and of the Council 2016a). Facing the **G**eneral **D**ata **P**rotection **R**egulation (GDPR[9]), data set contains PNR as an identifier arise a data privacy concern. De-identifying data by removing sensitive identifiers (e.g. PNR) and apply disclosure control methods can not completely remove the risk of privacy breach (Fienberg 2006; Hundepool et al. 2012). For future research with reproducible airline data, more privacy conscious matching procedures need to be standardised by the industry.

[9]See more details in The European Parliament and of the Council (2016b) and Tikkinen-Piri et al. (2018)

From the data technical side, the program *ONE Order* from the International Air Transport Association (2018b) visions a data communication standard to support secured and simplified data management by the year 2021. It aims to supersede the current booking and ticketing records including EMD and PNR as well as to combine the content of those records into a single retail and customer oriented order. On the one hand, the data quality can be improved by avoiding mismatch and redundant or overlapping records. On the other hand, more data privacy protection can be achieved.

For the current and future data usage, although the pricing schemes in RM systems do not base on personal data directly and are anonymous[10] information, the practice of differential pricing would generally fall under the scope of GDPR, despite the degrees of price differentiation (Steppe 2017). Facing this and further challenges, it is suggested to the airline industry to either apply models based booking data in a distributed manner using summary statistics as discussed in Section 1.4.2 or consider the modelling approaches based on aggregated data, e.g., as in Chapter 3. Both approaches make innovative changes in the demand modelling step and the corresponding data collection step in the RM system (recall Figure 1.1) and steer clear of the accessing of raw data which arise privacy concern and still allow to access and quantify the WTP and PE of (and beyond) ancillary services.

Last but not least, for the completeness of use-case information, it is important to mention that Chapter 2 analyses intercontinental flights due to higher relevance of ASR, whereas Chapter 3 examined continental markets, because branded fares were not available in intercontinental markets yet at the time of the research. Thus, the generalisation of interpreted results must consider this limitation, despite the modelling approach is universally applicable to the data in the respective markets.

[10]These data are not related to an identified person and are anonymous from a statistical perspective, but not necessarily from a data protection perspective.

# References


Adams, W. J. and Yellen, J. L. (1976). "Commodity Bundling and the Burden of Monopoly". In: *The Quarterly Journal of Economics* 90(3), pp. 475–498.

American Airlines (1987). *The Art of Managing Yield*. Tech. rep. Dallas, TX: American Airlines.

Balcombe, K., Fraser, I., and Harris, L. (2009). "Consumer willingness to pay for in-flight service and comfort levels: A choice experiment". In: *Journal of Air Transport Management* 15(5), pp. 221–226.

Barankin, E. W. and Maitra, A. P. (1963). "Generalization of the Fisher-Darmois-Koopman-Pitman Theorem on Sufficient Statistics". In: *Sankhyā: The Indian Journal of Statistics, Series A* 25(3), pp. 217–244.

Beckmann, M. J. (1958). "Decision and Team Problems in Airline Reservations". In: *Econometrica* 26(1), pp. 134–145.

Belobaba, P. (1987). "Air Travel Demand and Airline Seat Inventory Management". PhD thesis. MIT.

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Beris, A. N. and Giacomin, A. J. (2014). "πάντα ῥεῖ (everything flows): Motto for rheology". In: *Applied Rheology* 24(5), pp. 1–13.

Bockelie, A. and Belobaba, P. (2017). "Incorporating ancillary services in airline passenger choice models". In: *Journal of Revenue and Pricing Management* 16(6), pp. 553–568.

Boor, C. de (1972). "On calculating with B-splines". In: *Journal of Approximation Theory* 6(1), pp. 50–62.

Boor, C. de (2001). *A practical guide to splines*. New York, NY: Springer.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* Hoboken, NJ: John Wiley & Sons.

Breslow, N. E. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association* 88(421), pp. 9–25.

Caragea, D., Silvescu, A., and Honavar, V. (2004). "A Framework for Learning from Distributed Data Using Sufficient Statistics and its Application to Learning Decision Trees". In: *International Journal of Hybrid Intelligent Systems* 1(1-2), pp. 80–89.

Chiang, W.-C., Chen, J. C. H., and Xu, X. (2007). "An overview of research on revenue management: current issues and future research". In: *International Journal of Revenue Management* 1(1), pp. 97–128.

Choi, S. and Mattila, A. S. (2004). "Hotel revenue management and its impact on customers' perceptions of fairness". In: *Journal of Revenue and Pricing Management* 2(4), pp. 303–314.

Chu, E., Keshavarz, A., and Boyd, S. (2013). "A distributed algorithm for fitting generalized additive models". In: *Optimization and Engineering* 14(2), pp. 213–224.

Cleophas, C., Frank, M., and Kliewer, N. (2009). "Recent developments in demand forecasting for airline revenue management". In: *International Journal of Revenue Management* 3(3), pp. 252–269.

Correia, A., Pimpão, A., and Tão, M. (2012). "Willingness to pay for frills when travelling with low-cost airlines". In: *Tourism Economics* 18(6), pp. 1161–1174.

Cui, Y., Duenyas, I., and Sahin, O. (2018). "Unbundling of Ancillary Service: How Does Price Discrimination of Main Service Matter?" In: *Manufacturing & Service Operations Management* 20(3), pp. 455–466.

Eilers, P. H. C. and Marx, B. D. (1996). "Flexible Smoothing with B-splines and Penalties". In: *Statistical Science* 11(2), pp. 89–121.

Ellison, G. (2005). "A Model of Add-On Pricing". In: *The Quarterly Journal of Economics* 120(2), pp. 585–637.

Executive Office of the President of the United States (2015). *Big data and differential pricing.* Tech. rep., pp. 1–21.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression.* Berlin, Germany: Springer.

Fienberg, S. E. (2006). “Privacy and Confidentiality in an e-Commerce World : Data Mining , Data Warehousing , Matching and Disclosure Limitation”. In: *Statistical Science* 21(2), pp. 143–154.

Fiig, T., Isler, K., Hopperstad, C., and Olsen, S. S. (2012). “Forecasting and optimization of fare families”. In: *Journal of Revenue and Pricing Management* 11(3), pp. 322–342.

Fishburn, P. C. (1970). *Utility Theory for Decision Making*. Huntington, NY: Robert E. Krieger.

Galton, F. (1886). “Regression towards mediocrity in hereditary stature”. In: *Journal of the Anthropological Institute* 15, pp. 246–263.

Garrow, L. A., Hotle, S., and Mumbower, S. (2012). “Assessment of product debundling trends in the US airline industry: Customer service and public policy implications”. In: *Transportation Research Part A: Policy and Practice* 46(2), pp. 255–268.

Geraghty, M. K. and Johnson, E. (1997). “Revenue management saves National Car Rental”. In: *Interfaces* 27(1), pp. 107–127.

Gruß, C. (2008). *Revenue-Management in der Automobilindustrie*. Wiesbaden, Germany: Gabler.

Han, C.-P. and Li, J. (2009). “Evaluating Estimation Techniques of Transportation Price Elasticity”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2115(1), pp. 94–101.

Hastie, T. and Tibshirani, R. (1986). “Generalized Additive Models”. In: *Statistical Science* 1(3), pp. 297–318.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and Wolf, P.-P. de (2012). *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons.

Huntington, P. A. (1993). “Ticket pricing policy and box office revenue”. In: *Journal of Cultural Economics* 17(1), pp. 71–87.

IdeaWorks (2018). *Airline Ancillary Revenue Projected to Be $92.9 Billion Worldwide in 2018*. Retrieved November 30, 2018. Available: `https://www.ideaworkscompany.com/wp-content/uploads/2018/11/Press-Release-133-Global-Estimate-2018.pdf`.

Iliescu, D. C., Garrow, L. A., and Parker, R. A. (2008). “A hazard model of US airline passengers’ refund and exchange behavior”. In: *Transportation Research Part B: Methodological* 42(3), pp. 229–242.

International Air Transport Association (2018a). *Fact Sheet: Industry Statistics*. Retrieved December 3, 2018. Available: `https://www.iata.org/pressroom/facts_figures/fact_sheets/Documents/fact-sheet-industry-facts.pdf`.

International Air Transport Association (2018b). *One Order: Program Strategy Paper*. Tech. rep. May, pp. 1–24.

Kimms, A. and Müller-Bungart, M. (2007). "Revenue management for broadcasting commercials: the channel's problem of selecting and scheduling the advertisements to be aired". In: *International Journal of Revenue Management* 1(1), pp. 28–44.

Knight, K. (2000). *Mathematical Statistics*. Boca Raton, FL: Chapman and Hall/CRC.

Ladany, S. P. and Arbel, A. (1991). "Optimal cruise-liner passenger cabin pricing policy". In: *European Journal of Operational Research* 55(2), pp. 136–147.

Lee, J. Y. L., Brown, J. J., and Ryan, L. M. (2017). "Sufficiency Revisited : Rethinking Statistical Algorithms in the Big Data Era". In: *The American Statistician* 71(3), pp. 202–208.

Lin, N. and Xi, R. (2011). "Aggregated estimating equation estimation". In: *Statistics and Its Interface* 4(1), pp. 73–83.

Littlewood, K. (1972). "Forecasting and control of passenger bookings." In: *Proceedings of the AGIFORS Symposium* 4(2), pp. 95–128.

Marcus, B. and Anderson, C. K. (2008). "Revenue management for low-cost providers". In: *European Journal of Operational Research* 188(1), pp. 258–272.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London, UK: Chapman and Hall.

McFadden, D. (1973). "Conditional logit analysis of qualitative choice behaviour". In: *Frontiers in Econometrics*. New York, NY: Academic Press. Chap. 4, pp. 105–142.

McFadden, D. (1981). "Econometric models of probalistic choice". In: *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press. Chap. 5, pp. 198–272.

McFadden, D. (1984). "Econometric analysis of qualitative response models". In: *Handbook of Econometrics*. Amsterdam, Netherlands: Elsevier. Chap. 24, pp. 1395–1457.

McGill, J. I. and van Ryzin, G. (1999). "Revenue Management: Research Overview and Prospects". In: *Transportation Science* 33(2), pp. 233–256.

Menezes, A. G. de and Vieira, J. C. (2008). "Willingness to pay for airline services attributes : evidence from a stated preferences choice game". In: *European Transport* 39, pp. 1–13.

Morrison, S. and Winston, C. (1986). *The economic effects of airline deregulation*. Washington, DC: Brookings Institution Press.

Müller-Bungart, M. (2007). *Revenue Management with Flexible Products*. Berlin, Germany: Springer.

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135(3), pp. 370–384.

Oxford Dictionaries (2018). *Definition of ancillary*. Retrieved November 30, 2018. Available: https://en.oxforddictionaries.com/definition/ancillary. (Visited on 11/30/2018).

Pigou, A. C. (1932). *The Economics of Welfare*. 4th ed. London, UK: MacMillian.

Pinheiro, J. C. and Chao, E. C. (2006). "Efficient laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models". In: *Journal of Computational and Graphical Statistics* 15(1), pp. 58–81.

Poelt, S. (2016). "History of revenue management - From leg to O&D". In: *Journal of Revenue and Pricing Management* 15(3-4), pp. 236–241.

Ratliff, R. and Gallego, G. (2013). "Estimating sales and profitability impacts of airline branded-fares product design and pricing decisions using customer choice models". In: *Journal of Revenue and Pricing Management* 12(6), pp. 509–523.

Rothstein, M. (1985). "OR and the Airline Overbooking Problem". In: *Operations Research* 33(2), pp. 237–248.

Ruhnau, T. (2012). *Auktionen und Revenue Management in der Automobilindustrie*. Wiesbaden, Germany: Springer.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Schoenberg, I. J. (1946a). "Contributions to the problem of approximation of equidistant data by analytic functions: Part a. – on the problem of smoothing or graduation. a first class of analytic approximation formulae." In: *Quarterly of Applied Mathematics* 4(1), pp. 45–99.

Schoenberg, I. J. (1946b). "Contributions to the problem of approximation of equidistant data by analytic functions: Part b. – on the problem of osculatory interpolation. a second class of analytic approximation". In: *Quarterly of Applied Mathematics* 4(2), pp. 112–141.

Shen, Z.-J. M. and Su, X. (2007). "Customer Behavior Modeling in Revenue Management and Auctions: A Review and New Research Opportunities". In: *Production and Operations Management* 16(6), pp. 713–728.

Siddappa, S., Rosenberger, J. M., and Chen, V. C. P. (2008). "Optimising airline overbooking using a hybrid gradient approach and statistical modelling". In: *Journal of Revenue and Pricing Management* 7(2), pp. 207–218.

Simon, H. (1989). *Price management*. Amsterdam, Netherlands: Elsevier.

Stefanescu, C. (2009). "Multivariate Customer Demand: Modeling and Estimation from Censored Sales". In: *SSRN*, pp. 1–35.

Steppe, R. (2017). "Online price discrimination and personal data: A General Data Protection Regulation perspective". In: *Computer Law & Security Review: The International Journal of Technology Law and Practice* 33(6), pp. 768–785.

Strauss, A. K., Klein, R., and Steinhardt, C. (2018). "A review of choice-based revenue management: Theory and methods". In: *European Journal of Operational Research* 271(2), pp. 375–387.

Talluri, K. and van Ryzin, G. (2004). *The Theory and Practice of Revenue Management*. New York, NY: Springer.

The European Parliament and of the Council (2016a). "Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime". In: *Official Journal of the European Union* L 119, pp. 132–149.

The European Parliament and of the Council (2016b). "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)". In: *Official Journal of the European Union* L 119, pp. 1–88.

Tikkinen-Piri, C., Rohunen, A., and Markkula, J. (2018). "EU General Data Protection Regulation : Changes and implications for personal data collecting companies". In: *Computer Law & Security Review: The International Journal of Technology Law and Practice* 34(1), pp. 134–153.

Train, K. (2009). *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge, UK: Cambridge University Press.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge, UK: Cambridge University Press.

Vinod, B. (2005). “Retail revenue management and the new paradigm of merchandise optimisation”. In: *Journal of Revenue and Pricing Management* 3(4), pp. 358–368.

Vinod, B. and Moore, K. (2009). “Promoting branded fare families and ancillary services: Merchandising and its impacts on the travel value chain”. In: *Journal of Revenue and Pricing Management* 8(2-3), pp. 174–186.

Vulcano, G., Ryzin, G. van, and Chaar, W. (2010). “Choice-Based Revenue Management: An Empirical Study of Estimation and Optimization”. In: *Manufacturing & Service Operations Management* 12(3), pp. 371–392.

Wittmer, A. and Rowley, E. (2014). “Customer value of purchasable supplementary services: The case of a European full network carrier’s economy class”. In: *Journal of Air Transport Management* 34(C), pp. 17–23.

Wolfinger, R. and O’Connell, M. (1993). “Generalized Linear Mixed Models: A Pseudo-Likelihood Approach”. In: *Journal of Statistical Computation and Simulation* 48(3-4), pp. 233–243.

Wood, S. (2006). *Generalized Addtive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Xi, R., Lin, N., and Chen, Y. (2009). “Compression and Aggregation for Logistic Regression Analysis in Data Cubes”. In: *IEEE Transactions on Knowledge and Data Engineering* 21(4), pp. 479–492.

# Chapter 2

# “A la carte pricing” with semi-individual booking data

Chapter 2 presents a practice oriented research article for the application of GAM and GLMM to the “a la carte” pricing problem. The joint modelling approach answers questions of whether, when and which seats are selected by passengers.

**Contributing article:**

Shao, S., Kauermann, G., Smith, M. S. (2018) Whether, When and Which: Modelling Advanced Seat Reservation by Airline Passengers. Submitted to *Transportation Research Part A: Policy and Practice*, available at *SSRN*: `https://ssrn.com/abstract=3325346`.

**Author contributions:**

Shuai Shao 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 and Michael Stanley Smith provided valuable inputs to all sections of the article. All author revised and and proofread the manuscript.

# Whether, when and which: modelling advanced seat reservations by airline passengers

Shuai Shao, Göran Kauermann, Michael Stanley Smith

Submitted to

*Transportation Research Part A: Policy and Practice*

on

10. Sep. 2018

Shuai Shao is a doctoral student and Göran Kauermann is Chair of Applied Statistics, both at Department of Statistics, Ludwig-Maximilians-Universität München. Michael Stanley Smith is Chair of Management (Econometrics) at Melbourne Business School, University of Melbourne. Correspondence should be directed to Göran Kauermann per email at goeran.kauermann@stat.uni-muenchen.de or per post to Ludwigstraße 33, D-80539 München, Germany.

**Abstract**

Motivated by the growing importance of ancillary revenues in the airline industry, we propose a statistical model for the behavior of airline passengers making Advanced Seat Reservations (ASR). We focus on the questions of whether, when and which seats are selected. To address these questions, we employ a discrete time duration model, combined with a discrete choice model. Both employ unknown smooth covariate effects, that are estimated using contemporary P-spline methodology. This is applied to a large database of bookings on five intercontinental routes. By incorporating random effect terms to account for seat-specific heterogeneity, we find strong evidence of “middle seat avoiding” and “front seat preferring” effects. We also show that the willingness to pay for ASR depends on its price in relation to the ticket price, as well as on the distribution channel. These and other insights allow for product differentiation and variable pricing in ASR for each and every seat. In addition, the statistical model can also be used for other ancillary products — such as on-board dining and preferential baggage checking — allowing dynamic pricing of ancillary products in general.

**Keywords:** Advanced Seat Reservation; Airline Ancillary Pricing; Customer Behaviour Analyis, Discrete Choice Model; Discrete Time Duration Model; Willingness To Pay

# 1 Introduction

In the age of internet-based search engines, airlines are confronted with increased price transparency. For traditional airlines this has led to increased competition with low cost carriers. Together with numerous other market changes, such as the global economic slow down and new travel policies (mainly cost cutting) of companies, airlines have responded by offering new ancillary services to increase revenues. **A**dvanced **S**eat **R**eservation (ASR) — the ability to select specific seats prior to check-in — is one of these services. The aim of our study is explore the extent to which passengers are willing to pay for ASR using a new and large booking dataset.

To our knowledge, Lee and Luengo-Prado [2004] made the first contribution in this area by asking if passengers are willing to pay more for additional legroom by comparing two different settings of increased seat pitch. They found that the **W**illingness **T**o **P**ay (WTP) is higher if seat pitches are increased for some rows which can only be reserved by passengers who pay a full fare. This suggests that price discrimination is possible only if there is heterogeneity in the product value. Until recently, there has been very limited variability in airline seat selection, with only two products: legroom and standard seats. We show in this paper that there is substantial heterogeneity in customer preference for ASR at the individual seat level, which results in high variability in passengers' WTP for reserving each and every seat.

A number of studies have found heterogeneity in seat preference in other industries. For example, for train travel Wardman and Murphy [2015] show that seats and their configuration are fundamental part of the journey experience. They investigated how seating preferences depend on factors such as travel distance and journey purpose. Taking also the price component into account, Leslie [2004] studied profit implications of price discrimination for a Broadway play based on seat quality in theatres. Veeraraghavan and Vaidyanathan [2012] also measured seat value perceived by consumers at baseball stadiums.

However, before the recent introduction of ASR as an ancillary service in the airline industry, airline seats within the same booking class had the same price, and seat location in the plane was not considered. In direct contrast, it has long been the case that prices of seats in stadiums or theatres are valued higher when they offer a better view of the event. With different seat values based on seat location, the price of a selected seat (or more generally an ancillary product) can indicate how much a passenger values a particular seat. Espino et al. [2008] used an experiment in which Spanish travellers stated their preferences in six service attributes, including additional legroom for a short haul flight of two virtual airlines. Depending on the model specification, these authors found that the WTP for additional legroom varied between 15€ and 34€. In contrast, Garrow et al. [2012] found that passengers of Delta Air Lines were unwilling to pay for extra legroom on short flights, but were willing to purchase ASR on international markets. Balcombe et al. [2009] applied a Bayesian model in a related survey and found a comparable WTP for seats with more legroom. Studies from other scientific disciplines have also focused on legroom, e.g. from the ergonomic aspect (see Kremser et al. [2012] and Vink et al. [2012]).

ASR does not only apply to seats with extra legroom, and customer preferences in choosing between standard seats is also of interest. For instance Daft and Albers [2012] took both standard and legroom-seats into their profitability calculation for long haul flights of low-cost carriers. Caussade and Hess [2009] extracted WTP for some service attributes including standard and preferential (similar to legroom) seats from stated preference data in a branded fare context. Mumbower et al. [2015] contributed to the literature on ancillary fees by providing the first insights into the role of load factors and seat-map displays of all ASR-seats on customers’ purchasing behavior of seats with extra legroom using revealed preference data.

In this paper, we contribute to this research by considering the probability of making an ASR, as well as the preference for selecting between different seats. By focusing on standard ASR, and not legroom ASR, we want to answer three questions: whether and when passengers make use of standard ASR, and if so, which seats are preferred?

