Alma Mater Studiorum
Universitá di Bologna

DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
Ciclo XXVIII
Settore Concorsuale di afferenza: 13/D2 - STATISTICA ECONOMICA
Settore Scientifico disciplinare: SECS-S/03

Predicting nominal data in presence of poor information.

An application to air tickets.

Coordinatore:
Prof.ssa Alessandra Luati

Presentata da:
Annalisa Stacchini

Relatore:
Prof. Andrea Guizzardi

Co-relatore:
Prof. Michel Mouchart

Esame finale anno: 2016

Data is not information. Information is not knowledge. And knowledge is certainly not wisdom.

H. Gilbert Welch

Acknowledgements

I wish to acknowledge all the people that supported me during the Ph.D. and the elaboration of the present thesis. First my supervisor and co-supervisor, for the precious advices, the kind availability and irreplaceable help they provided to me. Many thanks to Professor David Draper for the truly kind hospitality at UCSC, his great availability and all the suggestions. I also thank all the professors of the Ph.D. board for having realized such a valuable and fascinating program.

Then, I acknowledge Seneca, in particular Ercolino Ranieri and Luigi Cristini, for having financed my scholarship, provided the datasets and chosen a topic that opened original research problems. I thank Thot for the business-specific loss function.

A further acknowledge is for my Ph.D. mates, especially Flavio Emanuele Pons, Saverio Ranciati and Matteo Farne', who gave me a lot of precious information and were always ready to answer my often weird questions.

Finally, a very special thanks to all of my family, especially to my father, mother and grandmother, for having supported me in all the possible ways and with great affection.

Contents

List of Figures

List of Tables

Introduction

The present thesis is written upon commission by a Travel Management Company (TMC). Among the various services and products offered by the TMC, an important business is constituted by the brokerage of business flights. A special contract, developed by the TMC, provides that the client company can agree to pay a fixed price for any flight with the wanted characteristics (on board class, route, airline). This arrangement puts the TMC in the position in which corporate travel departments are normally, having to minimize the cost of flights.

In fact, the management of business travel costs is becoming an important issue for more and more corporations, in increasingly globalized business environments. While in companies that do not require numerous business flights the ticket is often booked by the traveler himself (often in view of his comfort and frequent flyer program), corporations needing many flights tend to formulate a corporate travel policy, aiming at minimizing the cost of flights. Whence the need of a decision-making strategy able to help in this complex task, especially nowadays, when the bargaining power of corporations (with respect to the airlines) is very low and it is becoming impossible to get real quantity discounts, due to the global agreements among airlines and their strong revenue management strategies.

Most corporate travel departments, at least in