To shed light on these questions we consider a statistical modelling approach with multiple components. The probability that passengers make a seat reservation is modelled in the first part as a discrete time-to-event model, see e.g. Tutz and Schmid [2016]. We include non-parametric smooth terms in the model to accommodate the influence of continuous covariates, such as time of year to capture seasonality and time to departure. We make use of penalized spline estimation following the original ideas of Eilers and Marx [1996] and the further developments of Ruppert et al. [2003] and Wood [2017]. The second part of modelling focuses on seat selection, conditional on an ASR being made. For this, we employ a multinomial choice model (see e.g. Train [2009] and Fahrmeir and Tutz [2001]). Here, a random effect for seat number is included to account for seat heterogeneity. All models fall within the framework of generalized additive mixed models originally proposed in Hastie and Tibshirani [1990] and extensively extended; see Wood [2017] for a recent exposition.

A rich set of flight, booking and seat-specific factors are found to determine customer preference for ASR. As for flight-specific factors, departure day of year is a strong seasonal component. Moreover, bookings that include multiple passengers are more likely to reserve seats in advance. Compared to bookings made via intermediaries, passengers who book directly by airlines are more likely to reserve seats in advance. Seat-specific factors reveal that passengers prefer front rows and avoid middle seats. Another key result is that the price sensitivity for seat reservation depends on ticket price and decays towards departure. Last, we show that the revenue implications of adopting dynamic pricing for ASR based on our statistical model are substantial in dollar terms.

The remainder of this paper is organized as follows. In Section 2 we introduce the data using in the study. We outline the statistical model in Section 3, and provide an overview of the estimation method employed in Section 4. In Section 5 we discuss the empirical results, and run a comparison of the predicted and realised revenues as model validation in Section 6. We focus on the implications for ancillary revenue generation in Section 7 before concluding in Section 8.

## 2 Data

We analyse five intercontinental routes of a major European airline which wishes to remain anonymous, so that throughout this paper we refer to it as "AirABC". The data were collected from destinations $A$, $B$ and $C$ in South-America, $D$ in Asia and $E$ in North-America. All these routes were served by flights originating from the same European city by the same aircraft type (Boeing 747-8) with an identical seat-map, which is depicted in Figure 1.

The data were collected from economy class passengers who departed between February 2015 and December 2016. The data can be divided into two time periods. The first period is between February 2015 and January 2016, where ASR was offered at a constant price. From February 2016 to December 2016, the prices for ASR were experimentally varied in route $A$ and $B$ in an attempt to access and quantify price elasticity. Such experiments are rare in the airline industry and experimental pricing on a grand scale is difficult and in our case conflicted with transparent pricing communication for customers.

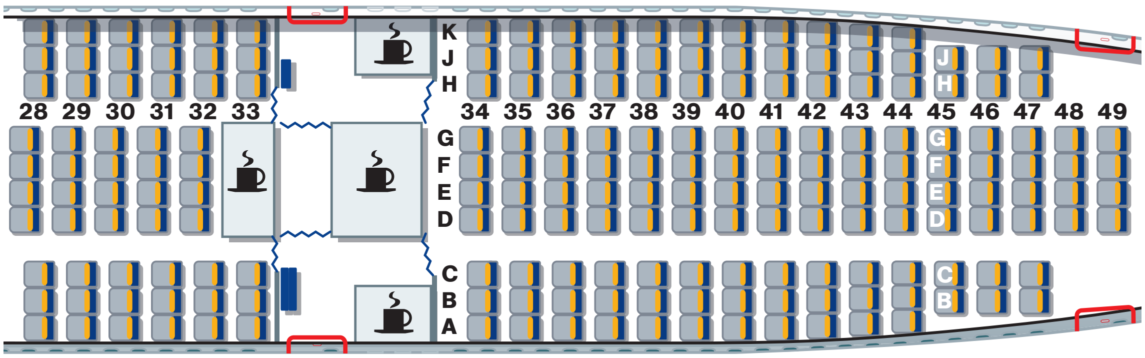


Figure 1: Seat-map of the analysing economy cabin of Boeing 747-8 by AirABC.

We therefore use the log relative price for ASR as the price variable, that is defined as

$$p = \log\left(\frac{\text{Price of ticket} + \text{Price of ASR}}{\text{Price of ticket}}\right).$$

Treating the ticket price, and hence $p$, as an exogenous variable is problematic when estimating price elasticity using demand data. This can lead to biased estimates, see e.g. Davidson and Mackinnon [1993] Davidson and Mackinnon [1999], Wooldridge [2012] or Petrin and Train [2010]. In this paper, however, we condition on bookings, because ASR can only be carried out once a ticket is booked. This implies that the ticket price is fixed and the endogenous relationship between booking process and price is accounted for. We therefore can treat ticket price — and also $p$ — as exogenous when modelling ASR.

The observational unit in our study is the booking of an itinerary. A single itinerary booking can include multiple passengers, as well as multiple flight segments. Therefore, the entire trip of a booking may include more than the considered intercontinental flight (i.e. route $A$, $B$, $C$, $D$ or $E$), such as a connecting flight before or after the intercontinental flight. However, in this study we only consider the outbound flights for the five routes (i.e. without inbound flights if there are any in this booking), so that no dependencies between the flights within one booking are considered. If we did not, we would have to account for the dependence between the inbound and outbound legs of a return flight. We are aware of the fact that this is a major simplification of ASR. But for the purpose of presentation and model validation as well as interpretation, it is helpful to simplify the task.

In total we have $485,279$ observations of bookings: $254,849$ from the data in Period I (February 2015 to January 2016) and $230,430$ from the data in Period II (February 2016 to December 2016). The data for routes $A$ and $B$, where the prices for ASR were varied in an experimental setup, will allow us to check the validity of using the log relative price as our price variable, by computing the out-of-sample prediction error for the experimental data in Period II. In Table 1 we give the detailed numbers of booking for the different routes in each and both data periods.

| Route: | | A | B | C | D | E | Total |
|---|---|---|---|---|---|---|---|
| Data Period: | I | 49,687 | 48,017 | 54,624 | 41,628 | 60,893 | 254,849 |
| | II | 42,883 | 44,838 | 52,138 | 36,707 | 53,864 | 230,430 |
| | I & II | 92,570 | 92,855 | 106,762 | 78,335 | 114,757 | 485,279 |

Table 1: Number of bookings in each data period (rows) and route (columns).

## 2.1 Covariate Variables

We include a number of covariates in our analysis in order to identify the driving factors behind ASR. These covariates can be divided into three categories; namely flight-specific, booking-specific and seat-specific quantities. To begin with the first, let $f$ be the index for a particular flight and $x_f^{\text{flight}}$ be the flight-specific covariates. In our study these are

$$\begin{aligned} x_f^{\text{flight}} &= \{\text{departure day of year, departure day of week}\} \\ &= \{d_f, \text{WDAY}_f\}. \end{aligned}$$

Note that $d \in \{1, \dots, 366\}$ and WDAY has *Monday* as a reference category. Note that our data covers 2016, which was a leap year with 366 days. Throughout this paper, we label categorical variables (such as departure day of week) with capitalized abbreviations, and other variables with single lower case letters. The specific levels of our categorical variables are written in italics.

Let $i = 1, \dots, n_f$ be the index of the bookings for a particular flight $f$ and $x_{f,i}^{\text{booking}}$ be booking-specific covariates for bookings $i = 1, \dots, n_f$. We include:

$$\begin{aligned} x_{f,i}^{\text{booking}} = \{&\text{days to departure when booking ticket, days to departure when selecting seat,} \\ &\text{log relative price for ASR, multiple passenger booking, distribution channel}\} \\ = \{&t_{f,i}, s_{f,i}, p_{f,i}, \text{MULTI}_{f,i}, \text{CHNL}_{f,i}\}. \end{aligned}$$

Days to departure is a negative ordinal variable ranging from $-365$ to 0, where 0 is the day of departure. A seat can only be reserved if a booking has already been made, so that $t_{f,i} \leq s_{f,i}$, where $t_{f,i} = s_{f,i}$ occurs if the booking and ASR have been carried out at the same time. Depending on the route and airline, seat reservations are free for all passengers via online check-in at some time point close to departure. In our data the online check-in process starts two days before departure, after which seat selection is free of charge. Thus, we do not include observations where the ticket is booked at $t_{f,i} > -2$. Furthermore, we consider all seats selected during the online check-in process as <u>not</u> ASR by definition. In this case we set $s_{f,i} = 0$, indicating a booking without ASR. We give more details subsequently. The covariate MULTI is a dummy variable for whether, or not, there were multiple passengers in the booking. The covariate CHNL indicates the distribution channel of the booking. There are four channels here: *Direct*, which means the ticket was booked by AirABC directly[1]; *Chains*, which means the ticket was booked through traditional travel chains; *OTA*, which means the ticket was booked by an **O**nline **T**ravel **A**gency (OTA); and *Others*, which includes all other options. A summary of all covariates except departure day of year[2] is given in Table 2. Note that the total count by variables for ASR differs from the total count by variables for bookings.

Finally, seat-specific covariates are denoted by $x_l^{\text{seat}}$, where subscript $l$ indexes the possible seats for ASR, with $l = 1, ..., L$. We include the following seat-specific covariates:

$$\begin{aligned} x_l^{\text{seat}} &= \{\text{window, middle or aisle seat, plane section of seat, special seat}\} \\ &= \{\text{WMA}_l, \text{SECT}_l, \text{SPEC}_l\}. \end{aligned}$$

In our case there are $L = 144$ seats being offered for ASR on every flight. The number of reservations recorded for each of these seats in our data are depicted in Figure 2. We take the *Middle* seats as the reference category of covariate WMA. We also separate the cabin into three sections, with reference category *Front* for rows $28 - 33$, and categories *Middle* and *Back* for rows $39 - 44$ and $45 - 49$, respectively. The seats $B$, $C$, $H$, and $J$ in rows $45 - 47$ are special seats because there are only two adjacent seats and they are not in a row of three seats; see also Figure 1. We will discuss the effects of these and other covariates in detail in Section 5.3.

[1] This includes bookings via call-center, city-office, official homepage or mobile application.
[2] A summary of this covariate would not be informative, since all routes were served on a daily basis.

| Numerical variables (booking-specific): | | Min. | $1^{st}$Qu. | Median | Mean | $3^{rd}$Qu. | Max. |
|---|---|---|---|---|---|---|---|
| Time point of booking ($t$) | | -365.0 | -81.0 | -36.0 | -58.7 | -15.0 | -2.0 |
| Time point of ASR ($s \neq 0$) | | -362.0 | -60.0 | -25.0 | -46.8 | -10.0 | -2.0 |
| Log relative price ($p$ in %) | | 0.0 | 1.3 | 2.3 | 2.1 | 3.1 | 5.7 |
| Categorical variables for bookings (flight-specific): | | | | | | Count | Prop. |
| Departure day of week (WDAY) | | | | | | | |
| | *Monday* | | | | | 70150 | 14.5% |
| | *Tuesday* | | | | | 66365 | 13.7% |
| | *Wednesday* | | | | | 64204 | 13.2% |
| | *Thursday* | | | | | 65901 | 13.6% |
| | *Friday* | | | | | 69591 | 14.3% |
| | *Saturday* | | | | | 75007 | 15.4% |
| | *Sunday* | | | | | 74061 | 15.3% |
| Categorical variables for bookings (booking-specific): | | | | | | | |
| Multiple passengers (MULTI) | | | | | | | |
| | *No* | | | | | 369248 | 76.1% |
| | *Yes* | | | | | 116031 | 23.9% |
| Distribution channel (CHNL) | | | | | | | |
| | *Direct* | | | | | 112712 | 23.2% |
| | *Chains* | | | | | 142073 | 29.3% |
| | *OTA* | | | | | 60988 | 12.6% |
| | *Others* | | | | | 169506 | 34.9% |
| Categorical variables for ASRs (seat-specific): | | | | | | | |
| Window, middle or aisle seat (WMA) | | | | | | | |
| | *Window* | | | | | 36376 | 27.3% |
| | *Middle* | | | | | 24452 | 18.4% |
| | *Aisle* | | | | | 72155 | 54.3% |
| Plane section (SECT) | | | | | | | |
| | *Front* | | | | | 76029 | 57.2% |
| | *Middle* | | | | | 26017 | 27.1% |
| | *Back* | | | | | 20937 | 15.7% |
| Special seat (SPEC) | | | | | | | |
| | *No* | | | | | 115587 | 86.9% |
| | *Yes* | | | | | 17139 | 13.1% |

Table 2: Summary of covariates across all five routes. The upper rows present summaries of the numerical covariates, which are all positively skewed. The lower rows present the summaries for the categorical covariates. Bookings are dominated by single passengers.

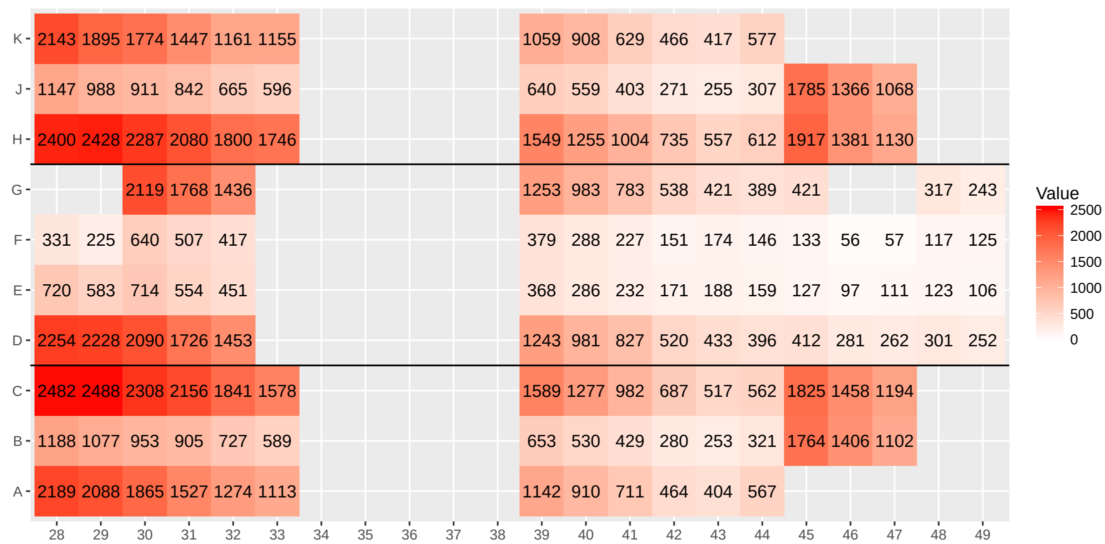


Figure 2: Total frequency of selected seats of all ASRs in our data. While in Figure 1 there are 198 seats, some are blocked out of the ASR-booking system because they are reserved for customers with special loyalty status or who need mobility assistance.

## 2.2 Response Variables

The first response variable is whether ASR has been made or not, which we denote as

$$y_{f,i}(s) = \begin{cases} 1, & \text{if ASR is made at the time point } s \\ 0, & \text{otherwise,} \end{cases}$$

where $s \in \{t_{f,i}, ..., -2\}$. In our statistical model, we further distinguish bookings where $s = t_{f,i}$, i.e. an ASR and booking have been made together at the same time point (about 12% of all bookings) or $t_{f,i} < s_{f,i} \leq -2$, i.e. an ASR was made after booking the ticket (about 8% of all bookings).

# 3 Model

As discussed in the introduction, we focus on the questions of: whether and when an ASR is made, and if so, which seat(s) are selected. To this end, we decompose the model into two parts. The first models the if and when an ASR is made, while the second accounts for seat selection.

## 3.1 Timing of ASR

We consider $y_{f,i}(s)$ as a stochastic process taken as a discrete time duration model. This accounts for the probability of making an ASR, given that an ASR has not been made before. As we have stated previously, the majority of ASRs are made at the time of booking, and we model these separately. This leads to a two component model for the probabilities:

$$\text{ASR Model 1 (ASR with booking): } \mathbb{P}[y_{f,i}(t_{f,i}) = 1] \text{ and} \tag{1}$$

$$\text{ASR Model 2 (ASR after booking): } \mathbb{P}[y_{f,i}(s) = 1 | y_{f,i}(t) = 0 \text{ for } t_{f,i} \leq t < s; s \leq -2], \tag{2}$$

where for simplicity, we omitted the covariates in the above notation. The ASR Model 1 is modelled using a logit model:

$$\begin{aligned} &\mathbb{P}[y_{f,i}(t_{f,i}) = 1] = \text{logit}^{-1}(\eta_1^{\text{A}}), \text{ where} \\ &\eta_1^{\text{A}} = \beta_{1,0} + \eta_1^{\text{flight}}(x_f^{\text{flight}}) + \eta_1^{\text{booking}}(x_{f,i}^{\text{booking}}). \end{aligned} \tag{3}$$

We use superscript A to refer to this aspect of the ASR model and the first subscript 1 represents being the first part of this model. As covariates we include flight- and booking-specific quantities. The first block of covariates enter the model in a semi-parametric manner through

$$\eta_1^{\text{flight}}(x_f^{\text{flight}}) = \text{WDAY}_f \cdot \boldsymbol{\beta}_{1,1}^{\text{flight}} + m_1^{\text{flight}}(d_f), \tag{4}$$

where $m_1^{\text{flight}}$ is a smooth periodic function capturing seasonal variation. The dot between $\text{WDAY}_f$ and the coefficients $\boldsymbol{\beta}_{1,1}^{\text{flight}}$ denotes the usual expansion of a categorical variable (omitting the reference category). In the second block, booking-specific covariates are included through

$$\begin{aligned}\eta_1^{\text{booking}}(x_{f,i}^{\text{booking}}) = \mathbb{1}_{\{\text{MULTI}_{f,i}=1\}} \cdot \beta_{1,1}^{\text{booking}} + \text{CHNL}_{f,i} \cdot \boldsymbol{\beta}_{1,2}^{\text{booking}} + \\ p_{f,i} \cdot \beta_{1,3}^{\text{booking}} + m_1^{\text{booking}}(t_{f,i}),\end{aligned} \tag{5}$$

where $m_1^{\text{booking}}$ is a smooth unknown function.

Of particular interest is to quantify how the log relative price $p_{f,i}$ influences the probability to reserve a seat, for which we expect a negative coefficient $\beta_{1,3}^{\text{booking}}$. On the other hand, the effect of multiple passenger is expected to be positive, because we assume the passengers who travel together want to sit together, which can only be ensured through making an ASR.

The ASR Model 2 applies if customers do not reserve a seat at the time point of booking the ticket. They may do so at a later occasion. We model this by employing a discrete time-to-event model. To be specific, we again make use of a conditional logit model and set for $s \leq -2$:

$$\begin{aligned}&\mathbb{P}[y_{f,i}(s) = 1 | y_{f,i}(t) = 0 \text{ for } t_{f,i} \leq t < s; s \leq -2] = \text{logit}^{-1}[\eta_2^{\text{A}}(s)], \text{ where} \\ &\eta_2^{\text{A}}(s) = \beta_{2,0} + \eta_2^{\text{flight}}(x_f^{\text{flight}}) + \eta_2^{\text{booking}}\left[x_{f,i}^{\text{booking}}(s)\right].\end{aligned} \tag{6}$$

Again, superscript A refers to the ASR model and the first subscript 2 represents being the second part of this model. As with equation (4), the flight-specific effect $\eta_2^{\text{flight}}$ is

$$\eta_2^{\text{flight}}(x_f^{\text{flight}}) = \text{WDAY}_f \cdot \boldsymbol{\beta}_{2,1}^{\text{flight}} + m_2^{\text{flight}}(d_f). \tag{7}$$

The booking-specific effect $\eta_2^{\text{booking}}$ has a similar form to equation (5), but with the time point of booking $t_{f,i}$ replaced by the time point of seat reservation $s_{f,i}$, so that

$$\begin{aligned}\eta_2^{\text{booking}}\Big[x_{f,i}^{\text{booking}}(s)\Big] =& \mathbb{1}_{\{\text{MULTI}_{f,i}=1\}} \cdot \beta_{2,1}^{\text{booking}} + \text{CHNL}_{f,i} \cdot \boldsymbol{\beta}_{2,2}^{\text{booking}} + \\ & p_{f,i} \cdot \beta_{2,3}^{\text{booking}} + m_2^{\text{booking}}(s_{f,i}).\end{aligned} \tag{8}$$

Note that the log relative price $p_{f,i}$ is a time varying variable for routes $A$ and $B$, where the price for ASR has been varied. The smooth function $m_2^{\text{booking}}$ captures how the intensity for ASR booking changes with the days to departure, and can therefore be seen as a baseline intensity of an inhomogeneous process.

## 3.2 Seat Selection

If the passenger decides to reserve a seat, he or she needs to select one or more seats. This selection is modelled by a discrete choice model, where the choice set changes over time, because seats that have been reserved before are not any longer available. We make use of a multinomial logistic model with a varying consideration set. We can describe the seats with seat-specific covariates. But even if we do this, we observe from Figure 2 that some seats are preferred, although they are in principle similar to other seats. This indicates that we should not treat seats as homogeneous, but allow for seat-specific heterogeneity. Let $\mathcal{C}_f(t)$ be the consideration set of available seats on flight $f$ at time point $t$, where $t$

denotes the time to departure. We define with $\mathcal{S}_{f,i}$ the set of seats reserved on flight $f$ by passengers with booking $i$. Note that $\mathcal{S}_{f,i} = \emptyset$ if no ASR has been made (the most common case) and we do not model seat selection for these bookings. The set $\mathcal{S}_{f,i}$ contains a single element if a single seat has been reserved and multiple elements if multiple seats have been reserved at the same time. Note also that $\mathcal{S}_{f,i} \subseteq \mathcal{C}_f(\tilde{s}_{f,i})$, i.e. seats can only be reserved from available seats, where $\tilde{s}_{f,i}$ refers to the time point just prior to the seat reservation for booking $i$ on flight $f$. We model the selection of seats for which ASR is made as:

$$\mathbb{P}(\mathcal{S}_{f,i}) = \prod_{l \in \mathcal{S}_{f,i}} \frac{\exp(\eta_l^{\mathrm{S}})}{\sum_{r \in \mathcal{C}_f(\tilde{s}_{f,i})} \exp(\eta_r^{\mathrm{S}})}, \text{ where}$$
$$\eta_l^{\mathrm{S}} = \eta_0^{\mathrm{seat}}(x_l^{\mathrm{seat}}) + \eta_l^{\mathrm{seat}}(x_{f,i}^{\mathrm{booking}}), \tag{9}$$

and superscript S denotes the label Seat Selection Model. The linear components $\eta_l^{\mathrm{S}}$ decompose to effects which depend on seat-specific covariates, i.e.

$$\eta_0^{\mathrm{seat}}(x_l^{\mathrm{seat}}) = x_l^{\mathrm{seat}} \cdot \beta_0^{\mathrm{seat}} \tag{10}$$

and two seat random effects

$$\eta_l^{\mathrm{seat}}(x_{f,i}^{\mathrm{booking}}) = b_{0l} + \mathbb{1}_{\{\mathrm{MULTI}_{f,i}=1\}} b_{1l}, \tag{11}$$

that capture seat-specific heterogeneity. The random effect $b_{0l}$ is the seat preference, while the random effect $b_{1l}$ captures seat preferences that occur when seats are reserved for couples or groups; i.e. for multiple passengers.

# 4 Estimation

The Log-likelihood of all model components together can now be written as

$$\begin{aligned}
\sum_f \sum_i \Bigg\{ & y_{f,i}(s = t_{f,i}) \cdot \eta_1^{\mathrm{A}} - \log\left[1 + \exp\left(\eta_1^{\mathrm{A}}\right)\right] + && \text{(ASR Model 1)} \\
& \mathbb{1}_{\{y_{f,i}(s=t_{f,i})=0\}} \sum_{s=t_{f,i}+1}^{\min(s_{f,i},-2)} \left\{ y_{f,i}(s) \cdot \eta_2^{\mathrm{A}}(s) - \log\left\{1 + \exp\left[\eta_2^{\mathrm{A}}(s)\right]\right\}\right\} + && \text{(ASR Model 2)} \\
& \mathbb{1}_{\{\mathcal{S}_{f,i} \neq \emptyset\}} \sum_{l \in \mathcal{S}_{f,i}} \left\{ \eta_l^{\mathrm{S}} - \log\Big[ \sum_{r \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \exp\left(\eta_r^{\mathrm{S}}\right)\Big]\right\}\Bigg\}, && \text{(Seat Selection Model)}
\end{aligned} \tag{12}$$

where the linear predictors are given above. The maximizer of the separate likelihoods equals the joint likelihood because of the separability of parameters in different components. ASR Models 1 and 2 are both generalized additive models with both mixed effects and semi-parametric smoothing components, which can be estimated using maximum likelihood. We use the *mgcv* package [Wood, 2011] for fitting in R [R Core Team, 2016]. The Seat Selection Model is a discrete choice model with varying choice set and we outlined how to estimate it below. Note that

$$\mathbb{P}(\mathcal{S}_{f,i}) = \prod_{l \in \mathcal{S}_{f,i}} \frac{\exp(\boldsymbol{x}_l \boldsymbol{\beta}_0 + \boldsymbol{b}_{0l})}{\sum_{r \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \exp(\boldsymbol{x}_r \boldsymbol{\beta}_0 + \boldsymbol{b}_{0r})}. \tag{13}$$

The Log-likelihood for this component can be written as

$$l(\boldsymbol{\theta}) = \sum_f \sum_i \sum_{l \in \mathcal{S}_{f,i}} \left\{ \boldsymbol{Z}_{f,i,l}\boldsymbol{\theta} - \log\left[\sum_{r \in \mathcal{C}_f(\tilde{s}_{f,i})} \exp(\mathbf{Z}_{f,i,r}\boldsymbol{\theta})\right]\right\}, \tag{14}$$

where $\boldsymbol{Z}_{f,i,l} = (\boldsymbol{x}_l, \boldsymbol{u}_l, \mathbb{1}_{\{\mathrm{MULTI}_{f,i}=1\}}\boldsymbol{u}_l)$, $\boldsymbol{u}_l$ is the indicator vector having entry 1 at the $l$-th position and 0 otherwise and $\boldsymbol{\theta} = (\boldsymbol{\beta}_0^T, \boldsymbol{b}^T)^T$ refers to the parameter of this component, where $\boldsymbol{b} = (b_{01}, b_{02}, ..., b_{0L}, b_{11}, b_{12}, ..., b_{1L})$. We consider vector $\boldsymbol{b}$ as random and impose a prior distribution on $\boldsymbol{b}$, that is

$$\begin{pmatrix} b_{01} \\ b_{02} \\ \cdots \\ b_{0L} \end{pmatrix} \sim N(0, \sigma_0^2 \boldsymbol{I}_L) \text{ and } \begin{pmatrix} b_{11} \\ b_{12} \\ \cdots \\ b_{1L} \end{pmatrix} \sim N(0, \sigma_1^2 \boldsymbol{I}_L). \tag{15}$$

This leads to the marginal likelihood of a generalized mixed model:

$$l(\boldsymbol{\beta}_0, \boldsymbol{\sigma}^2) = \log \int \exp\left[l(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{D}\boldsymbol{\theta}\right] \cdot \left|\sigma_0^2 \boldsymbol{I}_L\right|^{-\frac{1}{2}} \cdot \left|\sigma_1^2 \boldsymbol{I}_L\right|^{-\frac{1}{2}} \, d\boldsymbol{b}. \tag{16}$$

where $\boldsymbol{D}$ is of block structure $\boldsymbol{D} = \text{blockdiag}(\mathbf{0}, \frac{1}{\sigma_0^2}\boldsymbol{I}_L, \frac{1}{\sigma_1^2}\boldsymbol{I}_L)$ and $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)$. We rewrite the $\exp(\cdot)$ component to a penalized likelihood

$$l_{\text{p}}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2) = l(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{D}\boldsymbol{\theta}. \tag{17}$$

Following and extending Breslow and Clayton [1993] we integrate out $\boldsymbol{b}$ and obtain with Laplace approximation the approximate likelihood

$$l(\boldsymbol{\beta}_0, \boldsymbol{\sigma}^2) \approx -\frac{L}{2}\sigma_0 - \frac{L}{2}\sigma_1 - \frac{1}{2}\log\left|\frac{\partial^2 l_{\text{p}}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\sigma}^2)}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}\right| + l_{\text{p}}(\tilde{\boldsymbol{\theta}}), \tag{18}$$

where $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\beta}_0, \tilde{\boldsymbol{b}})$ and $\tilde{\boldsymbol{b}}$ denote the solution to

$$\frac{\partial l_{\text{p}}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)}{\partial \boldsymbol{b}} = 0. \tag{19}$$

We thereby assume that the determinant in equation (18) depends only weakly on $\boldsymbol{\theta}$, see Breslow and Clayton [1993]. Note that $\tilde{\boldsymbol{b}}$ implicitly depend on $\boldsymbol{\beta}_0$ and $\boldsymbol{\sigma}^2$. With $\hat{\boldsymbol{\theta}}(\boldsymbol{\sigma}^2)$ we define the final estimate of

$$\frac{\partial l_{\mathrm{p}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\sigma}^2)}{\partial \boldsymbol{\theta}} = 0, \tag{20}$$

so that $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{b}})$. Following Kauermann et al. [2009] we can estimate $\sigma_0^2$ and $\sigma_1^2$ using the approximation

$$\hat{\sigma}_0^2 \approx \frac{\hat{\boldsymbol{b}}_0^T \hat{\boldsymbol{b}}_0}{df_0}, \tag{21}$$

with

$$df_0 = \mathrm{tr}\left\{ \left[ \left( \frac{\partial^2 l_{\mathrm{p}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\sigma}^2)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1} \left( \frac{\partial^2 l_{\mathrm{p}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\sigma}^2 = \infty)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) \right]_0 \right\}, \tag{22}$$

where the index 0 extract the columns belonging to $\boldsymbol{b}_0$, see also Schulze Waltrup and Kauermann [2017]. Note that (21) is not an analytic estimating equation since both sides depend on $\sigma_0^2$. An analogous formula holds for $\boldsymbol{b}_1$ and the estimation of $\sigma_1^2$. The maximiser $\hat{\boldsymbol{\theta}}$ of equation (17) can be seen as a parameter estimate $\hat{\boldsymbol{\beta}}_0$ as well as a posterior prediction for the random coefficients $\hat{\boldsymbol{b}}$. More details are given in the Appendix A.1.

# 5 Empirical Analysis

We estimate the models for each of the five routes separately. For conciseness, we report the results in detail for route $A$, and compare revenue implications for the five routes in Section 7.

## 5.1 Factors Affecting ASR when Booking

The parameter estimates for the ASR Model 1 fitted to the bookings on route $A$ during data Period I, are given in Table 3. We see for departure day of week (WDAY), there are significant positive coefficients for *Friday*, *Saturday* and *Sunday*, relative to the reference category *Monday*. The effects of different distribution channels (CHNL) show that the chance of making an ASR decreases when the ticket is not booked directly by the airline; especially when it was booked by an *OTA*. This is because many online agents do not show ASR as an option for ancillary service to the customer. Moreover, customers buying tickets from tour operators (which belongs to the category *Others*) often get an all-inclusive offer, which may not have ASR as an option. Besides the common cost-based argument described in Granados et al. [2012], this insight is another reason why airlines increasingly require a distribution cost charge for bookings made through global distribution systems, but not directly through airlines themselves. Furthermore, we observe a significant positive effect of multiple passengers (MULTI). This shows that customers who travel together (often couples and family members) are more likely to make an ASR to ensure sitting together. The effect of log relative price ($p$) can now be interpreted as the WTP, with an increase of 1% log relative price decreasing the chance of making an ASR by almost a half ($e^{-60.302 \times 0.01} = 0.547$). Except for WDAY, all estimated parametric effects are similar for the five routes $A$-$E$, as shown in Table 7 of Appendix A.2.

| Parameter | | Est. | SE |
|---|---|---|---|
| Intercept | | ***-0.413 | 0.057 |
| Departure day of week (WDAY, Ref.: *Monday*) | | | |
| | *Tuesday* | 0.023 | 0.061 |
| | *Wednesday* | 0.072 | 0.062 |
| | *Thursday* | -0.041 | 0.062 |
| | *Friday* | ***0.221 | 0.057 |
| | *Saturtay* | ***0.335 | 0.056 |
| | *Sunday* | ***0.233 | 0.056 |
| Multiple passengers (MULTI) | | ***0.220 | 0.037 |
| Distribution channel (CHNL, Ref.: *Direct*) | | | |
| | *Chains* | ***-0.817 | 0.041 |
| | *OTA* | ***-2.738 | 0.095 |
| | *Others* | ***-1.472 | 0.039 |
| Log relative price (p) | | ***-60.302 | 1.700 |

Table 3: Coefficient estimates for ASR Model 1 fitted to the bookings on route $A$ during data Period I. The reference categories are also reported for the categorical variables. Standard errors are also reported, and parameters that are significantly different from zero at the 1% level are denoted with three stars.

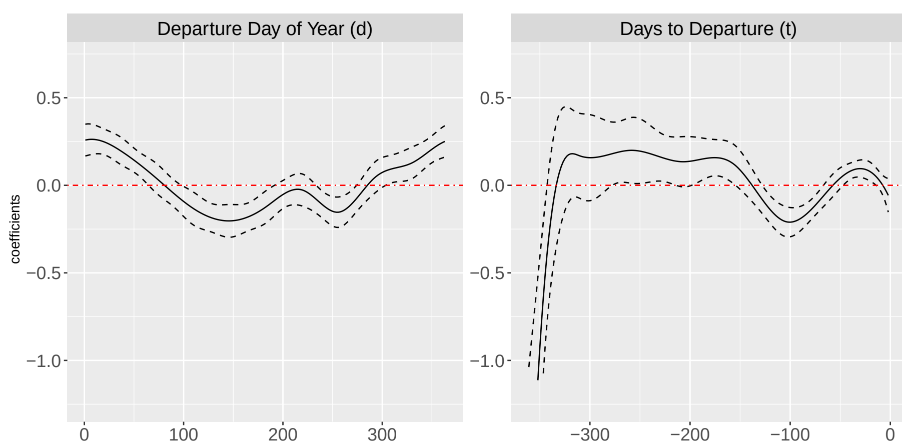


Figure 3: Estimated smooth effects for route $A$ in ASR Model 1. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(t)$ effect. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

The estimated smooth semi-parametric effects for route $A$ are plotted in Figure 3. There is a strong seasonal component (departure day of year) for the propensity to make an ASR at the same time as a ticket booking. This is mirrored for all other routes except route $C$; see Figure 8 - 11 of Appendix A.3. For the days to departure ($t$) we observe a oscillating effect with a strong propensity to make an ASR (joint with booking) between 50 and 10 days prior to departure. The same result is found for the other four routes in our study.

## 5.2 Factors Affecting ASR after Booking

The parameter estimates for the ASR Model 2 fitted to the bookings on route $A$ during data Period I are given in Table 4. The intercept is much smaller than that found in Table 3, which indicates that the chance of making an ASR is much smaller if customers did not make an ASR with their flight booking. In comparison to ASRs made at the time of booking, the effect of MULTI remains similar, but the effects of CHNL are weaker. This means that distribution channels play a less (negative) role here compared to the reference category *Direct*. This is because AirABC can reach out to customers after the tickets are booked and before the flight departs; e.g. via email advertisement. The baseline intensity of booking time $s$ of an ASR made after booking is illustrated on the right hand side of Figure 4. It shows that most ASRs are made in the month prior to departure. On the other hand, the seasonal effect is weaker than that for the ASR Model 1. The estimated effects — both categorical and smooth nonlinear — are similar for the other four routes, except for WDAY (see Table 8 and Figures 12 - 15 in Appendix A.2 and A.3).

| Parameter | | Est. | SE |
|---|---|---|---|
| Intercept | | ***-6.433 | 0.055 |
| Departure day of week (WDAY, Ref.: *Monday*) | | | |
| | *Tuesday* | 0.002 | 0.055 |
| | *Wednesday* | -0.048 | 0.056 |
| | *Thursday* | -0.029 | 0.055 |
| | *Friday* | ***0.266 | 0.052 |
| | *Saturday* | ***0.293 | 0.052 |
| | *Sunday* | ***0.180 | 0.054 |
| Multiple passengers (MULTI) | | ***0.265 | 0.030 |
| Distribution channel (CHNL, Ref.: *Direct*) | | | |
| | *Chains* | ***0.686 | 0.049 |
| | *OTA* | ***-0.205 | 0.060 |
| | *Others* | ***0.505 | 0.045 |
| Log relative price (p) | | ***-38.092 | 1.717 |

Table 4: Coefficient estimates for ASR Model 2 fitted to the bookings on route $A$ during data Period I. The reference categories are also reported for the categorical variables. Standard errors are also reported, and parameters that are significantly different from zero at the 1% level are denoted with three stars.

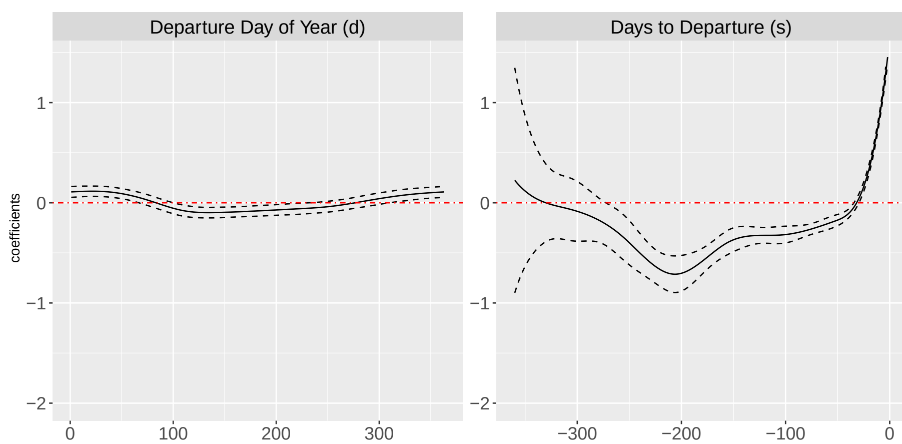


Figure 4: Estimated smooth effects for route $A$ in ASR Model 2. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(s)$ effect. The second can also be considered as the baseline purchasing intensity for this model. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

## 5.3 Factors Affecting Seat Selection

The parameter estimates for the Seat Selection Model are given in Table 5 and three observations can be made. First, the significant positive effects of *Aisle* and *Window* account for the preference of these seats compared to the reference category *Middle*. We name this behaviour as "middle seat avoiding". Secondly, the seats in the *Middle* section of the plane have significantly smaller probability of being reserved, than those in the *Front* of the plane. Moreover, this probability gets even smaller for the seats in the *Back* of the plane. This phenomenon can be described as "front seat preferring" and it is quite strong for intercontinental flights, presumably because passengers want to exit the plane quicker after landing at their destination. Last, the significant positive effects of special pair seats (SPEC) in the back of the plane can be explained by the generous extra space on the fuselage side of the seats, as illustrated in Figure 1 and discussed in some frequent flyer forums, e.g. Miller [2012] and Chan [2013].

| Parameter | | Est. | SE |
|---|---|---|---|
| Window, middle or aisle seat (WMA, Ref.: *Middle*) | | | |
| | *Aisle* | ***1.812 | 0.042 |
| | *Window* | ***1.614 | 0.045 |
| Plane section (SECT, Ref.: *Front*) | | | |
| | *Middle* | ***-0.862 | 0.033 |
| | *Back* | ***-2.133 | 0.085 |
| Special seat (SPEC) | | ***1.908 | 0.088 |

Table 5: Coefficient estimates for the Seat Selection Model fitted to the bookings on route $A$ during data Period I. The reference categories are also reported for the categorical variables. Standard errors are also reported, and parameters that are significantly different from zero at the 1% level are denoted with three stars.

Figure 5 gives the heat-map of the estimated seat random effects for ASRs made by single passengers $b_{1l}$, and multiple passengers $b_{0l}$. The latter shows that the "middle seat avoiding" effect is neutralised with multiple passenger bookings. Adding all effects


Figure 5: Heat-map of estimated random intercepts $b_{0l}$ (upper panel) and $b_{1l}$ (lower panel) in the Seat Selection Model for route $A$.

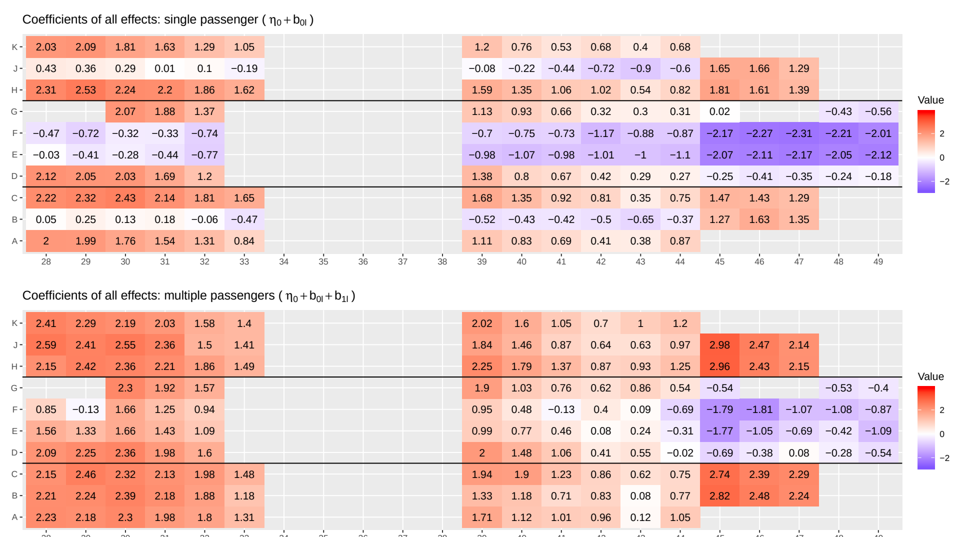


Figure 6: The overall heat-maps from the Seat Selection Model for route $A$. The upper panel gives the map for ASRs made by single passengers and the lower panel gives the map for ASRs made by multiple passengers.

($\eta_0^{\text{seat}}$, $b_{0l}$ and $b_{1l}$) together provides an overall heat-map for the popularity of seats in the aircraft. Figure 6 shows this for route $A$, and for a comparison with the routes in both data periods see Tables 9 - 10 and Figures 16 - 19 in Appendix A.2 and A.3. As a possible application to forecasting, one can weight the heat-maps of single and multiple passengers depending on the expected mixture of future arrival demand in order to get a generic heat-map of attractiveness of all the seats in the aircraft.

# 6 Model Validation

To validate the above model we investigate the out-of-sample predictive accuracy of fitted ASR Models 1 and 2. We fit the models to the data in Period I and then predict the expected revenue from ASR for the bookings in Period II by multiplying the predicted probability of an ASR for each booking by the price $p^{ASR}$. Note that both the point forecasts and prediction intervals can be computed. Our models break expected revenue into two parts: that arising from customers who made an ASR at the time of booking (ASR Model 1), and the revenue from customers who made an ASR afterwards (ASR Model 2). An expression for overall expected revenues is therefore:

$$\widehat{\text{REV}} = \sum_f \sum_i \left[ \hat{\mathbb{P}}^1_{f,i} \cdot p^{\text{ASR}}_{f,i} + (1 - \hat{\mathbb{P}}^1_{f,i}) \cdot \sum_{s=t_{f,i}+1}^{-2} \hat{\mathbb{P}}^2_{f,i} \cdot p^{\text{ASR}}_{f,i}(s) \right], \tag{23}$$

where $\hat{\mathbb{P}}^1_{f,i}$ and $\hat{\mathbb{P}}^2_{f,i}$ are the predicted probabilities from equation (1) and (2), respectively, and the summation is over all flights and bookings in the out-of-sample period. Note, for the expected revenues in ASR Model 2, we have to condition on the complementary probability $1 - \hat{\mathbb{P}}^1_{f,i}$, and sum over the purchasing probabilities up to two days before departure; i.e. exactly the last day before the check-in process begins and ASR becomes free.

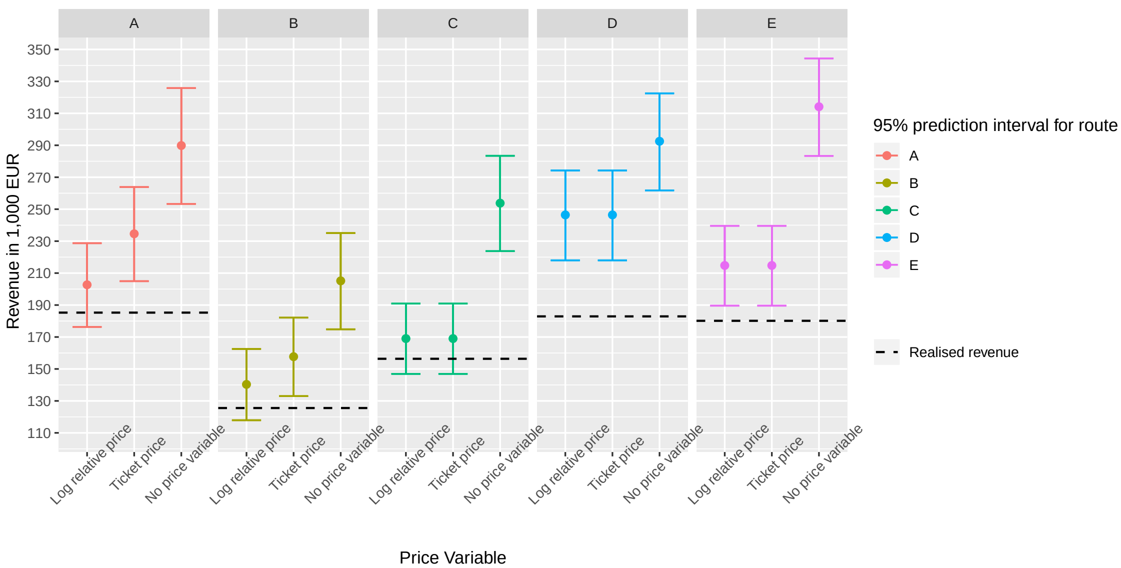


Figure 7: 95% prediction intervals of expected ASR revenue (vertical solid lines), and realised (horizontal dashed lines) ASR revenue for bookings in Period II.

Figure 7 plots the resulting predicted values of revenue for each of the five routes and the corresponding prediction variability based on the model. The realised revenues are drawn as dashed lines in each route. Furthermore, there are three kinds of expected revenue computations based on three different price variables, which are used to modify our model for fitting and predicting. The first uses log relative price $p = \log\left(\frac{p^{\mathrm{TKT}}+p^{\mathrm{ASR}}}{p^{\mathrm{TKT}}}\right)$ as the price variable. This is the variant introduced in this paper. The second uses the ticket price $p^{\mathrm{TKT}}$ as the price variable, while the third includes no price variable at all. For the latter case, the predictive interval of expected revenue does not cover the realised revenue for any of the five routes. Because the price of ASR $p^{\mathrm{ASR}}$ did not vary in Period II for routes $C$, $D$ and $E$, the predictions using the nominal and relative ticket prices are the same. However, on routes $A$ and $B$, the use of the log relative price instead of the nominal price greatly improves the accuracy of the revenue prediction. These results suggest that it is useful to use the logarithm of the price of ancillary services, relative to the ticket price, to estimate the WTP. This has the substantial advantage that it is not necessary to run pricing experiments for ancillary products to obtain variation in the nominal price. In particular, airlines can price the ancillary products dynamically relative to the paid ticket price by combining our ASR Models 1 and 2.

# 7 Economic Insights and Managerial Implications

It has been more than three decades since American Airlines [1987] described revenue management as the discipline "to maximize passenger revenue by selling the right seats to the right customers at the right time". The original definition has been extended to include at least two more important factors: with the right price and in a right combination by Cross [1997]. The key of this discipline is understanding customers' perception of product value, so that price can be varied in order to maximise revenue. Before ASR was introduced, airlines were not selling the specific seat(s), but one or other seat(s). It has become popular — including at AirABC — to sell legroom and standard ASR as an ancillary product. In this section we will discuss the benefits of extending this to allow for variable pricing based on the above models that use flight, booking and seat-specific covariates, compared to a single pricing policy for standard seats.

| Route: | A | B | C | D | E |
|---|---|---|---|---|---|
| ASR Model 1: | -60.302 | -50.668 | -72.656 | -54.140 | -83.431 |
| ASR Model 2: | -38.092 | -30.151 | -30.105 | -41.962 | -45.313 |

Table 6: Estimated effect of log relative price in all five routes in data Period I, which can be interpreted as price elasticities.

We first look further at the price elasticities for ASR using the log relative price $p$.

The negative coefficients of log relative price in ASR Models 1 and 2 can be interpreted as transformed relative price elasticities, and are summarized in Table 6, where the first column gives the effects already reported for route $A$ in Tables 3 and 4 above. These values can be used for predicting the demand for ASR. Before doing so, we stress an interesting insight: the price elasticity is weaker in ASR Model 2 than in ASR Model 1. That is, if an ASR is made after booking the ticket, passengers are less price sensitive. This appears to be a behavioral effect, with the price paid for the ticket having a declining effect over time on the subsequent decision of whether to make an ASR. Moreover, the price of a flight ticket can be considered as a reference price, as defined by Fibich et al. [2005], where its effect decreases with time.

A second insight is that the booking channel plays an important role. In particular, tickets that are not sold directly by AirABC, rarely are combined with an ASR at the time of booking. This suggests that management should endeavour to seek ways to increase ASR for tickets booked via these channels. The program, “New Distribution Capability”, recently launched by IATA is directed to meeting this objective (see Hoyles [2015]).

Finally, the third insight is that customers find that seats are heterogeneous. This is partially due to variations in seat-specific features, and partially due to heterogeneity in customer preferences. These are substantial as demonstrated by the heat-map in Figure 6. In this figure, the overall seat effects are depicted on the logarithmic scale. For example, ASR for seat 28A is approximately 20 times more attractive to single passengers than seat 43E ($e^{2-(-1)} = 20.09$). The strategic implication for pricing and revenue management is to give different prices to all seats for ASR according to their estimated value.

# 8 Discussion

Understanding customers' needs means that airlines can offer them the right product. Therefore, product differentiation has become one of the key success factors in many industries. The results of our analysis suggest that the ASR for standard seats — which are currently sold at the same price — are valued differently by customers. By adopting the characteristics of the seat heat-map, airlines have many possibilities to differentiate these products. For example, they could bundle bookings for multiple passengers with ASR, especially for those customers who seek to sit together in the special seats in rows $45 - 47$ in this cabin configuration. Another example is to give different prices to aisle or window seats, as well as for different plane sections. One result of doing so, is that booking data can be collected where the price per standard seat varies. Such data can be used to estimate our proposed ASR Model using the nominal ticket price, rather than the relative price. This will enable us to compute nominal WTP for each seat.

In practice, airlines have pursued two forms of price discrimination. The first involves adopting different prices for the same product through the use of booking classes and fare rules that cater to different customer segments in the market. The second is product differentiation through offering different products to different customers. For the new rising star of ancillary services, adopting nuanced price discrimination can make it a major driver of profitability. While we consider ASR in detail here, our work provides a framework for modelling customer preferences for other ancillary products using flight, booking and seat (or potentially other product) specific information. Particularly reasonable is the use of log relative prices, rather than nominal prices. This is because it exploits the variation in the already dynamically-varied ticket price to allow for estimation of the (relative) WTP or price elasticity, without requiring experimental pricing of ancillary products. One interesting area for future research is to explore optimal pricing of bundles of ancillary services. For this type of analysis, we suggest focusing on the data generating process as we have done in this paper, but also considering the relationships between products in a bundle.

# References

American Airlines (1987). Annual report. Technical report, Forth Worth Airport, Dallas.

Balcombe, K., Fraser, I., and Harris, L. (2009). Consumer willingness to pay for in-flight service and comfort levels: A choice experiment. *Journal of Air Transport Management*, 15(5):221–226.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Caussade, S. and Hess, S. (2009). An investigation into air travellers willingness to pay for ancillary service attributes within a branded fare context. In *European Transport Conference, 2009*. Citeseer.

Chan, C. (2013). The most popular seats on an airplane are at the very back. https://gizmodo.com/5994456/the-most-popular-seats-on-an-airplane-are-at-the-very-back, last accessed on 2018-07-22.

Cross, R. (1997). *Revenue Management: Hard-Core Tactics for Market Domination.* Crown Business, New York, reprint edition.

Daft, J. and Albers, S. (2012). A profitability analysis of low-cost long-haul flight operations. *Journal of Air Transport Management*, 19(10):49–54.

Davidson, R. and Mackinnon, J. (1993). *Estimation and Inference in Econometrics.* Oxford University Press, 1 edition.

Davidson, R. and Mackinnon, J. (1999). *Econometric Theory and Methods.* Oxford University Press, 1 edition.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b -splines and penalties. *Statistical Science*, 11(2):89–121.

Espino, R., Martín, J. C., and Román, C. (2008). Analyzing the effect of preference heterogeneity on willingness to pay for improving service quality in an airline choice context. *Transportation Research Part E: Logistics and Transportation Review*, 44(4):593–606.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, NY, 2 edition.

Fibich, G., Arieh, G., and Lowengart, O. (2005). The dynamics of price elasticity of demand in the presence of reference price effects. *Journal of the Academy of Marketing Science*, 33(1):66–78.

Garrow, L. A., Hotle, S., and Mumbower, S. (2012). Assessment of product debundling trends in the us airline industry: Customer service and public policy implications. *Transportation Research Part A: Policy and Practice*, 46(2):255–268.

Granados, N., Kauffman, R. J., Lai, H., and Lin, H. C. (2012). À La Carte Pricing and Price Elasticity of Demand in Air Travel. *Decision Support Systems*, 53(2):381–394.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall/CRC, 1 edition.

Hoyles, Y. (2015). New Distribution Capability (NDC): Together let’s build airline retailing. Technical report, IATA.ORG.

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

Kremser, F., Guenzkofer, F., Sedlmeier, C., Sabbah, O., and Bengler, K. (2012). Aircraft seating comfort: the influence of seat pitch on passengers’ well-being. *Work*, 41(Supplement 1):4936–4942.

Lee, D. and Luengo-Prado, M. J. (2004). Are passengers willing to pay more for additional legroom? *Journal of Air Transport Management*, 10(6):377–383.

Leslie, P. (2004). Price discrimination in broadway theater. *RAND Journal of Economics*, 35(3):520–541.

Miller, S. (2012). Where to sit (and not to sit!) on the lufthansa 747-8i. https://blog.wandr.me/2012/06/where-to-sit-and-not-to-sit-on-the-lufthansa-747-8i/, last accessed on 2018-07-22.

Mumbower, S., Garrow, L. A., and Newman, J. P. (2015). Investigating airline customer's premium coach seat purchases and implications for optimal pricing strategies. *Transportation Research Part A: Policy and Practice*, 73:53–69.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47:3–13.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Schulze Waltrup, L. and Kauermann, G. (2017). Smooth expectiles for panel data using penalized splines. *Statistics and Computing*, 27(1):271–282.

Train, K. (2009). *Discrete Choice Methods With Simulation*. Cambridge University Press, 2 edition.

Tutz, G. and Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer International Publishing Switzerland, 1 edition.

Veeraraghavan, S. and Vaidyanathan, R. (2012). Measuring seat value in stadiums and theaters. *Production and Operations Management*, 21(1):49–68.

Vink, P., Bazley, C., Kamp, I., and Blok, M. (2012). Possibilities to improve the aircraft interior comfort experience. *Applied ergonomics*, 43(2):354–359.

Wardman, M. and Murphy, P. (2015). Passenger’s valuations of train seating layout, position and occupancy. *Transportation Research Part A: Policy and Practice*, 74:222–238.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R.* CRC Press, Taylor & Francis, 2 edition.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach.* South-Western, 5 edition.

# A Appendix

## A.1 Derivations and algorithm for estimation

To maximize the penalized likelihood $l_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, two derivations are requested.

**Score Function:**

$$
\begin{aligned}
\boldsymbol{s}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_f \sum_i \sum_{l \in \mathcal{S}_{f,i}} \{\boldsymbol{Z}_{f,i,l}^T - \frac{1}{\sum_{r \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})} \cdot \sum_{r \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} [\exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta}) \cdot \boldsymbol{Z}_{f,i,r}^T]\} \\
&= \sum_f \sum_i \sum_{l \in \mathcal{S}_{f,i}} \{\boldsymbol{Z}_{f,i,l}^T - \sum_{r \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \frac{\exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})}{\sum_{q \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})} \cdot \boldsymbol{Z}_{f,i,r}^T\} \\
&= \sum_f \sum_i \sum_{l \in \mathcal{S}_{f,i}} \{\boldsymbol{Z}_{f,i,l}^T - \sum_{r \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \pi_r \cdot \boldsymbol{Z}_{f,i,r}^T\} \qquad (24)
\end{aligned}
$$

where $\pi_r = \frac{\exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})}{\sum_{q \in \mathcal{C}_{f,i}(\tilde{s}_{f,i})} \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})}$ and with penalization: $\boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{s}(\boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{\theta}$ where $\boldsymbol{\lambda} = (\frac{1}{\sigma_0^2}, \frac{1}{\sigma_1^2})$.

**Fisher Information:**

$$
\begin{aligned}
\boldsymbol{F}(\boldsymbol{\theta}) &= -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \\
&= \sum_f \sum_i \sum_r \left\{ \frac{\sum_q \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta}) \cdot \exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})\boldsymbol{Z}_{f,i,r}^T \cdot \boldsymbol{Z}_{f,i,r} - \exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})\boldsymbol{Z}_{f,i,r}^T \cdot \sum_q [\exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta}) \cdot \boldsymbol{Z}_{f,i,q}]}{\sum_q \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta}) \cdot \sum_q \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})} \right\} \\
&= \sum_f \sum_i \sum_r \left\{ \frac{\exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})\boldsymbol{Z}_{f,i,r}^T \boldsymbol{Z}_{f,i,r}}{\sum_q \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})} - \frac{\exp(\boldsymbol{Z}_{f,i,r}\boldsymbol{\theta})\boldsymbol{Z}_{f,i,r}^T \cdot \sum_q [\exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})\boldsymbol{Z}_{f,i,q}]}{\sum_q \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta}) \cdot \sum_q \exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})} \right\} \\
&= \sum_f \sum_i \sum_r \left\{ \pi_r \boldsymbol{Z}_{f,i,r}^T \boldsymbol{Z}_{f,i,r} - \pi_r \boldsymbol{Z}_{f,i,r}^T \cdot \sum_q \frac{\exp(\boldsymbol{Z}_{f,i,q}\boldsymbol{\theta})}{\sum_j \exp(\boldsymbol{Z}_{f,i,j}\boldsymbol{\theta})} \cdot \boldsymbol{Z}_{f,i,q} \right\} \\
&= \sum_f \sum_i \sum_r \left\{ \pi_r \boldsymbol{Z}_{f,i,r}^T \boldsymbol{Z}_{f,i,r} - \pi_r \boldsymbol{Z}_{f,i,r}^T \cdot \sum_q \pi_q \cdot \boldsymbol{Z}_{f,i,q} \right\} \\
&= \sum_f \sum_i \sum_r \left\{ \pi_r \boldsymbol{Z}_{f,i,r}^T (\boldsymbol{Z}_{f,i,r} - \sum_q \pi_q \boldsymbol{Z}_{f,i,q}) \right\}. \qquad (25)
\end{aligned}
$$

With Penalization: $\boldsymbol{F}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{D}$

We use Newton-Raphson-Algorithm

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{F}^{-1}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\lambda}^{(k)})\ \boldsymbol{s}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\lambda}^{(k)}) \tag{26}$$

to estimate the parameters and for the asymptotic variance of estimation we use the sandwich estimator

$$\widehat{\mathrm{se}(\boldsymbol{\theta})} = \sqrt{\mathrm{diag}[\boldsymbol{F}^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})\ \boldsymbol{F}(\boldsymbol{\theta})\ \boldsymbol{F}^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})]}, \tag{27}$$

where

$$\frac{\hat{\theta} - \theta}{\widehat{se(\theta)}} \sim t(n - p', 0). \tag{28}$$

## A.2 Additional Tables

| Route | A | | B | | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| Intercept | ***-0.413 | 0.057 | ***-0.275 | 0.056 | 0.090 | 0.052 | **0.150 | 0.051 | 0.055 | 0.040 |
| Departure day of week (WDAY, Ref.: *Monday*) | | | | | | | | | | |
| *Tuesday* | 0.023 | 0.061 | *-0.138 | 0.059 | 0.074 | 0.057 | 0.007 | 0.052 | **-0.131 | 0.043 |
| *Wednesday* | 0.072 | 0.062 | ***-0.201 | 0.059 | -0.035 | 0.057 | -0.011 | 0.052 | ***-0.150 | 0.045 |
| *Thursday* | -0.041 | 0.062 | ***-0.369 | 0.061 | -0.035 | 0.056 | 0.047 | 0.051 | ***-0.129 | 0.044 |
| *Friday* | ***0.221 | 0.057 | ***-0.272 | 0.057 | 0.060 | 0.053 | 0.023 | 0.050 | -0.063 | 0.043 |
| *Saturday* | ***0.335 | 0.056 | ***-0.264 | 0.056 | 0.076 | 0.050 | -0.041 | 0.048 | **0.117 | 0.039 |
| *Sunday* | ***0.233 | 0.056 | -0.891 | 0.055 | ***0.173 | 0.049 | 0.013 | 0.048 | -0.049 | 0.040 |
| Multiple passengers (MULTI) | ***0.220 | 0.037 | ***0.140 | 0.040 | ***0.249 | 0.038 | ***0.564 | 0.030 | **0.097 | 0.033 |
| Distribution channel (CHNL, Ref.: *Direct*) | | | | | | | | | | |
| *Chains* | ***-0.817 | 0.041 | ***-0.375 | 0.042 | ***-0.811 | 0.039 | ***-0.840 | 0.040 | ***-0.282 | 0.031 |
| *OTA* | ***-2.738 | 0.095 | ***-2.635 | 0.109 | ***-2.686 | 0.103 | ***-2.593 | 0.095 | ***-2.260 | 0.092 |
| *Others* | ***-1.472 | 0.039 | ***-1.176 | 0.043 | ***-1.581 | 0.039 | ***-0.479 | 0.037 | ***-0.738 | 0.036 |
| Log relative price (p) | ***-60.302 | 1.700 | ***-50.668 | 1.285 | ***-72.656 | 1.356 | ***-54.140 | 1.086 | ***-83.431 | 1.184 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Table 7: Coefficient estimates for ASR Model 1 fitted to the bookings on each route during data Period I.

| Route | A | | B | | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| Intercept | ***-6.433 | 0.065 | ***-6.685 | 0.081 | ***-6.470 | 0.067 | ***5.938 | 0.079 | ***-5.726 | 0.055 |
| Departure day of week (WDAY, Ref.: *Monday*) | | | | | | | | | | |
| *Tuesday* | 0.002 | 0.055 | 0.016 | 0.075 | -0.080 | 0.063 | 0.060 | 0.077 | ***-0.201 | 0.055 |
| *Wednesday* | -0.048 | 0.056 | *0.176 | 0.073 | 0.033 | 0.060 | 0.091 | 0.075 | -0.062 | 0.055 |
| *Thursday* | -0.029 | 0.055 | *0.162 | 0.073 | 0.069 | 0.059 | 0.051 | 0.074 | **-0.154 | 0.054 |
| *Friday* | ***0.266 | 0.052 | ***0.247 | 0.068 | ***0.287 | 0.056 | 0.043 | 0.072 | -0.098 | 0.054 |
| *Saturday* | ***0.293 | 0.052 | *0.150 | 0.069 | ***0.204 | 0.055 | 0.018 | 0.069 | -0.014 | 0.052 |
| *Sunday* | ***0.180 | 0.054 | 0.091 | 0.073 | ***0.225 | 0.057 | 0.115 | 0.069 | 0.047 | 0.052 |
| Multiple passengers (MULTI) | ***0.265 | 0.030 | **0.110 | 0.041 | ***0.271 | 0.034 | ***0.283 | 0.040 | ***0.257 | 0.034 |
| Distribution channel (CHNL, Ref.: *Direct*) | | | | | | | | | | |
| *Chains* | ***0.686 | 0.049 | ***0.802 | 0.056 | ***0.793 | 0.051 | ***0.306 | 0.056 | ***0.413 | 0.039 |
| *OTA* | ***-0.205 | 0.060 | *-0.161 | 0.069 | ***-0.072 | 0.065 | ***-0.231 | 0.069 | ***-0.552 | 0.064 |
| *Others* | ***0.505 | 0.045 | ***0.690 | 0.050 | ***0.686 | 0.045 | ***0.328 | 0.053 | ***0.226 | 0.043 |
| Log relative price (p) | ***-38.092 | 1.717 | ***-30.151 | 1.660 | ***-30.105 | 1.494 | ***-41.962 | 1.553 | ***-45.313 | 1.379 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Table 8: Coefficient estimates for ASR Model 2 fitted to the bookings on each route during data Period I.

| Route | A | | B | | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| Window, middle or aisle seat (Ref.: *Middle*) | | | | | | | | | | |
| *Aisle* | ***1.813 | 0.042 | ***2.198 | 0.053 | ***2.235 | 0.045 | ***1.625 | 0.047 | ***2.227 | 0.035 |
| *Window* | ***1.614 | 0.045 | ***1.876 | 0.056 | ***1.866 | 0.047 | ***1.273 | 0.052 | ***2.002 | 0.037 |
| Plane section (Ref.: *Front*) | | | | | | | | | | |
| *Middle* | ***-0.862 | 0.033 | ***-1.161 | 0.039 | ***-1.015 | 0.033 | ***-1.329 | 0.037 | ***-0.959 | 0.025 |
| *Back* | ***-2.133 | 0.085 | ***-2.542 | 0.104 | ***-2.246 | 0.081 | ***-2.864 | 0.111 | ***-1.846 | 0.057 |
| Special seat | ***1.908 | 0.088 | ***1.814 | 0.109 | ***1.765 | 0.085 | *0.306 | 0.137 | ***1.226 | 0.061 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Table 9: Coefficient estimates for the Seat Selection Model fitted to the bookings on each route during data Period I.

| Route | A | | B | | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est. | SE | Est. | SE | Est. | SE | Est. | SE | Est. | SE |
| Window, middle or aisle seat (Ref.: *Middle*) | | | | | | | | | | |
| *Aisle* | ***1.798 | 0.052 | ***1.805 | 0.052 | ***2.262 | 0.050 | ***1.690 | 0.051 | ***2.323 | 0.045 |
| *Window* | ***1.612 | 0.056 | ***1.581 | 0.057 | ***1.983 | 0.053 | ***1.216 | 0.057 | ***2.120 | 0.047 |
| Plane Section (Ref.: *Front*) | | | | | | | | | | |
| *Middle* | ***-1.191 | 0.041 | ***-1.061 | 0.041 | ***-1.049 | 0.036 | ***-1.284 | 0.041 | ***-1.008 | 0.033 |
| *Back* | ***-1.967 | 0.093 | ***-1.786 | 0.089 | ***-1.913 | 0.082 | ***-2.587 | 0.109 | ***-1.842 | 0.073 |
| Special seats | ***1.803 | 0.098 | ***1.406 | 0.096 | ***1.372 | 0.087 | -0.177 | 0.150 | ***1.253 | 0.077 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Table 10: Coefficient estimates for the Seat Selection Model fitted to the bookings on each route during data Period II.

## A.3 Additional Figures

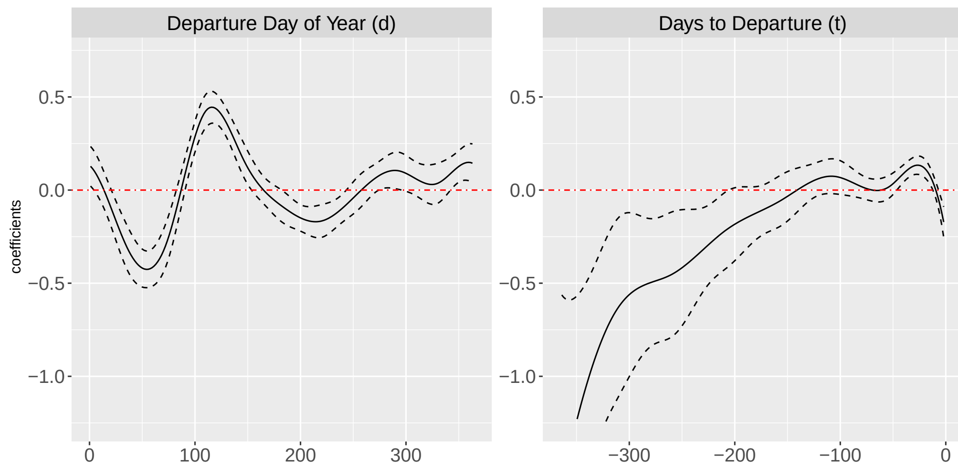


Figure 8: Estimated smooth effects for route $B$ in ASR Model 1. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(t)$ effect. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

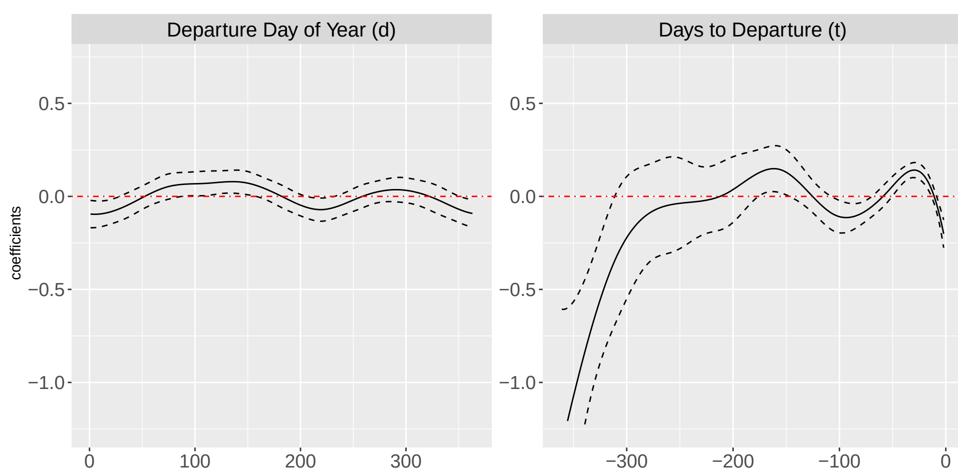


Figure 9: Estimated smooth effects for route $C$ in ASR Model 1. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(t)$ effect. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

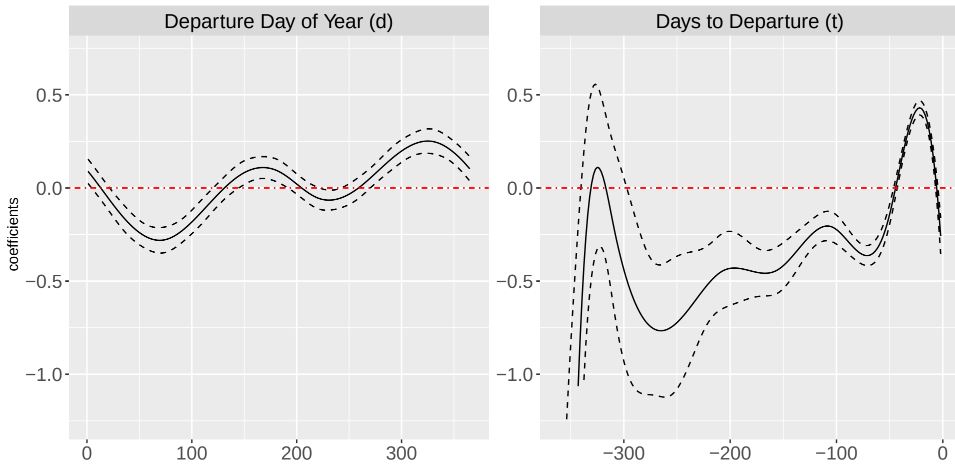


Figure 10: Estimated smooth effects for route $D$ in ASR Model 1. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(t)$ effect. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

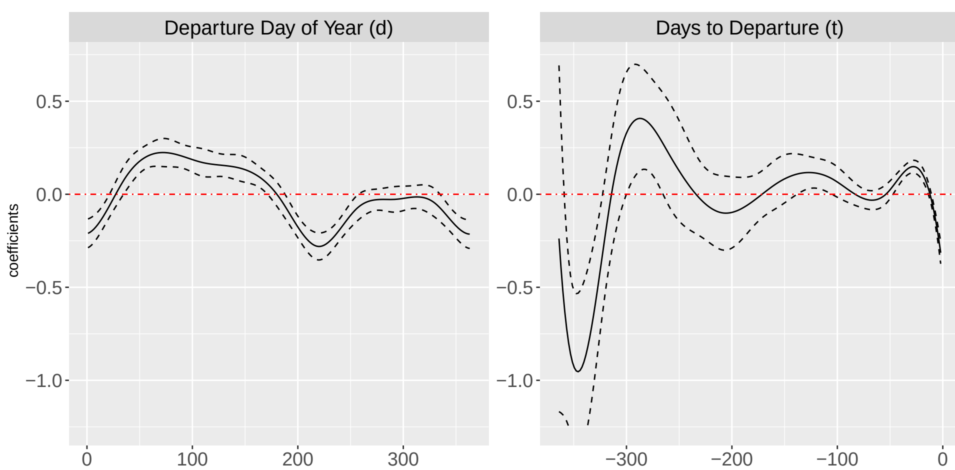


Figure 11: Estimated smooth effects for route $E$ in ASR Model 1. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(t)$ effect. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

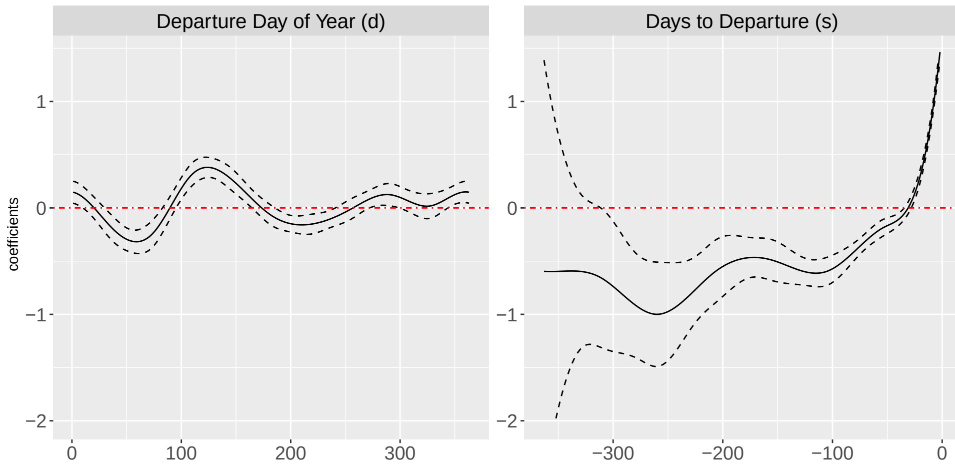


Figure 12: Estimated smooth effects for route $B$ in ASR Model 2. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(s)$ effect. The second can also be considered as the baseline purchasing intensity for this model. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

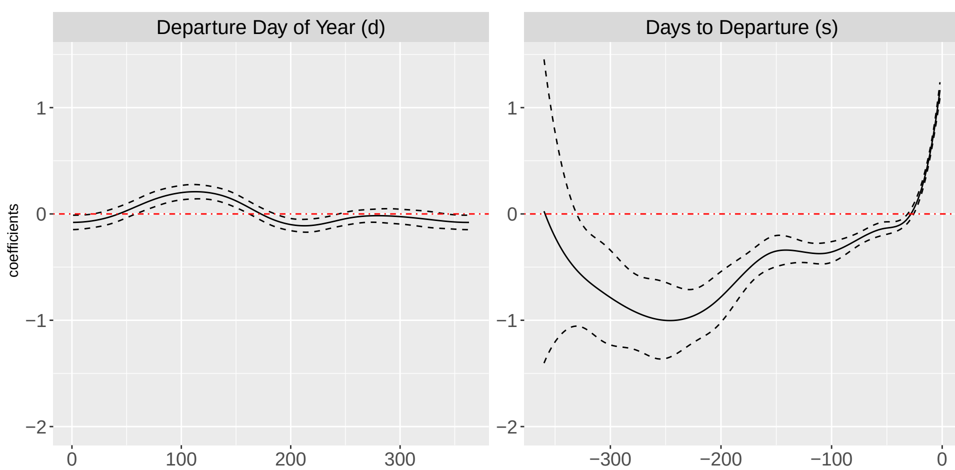


Figure 13: Estimated smooth effects for route $C$ in ASR Model 2. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(s)$ effect. The second can also be considered as the baseline purchasing intensity for this model. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

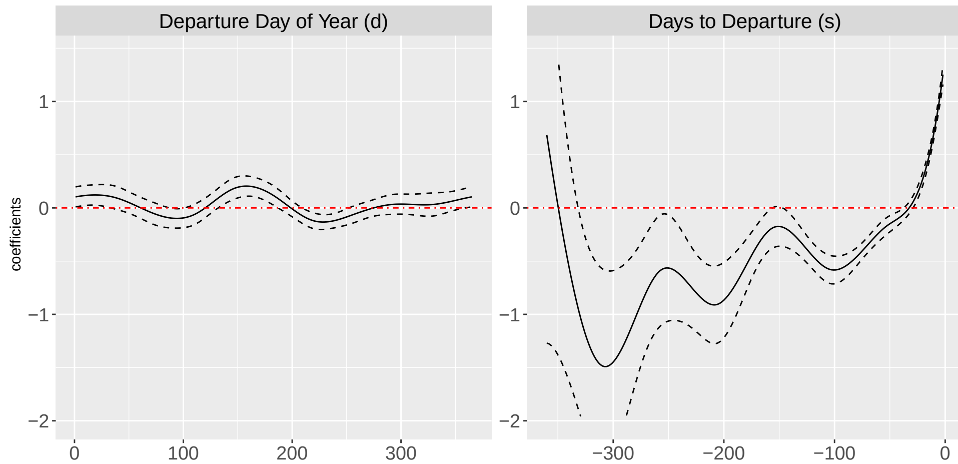


Figure 14: Estimated smooth effects for route $D$ in ASR Model 2. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(s)$ effect. The second can also be considered as the baseline purchasing intensity for this model. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

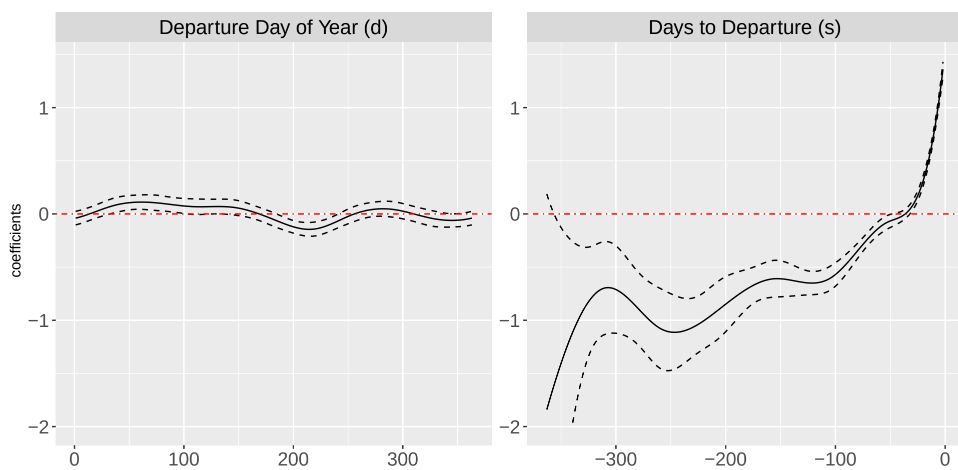


Figure 15: Estimated smooth effects for route $E$ in ASR Model 2. The left panel gives the departure day of year $m_2(d)$ effect, and the right panel gives the days to departure $m_2(s)$ effect. The second can also be considered as the baseline purchasing intensity for this model. The point estimates of the functions are given by the solid line, while the dashed lines are 95% confidence bands.

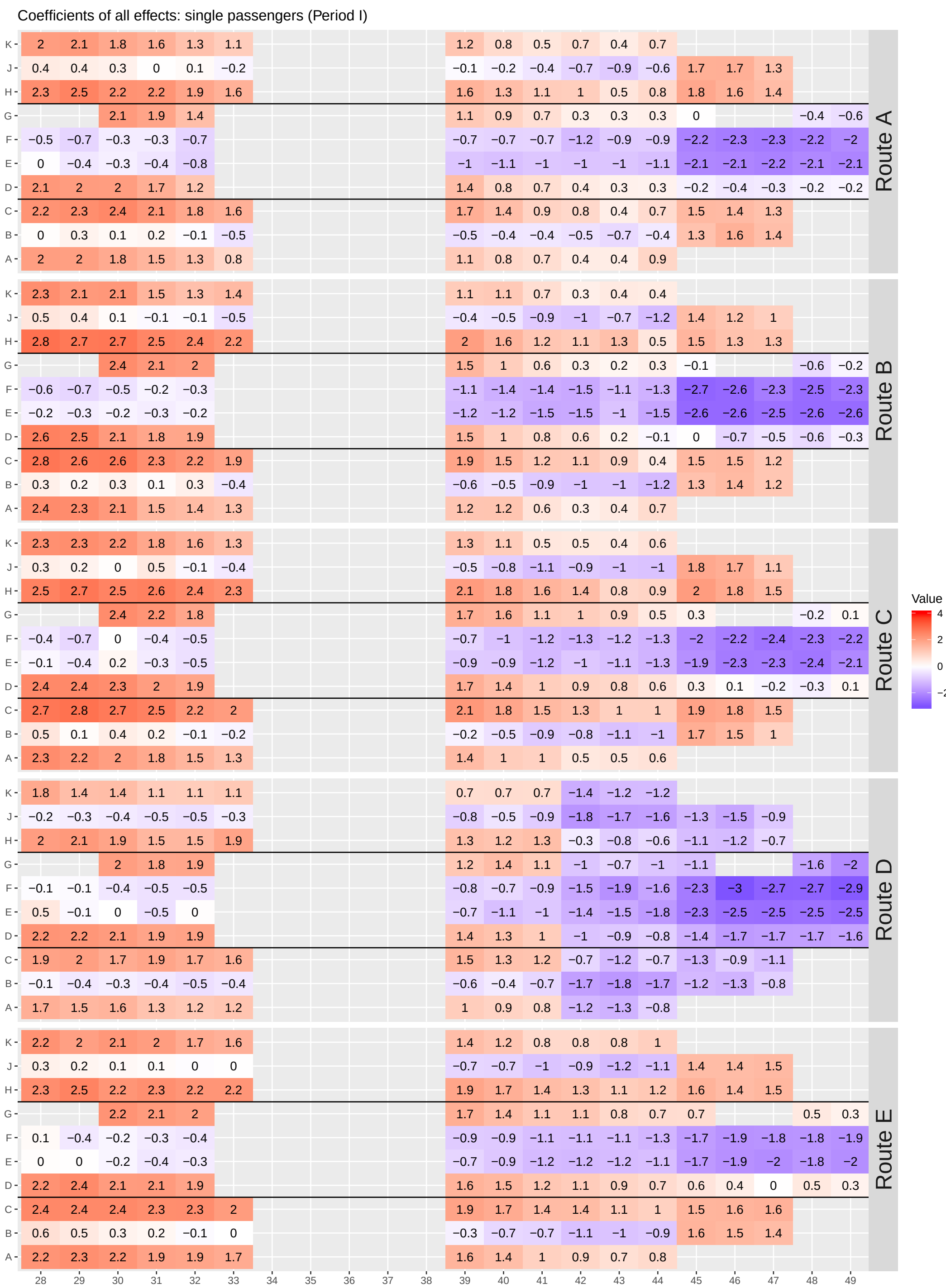


Figure 16: The overall heat-maps from the Seat Selection Model for single passengers in each route during data Period I.

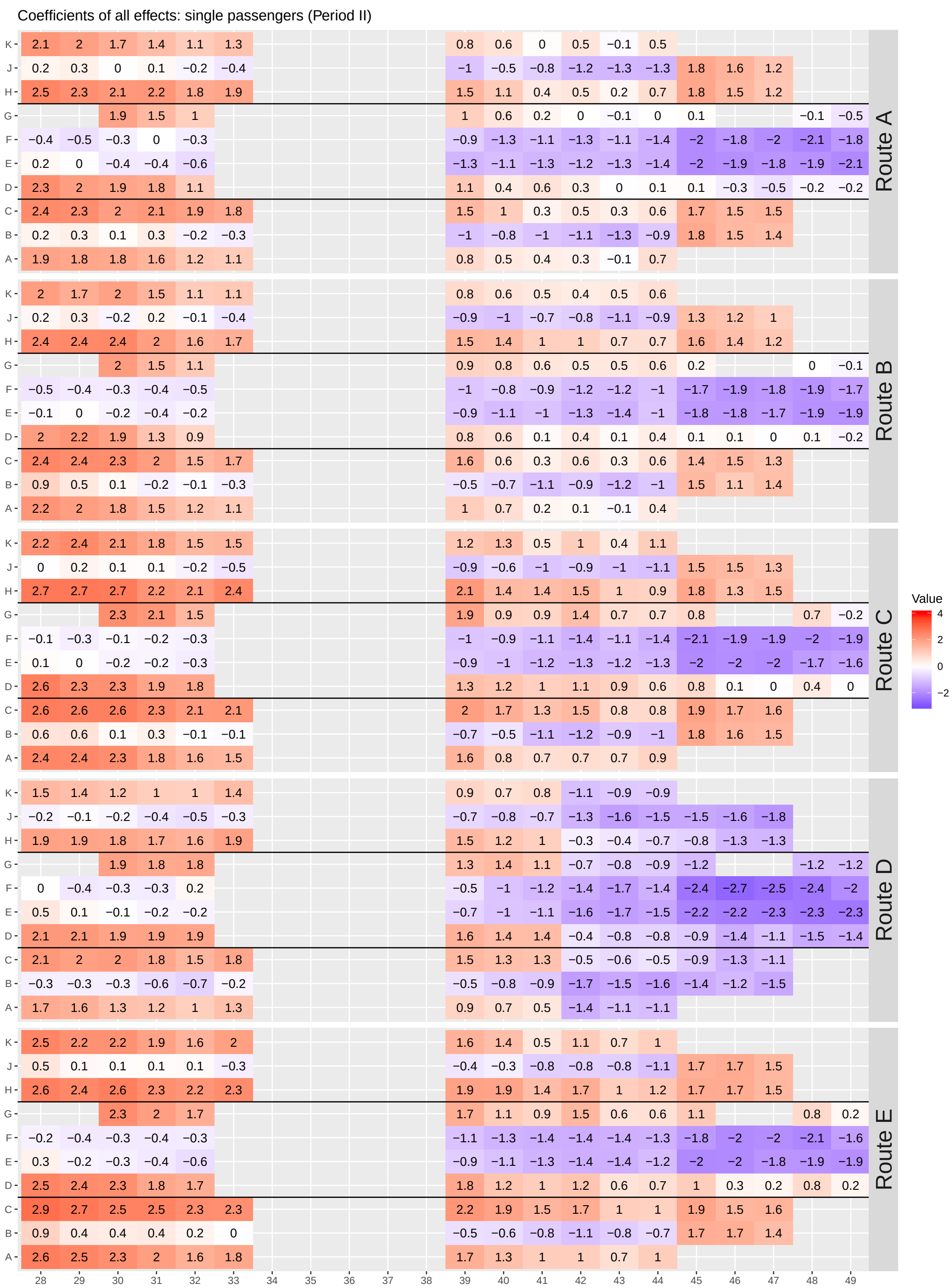


Figure 17: The overall heat-maps from the Seat Selection Model for single passengers in each route during data Period II.

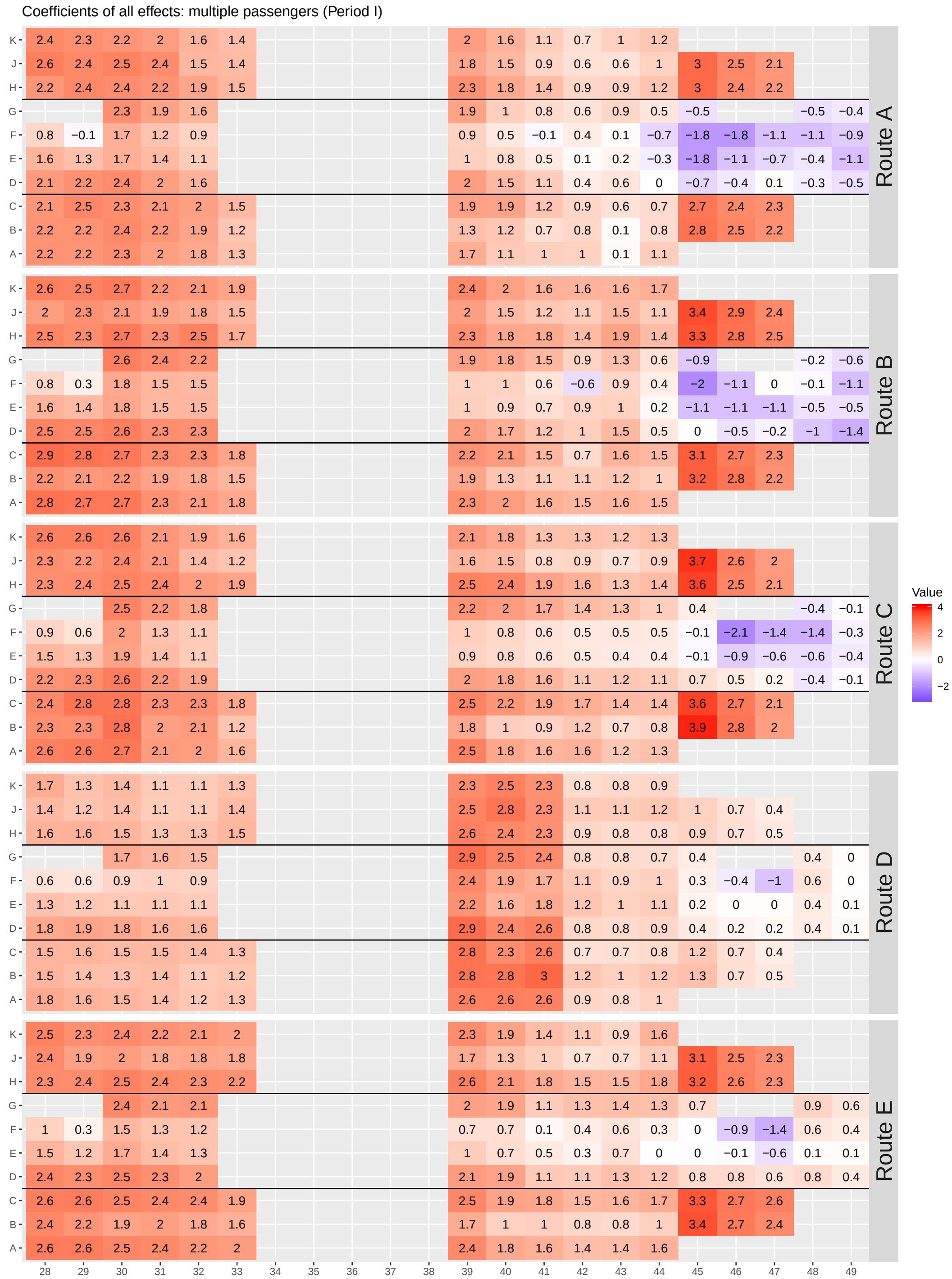


Figure 18: The overall heat-maps from the Seat Selection Model for multiple passengers in each route during data Period I.

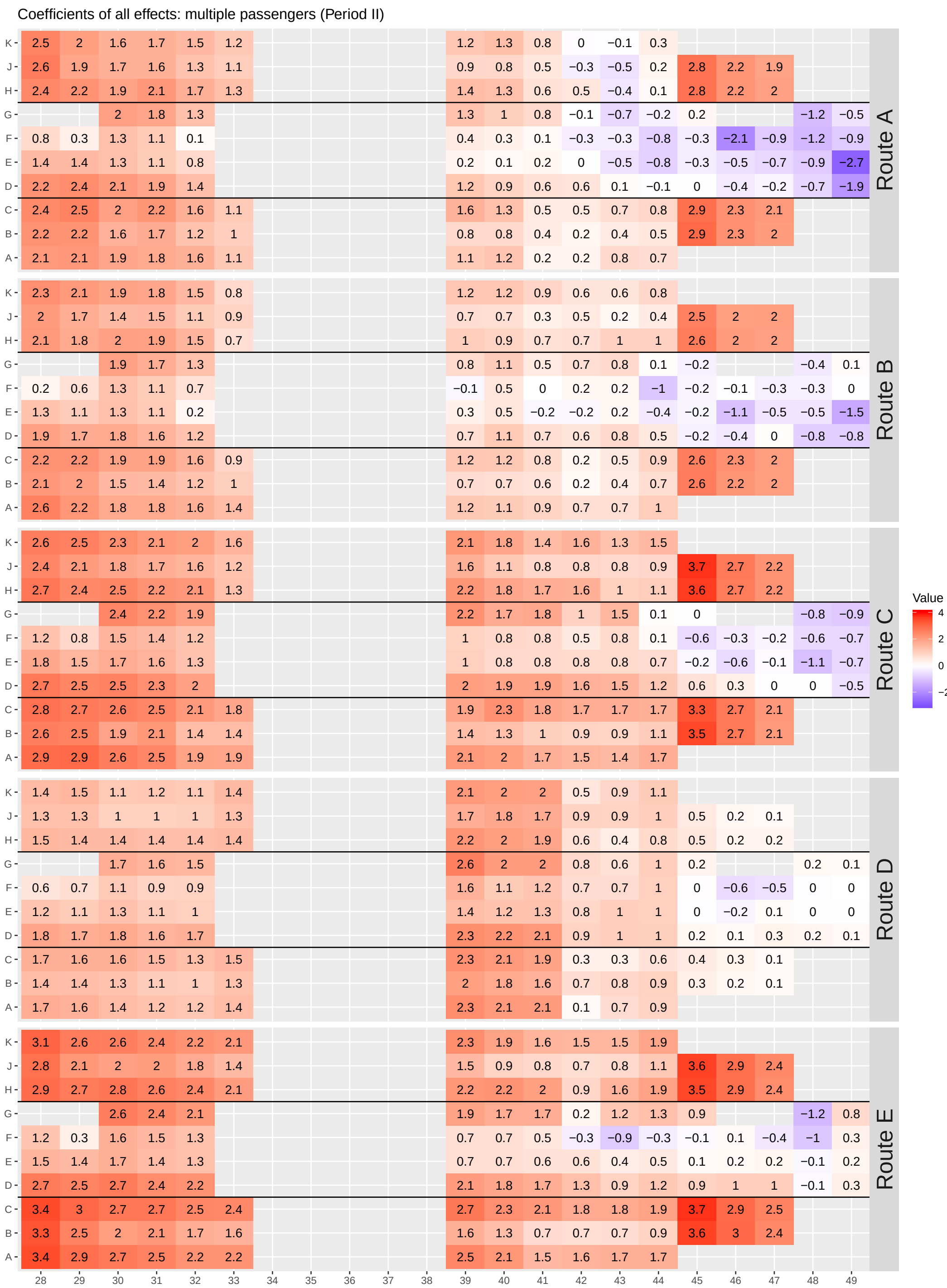


Figure 19: The overall heat-maps from the Seat Selection Model for multiple passengers in each route during data Period II.

# Chapter 3

# Bundle pricing with aggregated market data

Chapter 3 presents a applied research article, which concerns the bundle pricing problem. Considering the nature of customers' choices between purchasing no ancillaries, single ancillary items ("a la carte") or multiple ancillaries in a bundle; a discrete choice model based on multinomial distribution is proposed.

**Contributing article:**

Shao, S., Kauermann, G. (2019) Understanding price elasticity for airline ancillary services. *Journal of Revenue and Pricing Management*, available at: `http://link.springer.com/article/10.1057/s41272-018-00177-z`.

**Copyright:**



**Author contributions:**

Shuai Shao conceived the research question and was responsible for the experimental design and implementation. He drafted the manuscript including examples and visualisations. Göran Kauermann critically revised and contributed to the manuscript by giving valuable inputs on modelling and interpretation approaches, as well as proofreading.

Journal of Revenue and Pricing Management
https://doi.org/10.1057/s41272-018-00177-z

RESEARCH ARTICLE

# Understanding price elasticity for airline ancillary services

Shuai Shao[1] · Göran Kauermann[1]

Received: 1 November 2018 / Accepted: 20 December 2018


**Abstract**

Recently, the general trend in the airline industry has been to generate ancillary revenue by offering additional services. Instead of completely separating ancillary services from tickets as optional components, most of the traditional airlines offer the so-called branded fares which bundle some of the ancillary components to an inclusive fare preventing a possible negative impact on the customers' perception and brand image (mixed bundling). For instance, seat reservation and baggage transportation are often already included in the default fare. In this study, we analyse data to evaluate different bundle-pricing policies within the mixed bundling context. We use statistical regression methods to infer individual behaviour by analysing aggregated data on market level from a major European airline. We tackle the question of how to optimally price bundled fares. With the General Data Protection Regulation in place today, such high-level models which only require aggregated market data and no individual personal data are becoming more relevant for business analytics. We demonstrate how aggregate data still allow to investigate individual behaviour and our data analysis reveals the existence and variability of price elasticity. The results can help companies to segment their markets based on price elasticity and optimise their ancillary offerings accordingly.

**Keywords** Ancillary revenue · Branded fares · Bundle pricing · Mixed bundling · Price elasticity · Pricing policy evaluation

## Introduction

In the last years, ancillary revenue has become increasingly relevant in driving profits in the airline industry. Fierce competition has driven traditional airlines to consider unbundling their services in a way that low-cost-carriers (LCCs) do. LCCs are particularly known to offer additional services like airport check-in, luggage handling or seat reservation only "a la carte" (pure components), meaning each component requires to pay for it. In the past, such services were included in the fare by traditional airlines. However, these airlines lately follow the route of LCCs which allows them to keep the airfares competitive and generate additional revenues by selling ancillary services. Still, some classical airlines also re-bundled their unbundled components as the so-called branded fares. That is, they offer a bundle of pre-set ancillary services for a slightly reduced price. This approach can be seen as mixed bundling where the ancillary products are offered simultaneously "a la carte" and in a pre-set bundle. The key reason for this strategy is to prevent possible negative impact on customer perception and brand image while they can still use the cheapest branded fare as a tactic to compete with the LCCs.

The research on (mixed) bundling has been discussed extensively. Adams and Yellen (1976) were among the first to study product bundling. Venkatesh and Mahajan (2009) provided an extensive review on further literature regarding the design and pricing of bundles focusing on diverse rationales, guidelines and approaches. Focusing on revenue management, Gallego and Stefanescu (2012) discussed applications of bundles used in different industries. Ratliff and Gallego (2013) introduced a decision support framework for bundled product design applying to airline branded fares. They used survey data (stated preference) and pointed out that, due to different needs of the customer, the value derived from branded fares varies between market segments. The task of pricing and revenue management is to portray the values of ancillary up-sell (both "a la carte" and bundle) adequately. As listed by Bockelie and

✉ Shuai Shao
shuai.shao@stat.uni-muenchen.de

1 Department of Statistics, LMU München, Ludwigstraße 33, 80539 Munich, Germany

Published online: 21 January 2019

Belobaba (2017), numerous studies on the effect of ancillary fees have been carried out, yet very few of them based on historical ancillary sales (revealed preference). A reason for this might be that historical data on ancillary sales have not been maintained by many airlines for a long time. With the General Data Protection Regulation (GDPR) which came into effect on 25th May 2018, the collection of data on an individual/personal basis has become more restrictive, see e.g. Steppe (2017) and Tikkinen-Piri et al. (2018). We therefore propose a statistical approach using a novel interpretation of a standard statistical model which is based on aggregated market data. Though being applied to aggregate data, the model (for some extend) allows for the interpretation of individual behaviour.

In this paper, we analyse data from a major European airline, which wishes to remain anonymous. As result, we refer to it as "AirABC" throughout this paper. The data were collected on 40 different markets, meaning routes differentiated by origin and destination. These markets have been experimentally divided in three groups and in each of the groups a different price policy was applied. Our task is to access and quantify the customers' price elasticity with respect to flight-related ancillary services. This task can be tackled in different ways. A simple approach is to look at the changes regarding total ancillary revenue in the three market groups and make use of a traditional Analysis of Variance (ANOVA) approach, see e.g. Montgomery (2012). Applied to our data, we find no significant changes in the three groups with different pricing policies. Apparently, this is a rather simplistic approach which neglects the detailed structure in the data. We therefore employ a more elaborated statistical model which takes details of the three market groups into account. Thus, a more sophisticated data analysis is carried out in this paper which allows for deeper insight into customers' preference and behaviour.

We pursue a statistical model that looks at the different possible choices for a customer, that is buying bundles of ancillaries or booking ancillaries separately or no ancillary purchase at all. This leads to a discrete choice model, see e.g. Train (2009). We make use of a multinomial distribution as described in Fahrmeir and Tutz (2001) and take market specific covariates into account. In fact, for each market, we have the distribution of the customers' choices before and after the experimental price changes. We use the prior distribution of the customers' ancillary selection as covariate describing the market's overall potential of selling ancillary services. This leads to a multinomial model with multivariate covariates. The setting resembles the data analytic approaches carried out with the so-called ecological inference, see e.g. Achen and Shively (1995), King (1997), Wakefield (2004) or Klima et al. (2016). Ecological inference thereby describes statistical approaches to draw inference from aggregated data about individual behaviour. A common application area for such models is to examine voter migration based on aggregate data on the distribution of votes for a set of political parties. We adopt this approach to our setting in order to illuminate the customers' behaviour, where the price for ancillary bundles is changed. The hypothetical idea behind this is that we question how a "market specific standard customer" with a preference for ancillary services changes his/her preference if the price for the bundle changes.

The paper is organised as follows: In section "Data", we give a detailed description of the data and show the results of an ad hoc analysis of variance. Section "Models and results" describes the discrete choice model and its specification. Section "Interpretation and strategic recommendation" provides an interpretation of the resulting parameters using ideas of ecological inference. This approach also allows to give a strategic recommendation. For model evaluation, we compare each applied pricing policy with the expected and realised ancillary revenues in section "Policy evaluation and economical implication", before we conclude our results in section "Discussion".

## Data

The data that underlie our analysis were collected in the years 2016 and 2017. We look at bookings and their corresponding purchased ancillary services. All bookings can be divided into three mutually exclusive and exhaustive categories: Tickets with no ancillaries booked (N), those with ancillaries booked "a la carte" (A) and tickets booked with ancillaries offered in a bundle (B). In total, this leads to 145,243 tickets which were booked in 2016 (data period I) and 126,971 tickets booked in 2017 (data period II). In data period II, the prices for ancillary services have been changed, where in particular, the bundled fare was modified. This was done in an experimental setting, i.e. the routes have been chosen randomly. Since in data period I, the same pricing policy was applied on all 40 markets, we use these data as baseline, which expresses the market specific heterogeneity with respect to ancillary sales. This heterogeneity is visualised in Fig. 1 exemplary for 10 destinations which all departed from the same origin.

The percentage of bookings with purchased ancillary services (both in bundle and "a la carte") in data period I is given as greyscale. As can be seen from the figure, a homogeneous pricing policy may not justify the heterogeneous demand on ancillary services. AirABC therefore pursued a designed experimental setting of pricing policy in data period II. This is indexed by $p \in \mathcal{P} = \{``+", ``0", ``-"\}$: while "0" stands for no change in the bundle price, "+" indicates the bundle price was increased and "−" represents that the bundle price

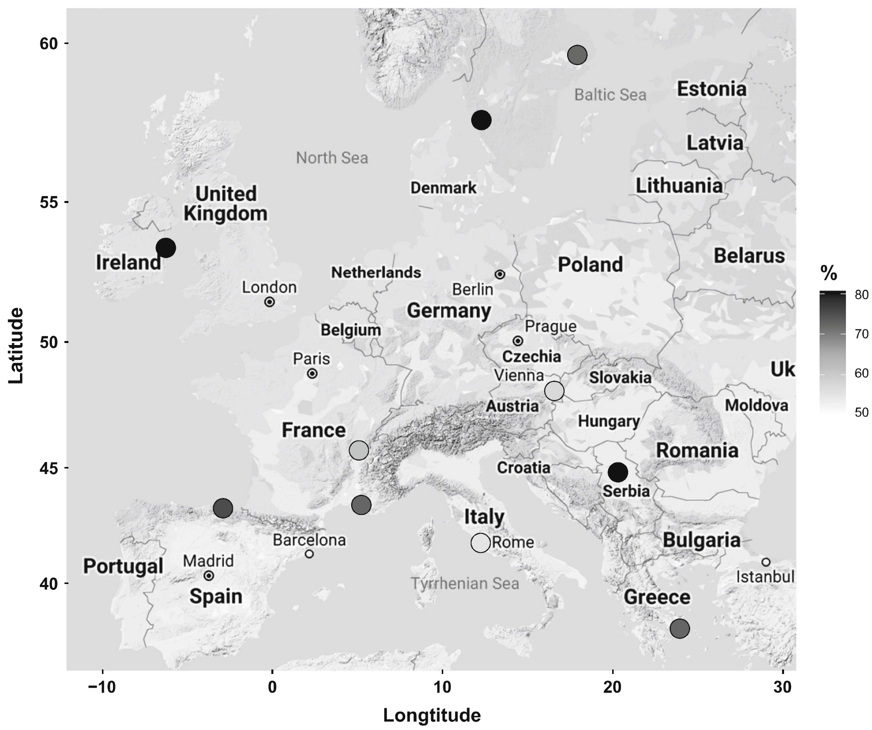


**Fig. 1** Variable percentages of ancillaries (both "a la carte" and bundle) purchased during data period I in 10 selected markets

was decreased. Note that the amounts of decrease and increase for the bundle price were the same and the prices for each ancillary item "a la carte" did not change between data period I and II. The experimental setting divides the 40 markets $\mathcal{M}$ with respect to the applied pricing policy in data period II which we denote as $\mathcal{M} = \cup_{p \in \mathcal{P}} \mathcal{M}^p$. A breakdown of the data to the different markets is given in Table 1.

The task is now to explore possible differences in the three market groups, which due to the experimental approach could be statistically accredited to price elasticity. A common practice in pricing and revenue management is a comparison in performances in both sales periods, mostly with the focus on changes in the revenue. We pursue this point here for demonstrative purposes and define the relative change in revenue through

$$z_m = \frac{\text{Ancillary revenue in data period II} - \text{Ancillary revenue in data period I}}{\text{Ancillary revenue in data period I}}\%, \tag{1}$$

where $m \in \mathcal{M}$. Putting these numbers into an ANOVA leads to $z_{p,m} = \mu + \beta_p + \varepsilon_{p,m}$ with side constraint $\sum_{p \in \mathcal{P}} \beta_p = 0$. We can now test the null hypothesis $H_0 : \beta_p = 0$ which, however, leads to no significant effect. The corresponding output is provided in Table 2.

**Table 1** Tables of absolute numbers of booking and marginal proportions in each group of markets under respective pricing policy ("+", "0", and "−") in data period I and II

| | No ancillaries | "A la carte" | Bundle | Total |
|---|---|---|---|---|
| $\mathcal{M}^{\text{“}+\text{”}}$, data period | | | | |
| I | 7814 (30%) | 3011 (12%) | 15,096 (58%) | 25,921 |
| II | 10,798 (38%) | 4273 (15%) | 13,035 (47%) | 28,006 |
| $\mathcal{M}^{\text{“}0\text{”}}$, data period | | | | |
| I | 25,578 (34%) | 8608 (11%) | 40,713 (54%) | 74,899 |
| II | 21,907 (35%) | 6408 (10%) | 33,442 (54%) | 61,757 |
| $\mathcal{M}^{\text{“}-\text{”}}$, data period | | | | |
| I | 18,147 (41%) | 5353 (12%) | 20,923 (47%) | 44,423 |
| II | 15,288 (41%) | 3507 (9%) | 18,413 (49%) | 37,208 |

**Table 2** ANOVA-Table of changes in ancillary revenue ($z_m$). Under this common practice, no significant effect among pricing policies could be determined

| | Degrees of freedom | Sum of squares | Mean squares | $F$ value | $P$ value |
|---|---|---|---|---|---|
| Price change ("+","0","−") | 3 | 5863 | 1954 | 1.47 | 0.24 |
| Residuals | 37 | 9178 | 1329 | | |

Such comparison is, as we will see, too simplistic since the detailed dynamics of up-sell behaviour is neglected. We therefore look again at Table 1. The data show the aggregated bookings decomposed by the ancillaries. Looking at the proportions in the table, we see a small variation in the margin proportions in the reference group $\mathcal{M}^{\text{“0”}}$, which is due to natural (price-independent) variation of demand in the markets. Moreover, we see evident differences between the pricing policies, e.g. percentage-wise in group $\mathcal{M}^{\text{“+”}}$ more customers chose "a la carte" and less customers chose bundles in data period II. Vice versa, in group $\mathcal{M}^{\text{“−”}}$ less customers chose "a la carte" and more customers chose the bundle in data period II. This gives a hint that price elasticity exists, which, however, needs to be accessed and quantified with more advanced statistical models.

## Models and results

In this section, we describe how the proposed model allows to infer individual behaviour from aggregated data. We therefore define the multinominal random variable

$$Y \in \{\mathrm{N} := \textbf{N}\text{o ancillaries}, \mathrm{A} := \text{“}\textbf{A}\text{ la carte”}, \mathrm{B} := \textbf{B}\text{undle}\} =: \mathcal{C} \tag{2}$$

giving the customer's choice from his/her choice-set $\mathcal{C}$ for a single booking. We model the probability $\mathbb{P}(Y = c) = \pi_c$ , where $c \in \mathcal{C}$. The distribution is allowed to depend on a set of covariates $x$, which is included in the model through

$$\pi_c = \mathbb{P}(Y = c|x) = \frac{\exp(\eta_c)}{\sum_{\tilde{c}} \exp(\eta_{\tilde{c}})}, \text{ with } \eta_c = \boldsymbol{x}\boldsymbol{\beta}_c, \tag{3}$$

where $x$ may also contain the intercept. Equation (3) corresponds to a multinomial regression model with category-specific parameter $\beta_c$. We denote with $\boldsymbol{y}_m = (y_{m,\mathrm{N}}, y_{m,\mathrm{A}}, y_{m,\mathrm{B}})$ the cumulated number of bookings in three categories/choices in market $m$ and assume that $\boldsymbol{y}_m$ follows a multinomial distribution

$$f(\boldsymbol{y}_m|\boldsymbol{\pi}) = \frac{T_m!}{y_{m,\mathrm{N}}!y_{m,\mathrm{A}}!y_{m,\mathrm{B}}!}\prod_{c\in\mathcal{C}} \pi_c^{y_{m,c}} m, \text{ with } T_m = \sum_{c\in\mathcal{C}} y_{m,c}. \tag{4}$$

Note that $T_m$ is the total number of bookings in market $m$ in the current data period. We consider bookings done in the experimental data period II as response variable $\boldsymbol{y}$ and take data from data period I as covariate variable $\boldsymbol{x}$. Apparently, we do not have covariates on an individual level but we do use market-specific covariates, namely the bookings carried out in data period I. To be specific, we denote with $\boldsymbol{x}_m = (x_{m,\mathrm{N}}, x_{m,\mathrm{A}}, x_{m,\mathrm{B}})$ the total number of bookings in data period I, which are transformed to proportion through

$$\tilde{\boldsymbol{x}}_m = \frac{\boldsymbol{x}_m}{T_{(x)m}} = \left(\frac{x_{m,\mathrm{N}}}{T_{(x)m}}, \frac{x_{m,\mathrm{A}}}{T_{(x)m}}, \frac{x_{m,\mathrm{B}}}{T_{(x)m}}\right), \tag{5}$$

where $T_{(x)m} = \sum_{r\in\mathcal{C}} x_{m,r}$. The covariate $\tilde{\boldsymbol{x}}_m$ gives the market specific covariate vector, which is the market-specific distribution of ancillary booking proportions prior to the experimental setting.

The central focus of our analysis lies on the quantification of the different pricing policies. That is, we want to investigate whether the three experimental groups differ and if so how. We do this by fitting an interaction model, meaning that all parameters may change between the experimental groups. This leads to the final model

$$\mathbb{P}(Y_{m,c}|\tilde{\boldsymbol{x}}_m, p) = \frac{\exp(\tilde{\boldsymbol{x}}_m^p \boldsymbol{\beta}_c^p)}{\sum_{\tilde{c}} \exp(\tilde{\boldsymbol{x}}_m^p \boldsymbol{\beta}_{\tilde{c}}^p)}. \tag{6}$$

The coefficients in Eq. (6) can be estimated using maximum likelihood. We use the *VGAM* package (Yee 2010) for fitting in R (R Core Team 2018). The estimates are listed in Table 3 including their standard errors and 95% confidence intervals.

Note that $\boldsymbol{\beta}_c^p = (\beta_{\mathrm{N},c}^p, \beta_{\mathrm{A},c}^p, \beta_{\mathrm{B},c}^p)$ and we take $c = \mathrm{N}$ as the reference category for response variable $y$. The interpretation of the parameters is generally clumsy. We therefore visualise the parameter estimates $\hat{\beta}_{r,c}^p$ for $r \in \{\mathrm{N}, \mathrm{A}, \mathrm{B}\}$ and $c \in \{\mathrm{N}, \mathrm{A}, \mathrm{B}\}$ in Fig. 2.

Except $\beta_{\mathrm{A,B}}^{\text{“+”}}$, all effects are significant within each group of markets. Moreover, comparing the fitted coefficients between the groups, we see that effects in markets under pricing policy "+" differ more from the "0" markets compared to the "−" markets. This gives us a hint that the price elasticity is not symmetric, bearing in mind that the amount of increased and decreased up-sell prices was the same. From Fig. 2, we also see that the estimates of $\beta_{\mathrm{N,A}}^p$ differ in the three experimental groups. Overall, all significant effects of $p$ are in parameter estimates $\hat{\beta}_{r,c}^p$ where $r \neq c$. This means that certain dynamics between the

**Table 3** Estimated coefficients, standard error and 95% confidence intervals of parameter $\beta^{p}_{\mathrm{r,c}}$ in three models

| p: | "+" | | | "0" | | | "−" | | |
|---|---|---|---|---|---|---|---|---|---|
| | Est. | SE | 95% CI | Est. | SE | 95% CI | Est. | SE | 95% CI |
| $\boldsymbol{\beta}^{p}_{\mathrm{N,A}}$ | − 5.12*** | 0.23 | [− 5.56; − 4.67] | − 3.84*** | 0.12 | [− 4.07; − 3.61] | − 3.03*** | 0.11 | [− 3.25; − 2.81] |
| $\boldsymbol{\beta}^{p}_{\mathrm{N,B}}$ | − 2.84*** | 0.15 | [− 3.14; − 2.53] | − 2.17*** | 0.07 | [− 2.30; − 2.03] | − 2.12*** | 0.06 | [− 2.25; − 2.00] |
| $\boldsymbol{\beta}^{p}_{\mathrm{A,A}}$ | 9.02*** | 0.94 | [7.17; 10.87] | 6.82*** | 0.38 | [6.08; 7.57] | 6.06*** | 0.40 | [5.27; 6.85] |
| $\boldsymbol{\beta}^{p}_{\mathrm{A,B}}$ | 0.22 | 0.57 | [− 0.89; 1.33] | 1.26*** | 0.23 | [0.81; 1.70] | 2.26*** | 0.24 | [1.80; 2.72] |
| $\boldsymbol{\beta}^{p}_{\mathrm{B,A}}$ | − 0.86*** | 0.21 | [− 1.27; − 0.45] | 1.31*** | 0.09 | [− 1.49; 1.13] | − 2.04*** | 0.12 | [− 2.26; − 1.81] |
| $\boldsymbol{\beta}^{p}_{\mathrm{B,B}}$ | 1.70*** | 0.14 | [1.43; 1.96] | 1.87*** | 0.05 | [1.76; 4.98] | 1.64*** | 0.06 | [1.52; 1.77] |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

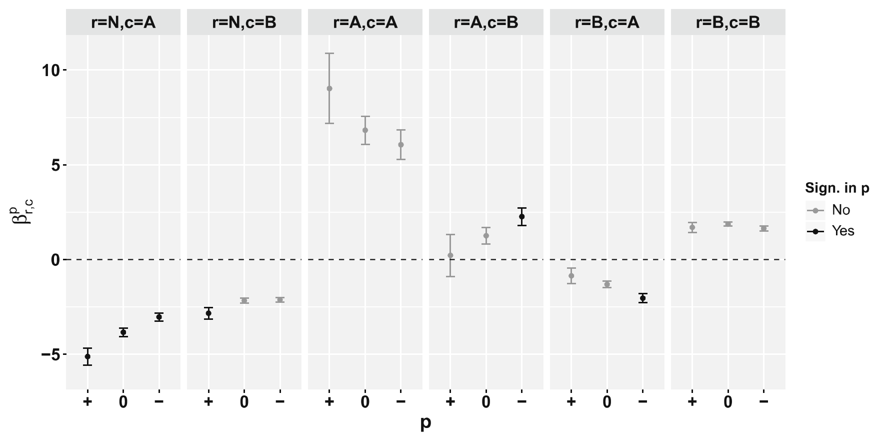


**Fig. 2** Estimated coefficients and their 95% confidence intervals. We compare the pricing policy $p$ in each parameter $\beta^{p}_{r,c}$: If a confidence interval does not overlap with others, the effect of $p$ is significant and the interval is drawn in black

preferred choices take place. We will access and quantify these dynamics in the next section.

## Interpretation and strategic recommendation

Models with multinomial responses often suffer from the large number of coefficients which impede interpretation. Moreover, due to the non-linear link function, it is difficult to provide a simple interpretation e.g. in terms of probability which is easily understandable for the end users in the industry. In this section, we introduce a different way to interpret the fitted models following a pricing and revenue management perspective. The intention is to exemplify the parameters through individual behaviour. Let therefore $T_{(x)m}$ remain the total number of bookings in data period I and $T_m$ is the total number of bookings in data period II, as defined previously. This implies that $\widehat{\mathbb{P}}(Y_m = c|\tilde{\boldsymbol{x}}_m) \cdot T_m$ will provide the predicted number of bookings in category $c$ in data period II, given the market-specific covariates $\tilde{\boldsymbol{x}}_m$. The idea is now to change $\tilde{\boldsymbol{x}}_m$ slightly to explore how changes in $\tilde{\boldsymbol{x}}_m$ will lead to changes in $\widehat{\mathbb{P}}(Y_m = c|\tilde{\boldsymbol{x}}_m) \cdot T_m$. To do so, we assume a single hypothetical additional booking in data period I. We start with an additional booking with no ancillaries booked, that is we increase $x_{m,\mathrm{N}}$ by one and keep $x_{m,\mathrm{A}}$ and $x_{m,\mathrm{B}}$ unchanged. This leads to the new covariate value $\tilde{\boldsymbol{x}}_{m|\mathrm{N}}$ defined through

$$\tilde{\boldsymbol{x}}_{m|\mathrm{N}} = \left( \frac{x_{m,\mathrm{N}} + 1}{T_{(x)m} + 1}, \frac{x_{m,\mathrm{A}}}{T_{(x)m} + 1}, \frac{x_{m,\mathrm{B}}}{T_{(x)m} + 1} \right). \tag{7}$$

We then also assume that we have an additional hypothetical booking in data period II. That is we take $\tilde{\boldsymbol{x}}_{m|\mathrm{N}}$ as input and calculate $\hat{\mathbb{P}}\big(Y_m = c|\tilde{\boldsymbol{x}}_{m|\mathrm{N}}\big) \cdot (T_m + 1)$ for $c \in \mathcal{C}$. The idea for this calculation is based on ecological reasoning, assuming that a single additional customer has booked a flight with no ancillaries ($c = \mathrm{N}$) in the previous time (data period I) and now (data period II) faces the selection $c \in \mathcal{C}$ again. Following this, we compare the fitted number of bookings $\hat{\mathbb{P}}(Y_m = c|\tilde{\boldsymbol{x}}_m) \cdot T_m$ with the above-calculated hypothetical number leading to

$$\widehat{\Delta y}_{m,c|\mathrm{N}} = \widehat{\mathbb{P}}\left(Y_m = c|\tilde{\boldsymbol{x}}_{m|\mathrm{N}}\right) \cdot (T_m + 1) - \widehat{\mathbb{P}}\left(Y_m = c|\tilde{\boldsymbol{x}}_m\right) \cdot T_m. \tag{8}$$

These numbers can be calculated for all three categories in $r \in \mathcal{C}$ for new covariate $\tilde{\boldsymbol{x}}_{m|\mathrm{r}}$ leading to $\widehat{\Delta \boldsymbol{y}}_{m,c|r}$ with $r = \mathrm{N, A, B}$. Note that $\widehat{\Delta y}_{m|r} = \sum_{c \in \mathcal{C}} \widehat{\Delta y}_{m,c|r} = 1$, since the hypothetical change in booking is based on one additional customer in each referring category $r$. The estimate $\widehat{\Delta y}_{m,c|r}$ can be interpreted as the unscaled choice $c$ in data period II in market $m$ conditioning on one hypothetical customer who would have occurred in the previous data period I and whose choice was $r$. Step by step, we can create a table of estimates with $r$ rows and $c$ columns. This leads to a customers' transition matrix from each referring previous choice $r$ to the current choice $c$. By averaging all market-specific estimates under each pricing policy group, we obtain

$$\widehat{\Delta \boldsymbol{y}}^p_{c|r} = \frac{\sum_{\mathcal{M} \in \mathcal{M}^p} \widehat{\Delta \boldsymbol{y}}_{m,c|r}}{|\mathcal{M}^p|}, \tag{9}$$

where $|\mathcal{M}^p|$ represents the number of markets under pricing policy $p$. The results are illustrated in Fig. 3.

We look at the transitions in markets under pricing policy "0" first. This will be our baseline for later interpretations and the matrix is dominated by its coefficients on the diagonal, which means customers would most likely keep their original choice. However, by $\widehat{\Delta y}^{\text{“0”}}_{\mathrm{B|A}} = 0.34$, we must not neglect the trend, that the bundled fare is becoming a popular choice for "a la carte" customers, expressed empirically by the change of booking proportions between 2016 and 2017. This transition is intensified by lower bundle price under policy "−" $\left(\widehat{\Delta y}^{\text{“}-\text{”}}_{\mathrm{B|A}} = 0.64\right)$ and plunged for higher bundle price under policy "+" $\left(\widehat{\Delta y}^{\text{“}+\text{”}}_{\mathrm{B|A}} = -0.28\right)$. In other words, for customer who bought ancillaries "a la carte" ($r = \mathrm{A}$), decreasing the up-sell price would lead to more transitions to bundles ($c = \mathrm{B}$) while increasing the up-sell price reduces this transition drastically. The asymmetry of price elasticity mentioned in the last section can also be found here again. In fact, this can be attended by the asymmetrical value function in prospect theory introduced by Kahneman and Tversky (1979). The absolute difference $\widehat{\Delta y}^p_{\mathrm{B|A}}$ resulting between "+" and "0" is about twice higher than between "−" and "0".

Keeping transition of markets under policy "0" as baseline, we interpret further transitions by comparing the matrices. For a hypothetical customer whose choice was no ancillary purchase ($r = \mathrm{N}$), the transition-vector $\widehat{\Delta \boldsymbol{y}}^{\text{“0”}}_{c|\mathrm{N}} = (0.92, 0.01, 0.07)$ represents his preference in data period II for each category N, A and B, respectively. This preference would have no relevant changes by decreasing the bundle price (i.e. $p = \text{“}-\text{”}$), since the transition-vector $\widehat{\Delta \boldsymbol{y}}^{\text{“}-\text{”}}_{c|\mathrm{N}} = (0.99, 0.04, -0.03)$ for pricing policy "−" stays similar. However, by increasing the bundle price (i.e. $p = \text{“}+\text{”}$) in contrast, the transition-vector becomes $\widehat{\Delta \boldsymbol{y}}^{\text{“}+\text{”}}_{c|\mathrm{N}} = (1.23, -0.19, -0.04)$ showing an increased preference for the fare with no ancillaries. This can be interpreted as stabilisation for their original choice, i.e. $r = c = \mathrm{N}$ for no ancillary purchase.

Furthermore, the preferences of former bundle customers ($r = \mathrm{B}$) are surprisingly little affected by price changes at this level. The coefficient $\widehat{\Delta \boldsymbol{y}}^p_{\mathrm{B|B}}$ for keeping bundle-buyers' original choice would increase by 0.04 for decreased bundle price and decrease by 0.03 for increased bundle price. Comparing the magnitude of changes to "a la carte" customers ($r = \mathrm{A}$), this suggests that the price elasticity of bundle customers is lower than that of "a la carte" customers at the level of the experimental setting.

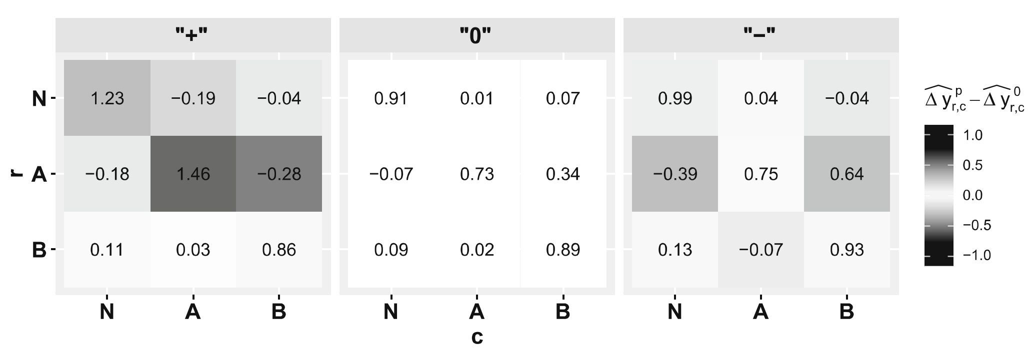


**Fig. 3** Estimates of hypothetical transition referring $r$ to current choice $c$ over all markets under each pricing policy $p \in \{\text{“}+\text{”}, \text{“0”}, \text{“}-\text{”}\}$. Greyscale visualises the absolute difference of the "+" and "−" group to the baseline group "0"

The final simple question that management would ask is "Well, should I increase or decrease the bundle price?" The answer is simple, too: "It depends.". Looking back to Fig. 1 which reminds us how heterogeneous the markets can be, the recommendation to AirABC is thus to use the findings in our interpretation depending on the market characteristics. On markets where the proportion of no ancillary buyers is high, AirABC should avoid reducing the bundle price in order to win more customers switching to buy bundles, since no relevant transition would be achieved through this strategy. On the other hand, due to the different price elasticity of customers of type "a la carte" and bundle, there is a certain potential of applying bundle price steering tools in revenue management on those markets with higher ancillary shares. To be more specific, the transition vectors $\widehat{\Delta \boldsymbol{y}}^{p}_{c|\mathrm{A}}$ and $\widehat{\Delta \boldsymbol{y}}^{p}_{c|\mathrm{B}}$ need to be taken into consideration for optimisation of bundle prices, as well as the market-specific shares of different customer types, i.e. $\tilde{\boldsymbol{x}}_m$. Note for a global optimisation of bundle prices, it will require more pricing points and corresponding experimental settings. The aim of this is therefore on the ex-post evaluation of applied pricing policy.

## Policy evaluation and economical implication

We extend the above results to make customers' behaviour more understandable. To do so, we use the idea of inferring causal impact, see e.g. Brodersen et al. (2015) and Varian (2016). We evaluate the applied pricing policies ("+" and "−" comparing to "0") by hypothetically calculating the expected revenue of markets in $\mathcal{M}^{\text{“+”}}$ and $\mathcal{M}^{\text{“−”}}$, if in fact pricing policy "0" would have been applied. To achieve this, we first predict market wise the numbers of booking in each category in data period II using the fitted coefficients for the market with pricing policy "0", that is we calculate

$$\hat{\boldsymbol{y}}_m^{\text{“0”}} = \mathbb{P}(\boldsymbol{Y}|\boldsymbol{x}_m, \hat{\boldsymbol{\beta}}^{\text{“0”}}) \cdot T_m, \tag{10}$$

where $m$ represents a market in $\mathcal{M}$"+" and $\mathcal{M}$"−", respectively. In a second step, we compute and aggregate the expected revenues in each group of markets, that is

$$\mathbb{E}^{\text{“0”}}(\mathrm{REV}^{\mathcal{M}^p}) = \sum_{m \in \mathcal{M}^p} \hat{\boldsymbol{y}}_m^{\text{“0”}} \text{€}^{\text{“0”}}, \tag{11}$$

where $\text{€}^{\text{“0”}} = (0, \text{€}_\mathrm{A}, \text{€}_\mathrm{B}^{\text{“0”}})$ is the column vector of prices for each category $c \in \mathcal{C}$ under pricing policy "0". Note that only the bundle price $\text{€}_\mathrm{B}^p$ differs between different pricing policy $p$. The expected revenue can also be calculated for the hypothetical cases if pricing policy "+" or "−" would have been applied, respectively. The results are listed in Table 4 and through this we can evaluate the effect of the applied pricing policies.

**Table 4** Comparison of realised and expected ancillary revenues, if each pricing policy "+", "0" or "−" was applied in each three groups of markets

| | $\mathcal{M}^{\text{“+”}}$ | $\mathcal{M}^{\text{“0”}}$ | $\mathcal{M}^{\text{“−”}}$ |
|---|---|---|---|
| Ancillary revenue in data period II | 389,970.0 | 764,960.0 | 328,800.0 |
| $\mathbb{E}^{\text{“+”}}(\mathrm{REV}^{\mathcal{M}^p})$ | *389,970.0* | *790,246.9* | 416,190.6 |
| $\mathbb{E}^{\text{“0”}}(\mathrm{REV}^{\mathcal{M}^p})$ | 369,592.8 | 764,960.0 | *417,716.7* |
| $\mathbb{E}^{\text{“−”}}(\mathrm{REV}^{\mathcal{M}^p})$ | 284,529.8 | 592,307.5 | 328,800.0 |

The maximal expected value is italics, which leads to the locally optimal policy

We can now tackle the question "what would have happened, if AirABC did not use the experimental setting of pricing policies in data period II?". That is, we calculate the expected revenue if only policy "0" was applied on all markets. However, since the pricing policy "0" was in fact applied on all markets in $\mathcal{M}^{\text{“0”}}$, we exclude these markets for above question and compare the respective expected revenue $\mathbb{E}^p(\mathrm{REV}^{\mathcal{M}^{\text{“0”}}})$ and $\mathbb{E}^p(\mathrm{REV}^{\mathcal{M}^{\text{“−”}}})$ depending on $p$. Starting with markets in $\mathcal{M}^{\text{“+”}}$, the expected ancillary revenue under policy "0" would be lower than the realised ancillary revenue, i.e. more ancillary revenue was achieved through policy "+". On the other hand, less ancillary revenue for markets in $\mathcal{M}^{\text{“−”}}$ was realised through policy "−" than the expected revenue under policy "0". This emphasises again, that the so-called promotion/bait[1] bargaining strategy is empirically and hence not economically advisable. This contradicts with the traditional opinion on marketing tactics as stated in Blattberg and Neslin (1990), even in the short term. Thus, we must not neglect the fact that airline ancillary bundles are not traditional commodities, certain established marketing tools may not work well for them per se.

As mentioned before, we can follow the idea of causal inference further and also calculate the expected ancillary revenue under any pricing policy. We thereby underlined the revenue optimal pricing policy in each group of markets. This gives us the evaluation, that for both groups $\mathcal{M}^{\text{“+”}}$ and $\mathcal{M}^{\text{“0”}}$, pricing policy "+" is considered as the revenue optimal pricing policy. For markets in $\mathcal{M}^{\text{“−”}}$, pricing policy "0" would be the best (moderately, compared to "+") choice regarding expected ancillary revenue.

[1] The sale of a product at a cut price, aiming to make up the decreased profit per unit by returning bigger gross profits.

## Discussion

Due to the lack of individual information in the data, it is not possible in our case to evaluate impact of pricing policy on individual behaviour in a traditional way. Thus, we suggest a statistical approach to model the different market-specific choices of individual customers based on aggregated market data. Based on the ideas of interpretations in ecological inference, we obtain hypothetical transition streams of customer depending on different pricing policy. The analysis of transitions and the comparison of realised and expected ancillary revenue mirror the estimated coefficients of our model. As a result, our recommendation for the stakeholders of ancillary pricing management is to study the customers price elasticity before implementing or investing in a dynamic pricing system for ancillaries. To sum up, it can be concluded that price elasticity exists partially and asymmetrically on the AirABC's markets under the three applied pricing policies. This bears risks as well as opportunities. Based on the empirical findings, the suitable strategy in our case is to increase the bundle price on those markets with higher share of ancillary purchase and do not decrease the up-sell price on those markets with lower ancillary shares, since no relevant transition to ancillary bundles can be won by lower prices.

The validity of identical characteristics in population as a key assumption should certainly be checked by further studies. We encourage comparative studies by using our model and a model based on individual covariates if possible. Crucial factors which are influencing consumers' behaviour depending on price while purchasing in a mixed bundling setting (e.g. in airline case: time to departure, departure day/time, number of passengers etc.) may provide more useful information in order to better understand their price elasticity. By comparing both approaches based on individual and market data, the validity of the conclusion can also be examined.

Regarding our model, certain modification can be considered, one possible alternative is to aggregate the estimation of market specific hypothetical changes $\widehat{\Delta \boldsymbol{y}}_{m,c|r}$ in Eq. (9) with weights of bookings in each market, i.e. $\widehat{\Delta \boldsymbol{y}}^p_{c|r} = \sum_{m \in \mathcal{M}^p} \widehat{\Delta \boldsymbol{y}}_{m,c|r} \cdot W_m$, where $W_m = \frac{T_m}{\sum_{m \in \mathcal{M}^p} T_m}$ or $W_m = \frac{T_{(x)m}}{\sum_{m \in \mathcal{M}^p} T_{(x)m}}$. The results of these alternatives mentioned above are given in Appendix. They are similar to our outcomes in section "Interpretation and strategic recommendation". After all, *All models are wrong but some of them are useful* (Box and Draper 1987). Any re-bundling needs to account for the price elasticity across market segments, and making up-sell economics work for airlines will require further research by the industry. We intend to not close but open a discussion for further research on methods to access price elasticity and evaluate pricing policy in this area.

## Appendix

See Figs. 4 and 5.

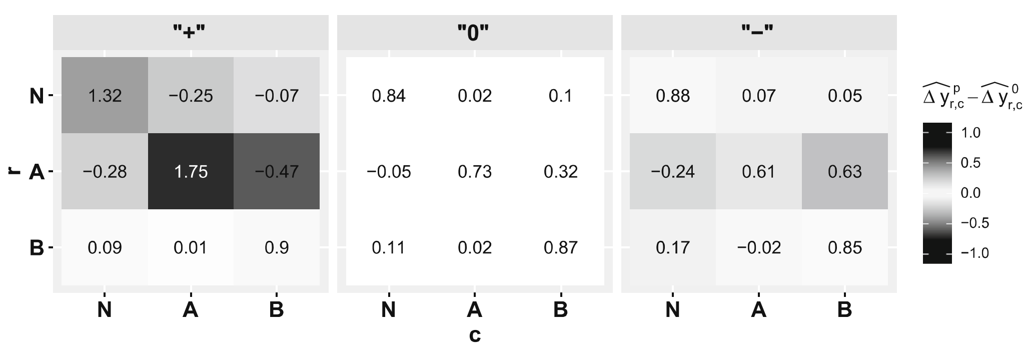


**Fig. 4** Estimates of hypothetical transitions by weighting Eq. (9) with $T_m$

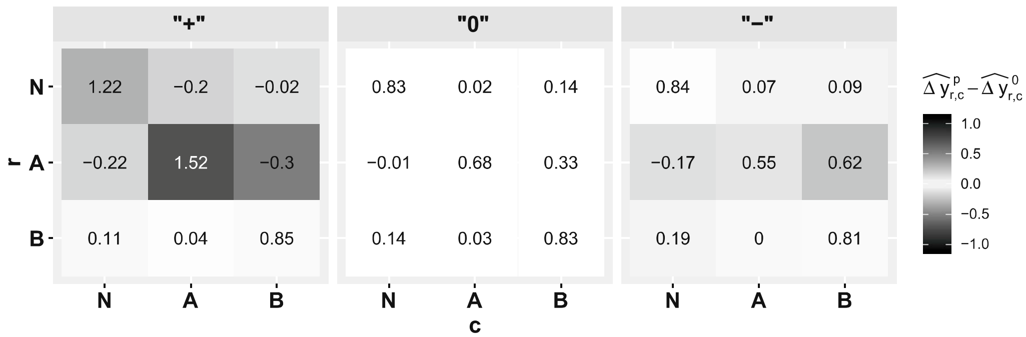


**Fig. 5** Estimates of hypothetical transitions by weighting Eq. (9) with $T_{(x)m}$

## References

Achen, C.H., and W.P. Shively. 1995. *Cross-Level Inference*. Chicago, IL: University of Chicago Press.

Adams, W.J., and J.L. Yellen. 1976. Commodity Bundling and the Burden of Monopoly. *The Quarterly Journal of Economics* 90 (3): 475–498.

Blattberg, R.C., and S.A. Neslin. 1990. *Sales Promotion: Concepts, Methods and Strategies*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Bockelie, A., and P. Belobaba. 2017. Incorporating Ancillary Services in Airline Passenger Choice Models. *Journal of Revenue and Pricing Management* 16 (6): 553–568.

Box, G.E.P., and N.R. Draper. 1987. *Empirical Model-Building and Response Surfaces*. Hoboken, NJ: Wiley.

Brodersen, K.H., F. Gallusser, J. Koehler, N. Remy, and S.L. Scott. 2015. Inferring Causal Impact Using Bayesian Structural Time-series Models. *Annals of Applied Statistics* 9 (1): 247–274.

Fahrmeir, L., and G. Tutz. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York, NY: Springer.

Gallego, G., and C. Stefanescu. 2012. Services engineering: Design and pricing of service features. In *The Oxford Handbook of Pricing Management*, ed. Ö. Özer and R. Phillips, 713–737. Oxford: Oxford University Press.

Kahneman, D., and A. Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47 (2): 263.

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.

Klima, A., P.W. Thurner, C. Molnar, T. Schlesinger, and H. Kuchenhoff. 2016. Estimation of Voter Transitions Based on Ecological Inference: An Empirical Assessment of Different Approaches. *AStA Advances in Statistical Analysis* 100 (2): 133–159.

Montgomery, D.C. 2012. *Design and Analysis of Experiments*. Hoboken, NJ: Wiley.

R Core Team (2018). R: A Language and Environment for Statistical Computing.

Ratliff, R., and G. Gallego. 2013. Estimating Sales and Profitability Impacts of Airline Brandedfares Product Design and Pricing Decisions Using Customer Choice Models. *Journal of Revenue and Pricing Management* 12 (6): 509–523.

Steppe, R. 2017. Online Price Discrimination and Personal Data: A General Data Protection Regulation Perspective. *Computer Law and Security Review* 33 (6): 768–785.

Tikkinen-Piri, C., A. Rohunen, and J. Markkula. 2018. EU General Data Protection Regulation: Changes and Implications for Personal Data Collecting Companies. *Computer Law and Security Review* 34 (1): 175–179.

Train, K.E. 2009. *Discrete Choice Methods with Simulation*, 2nd ed. Cambridge: Cambridge University Press.

Varian, H.R. 2016. Causal Inference in Economics and Marketing. *Proceedings of the National academy of Sciences of the United States of America* 113 (27): 7310–7315.

Venkatesh, R., and V. Mahajan. 2009. The Design and Pricing of Bundles: A Review of Normative Guidelines and Practical Approaches. In *Handbook of Pricing Research in Marketing*, ed. V.R. Rao, 232–257. Cheltenham: Edward Elgar Publishing.

Wakefield, J. 2004. Ecological Inference for 2 ×2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 167 (3): 385–445.

Yee, T.W. 2010. The VGAM Package for Categorical Data Analysis. *Journal of Statistical Software* 32 (10): 1–34.

**Shuai Shao** is a Doctoral Student and Göran Kauermann is Chair of Applied Statistics, both at Department of Statistics, Ludwig-Maximilians-Universität München. The department addresses the full range of statistical data analysis, focusing on the development and application of statistical methods in life science, social sciences and economic science. Their interdisciplinary research includes cooperation with both academic and industrial partners.

# Chapter 4

# Final remarks

The success of airlines has always been accompanied by new challenges and opportunities, as well as new methods and new disciplines. Given the new freedoms to set prices after the deregulation of markets in the late 70s, airlines rapidly embraced the concepts of differential pricing, in which different prices are offered not only for different physical products (e.g. First, Business and Economy cabins), but also identical services within the same cabin. Differential pricing involves both methods, product differentiation and price discrimination. Based on these methods, RM has since then developed various innovations on network optimisation as well as choice modelling, and established itself as a capacious discipline.

Since the beginning of this century, existing RM systems are facing new challenges. As a major recent development in the airline industry, ancillary revenue opportunities bring challenges in parallel with the big data revolution. This thesis encompassed different statistical methods applied to airline RM. It is demonstrated that regression models can be modified as tailor-made solutions for accessing WTP and quantifying PE. In the case of ASR, it is shown how a single ancillary item can be embedded in current RM systems with differential pricing based on semi-individual booking data. Furthermore, models based on aggregated market data are proposed to evaluate revenue-oriented ancillary bundle pricing policy. In this spirit, models applied in this thesis can also be utilised in other industries (e.g. hotels, car rentals, cruise companies) with the need for incorporating ancillary revenue to their RM systems. As the contributions in each chapter already contain specific remarks and research perspectives, global remarks on applying statistical methods to the transportation and tourism industry in connection to big data are given in following.

Originally[1], big data are defined by its big *volume*, *velocity* and *variety* (Laney 2001). Modern transportation and tourism as information-intensive industries are predestined to be affected by big data seeing that

- **volume**: the extensive network of destinations leads to large databases;
- **velocity**: booking data are often available on a real-time basis;
- **variety**: a rich set of ancillary services provides different types of information.

As King (2016) notes, "*big data is not about the data*", but much more about extracting valuable information from them. To achieve this goal, multiple scientific disciplines are in demand, or even competitive in a certain way. For instance take the generative and algorithmic culture defined by Breiman (2001): The former is more concerned with the data generating process from the statistical point of view; the latter has more focus on predictive capability and is more present in computer science.

While the majority of academical research in statistics is dedicated to further methodological development in a very advanced and sophisticated way, it is also essential to make use of the opportunities in interdisciplinary applications. In the action plan proposed by Cleveland (2001), it is even suggested to give interdisciplinary investigations the most resources. At long last, the benefits of the developed statistical methods cannot be demonstrated without application.

Finally, the author would like to emphasise the cited aphorism from Box and Draper (1987) again: since "*all models are wrong, but some are useful*", both generative and algorithmic culture can be useful for real-world problems in the big data era, depending on the research question, goal and complexity. Indeed, the emerging scientific discipline *data science* can also make valuable contributions to RM, since it combines but not replaces both cultures (Kauermann and Seidl 2018). For the collaboration with RM, statistical methods can be more useful by offering not only models but also the theoretical foundation behind, which can find their equivalence in the quantitative branches of economics; or more importantly, their equivalence in the standard steering parameter in RM systems (recalling the success story of RM stems from transferring theory to practice). With no doubt, developments in *interpretable machine learning* from the algorithmic culture can be useful in this way as well, see more details in Molnar (2019). In this spirit, this thesis can also be seen as an appealing research invitation to applying interpretable machine learning to RM.

[1] The original definition of big data with 3 Vs has been since then extended in many versions, e.g., additionally with veracity, value, validity, variability and volatility. Here, the respective justification is disclaimed. Instead, the focus lies on demonstrating how the transportation and tourism industry is affiliated with big data.

# References


Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Hoboken, NJ: John Wiley & Sons.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures". In: *Statistical Science* 16(3), pp. 199–215.

Cleveland, W. S. (2001). "Data science: An action plan for expanding the technical areas of the field of statistics". In: *International Statistical Review* 69(1), pp. 21–26.

Kauermann, G. and Seidl, T. (2018). "Data Science: a proposal for a curriculum". In: *International Journal of Data Science and Analytics* 6(3), pp. 195–199.

King, G. (2016). "Big Data Is Not About The Data!" In: *Computational Social Science: Discovery and Prediction*. Cambridge, UK: Cambridge University Press. Chap. Preface.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. February 2001, pp. 1–4.

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. `https://christophm.github.io/interpretable-ml-book`.

# Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 06.03.2019

Shuai Shao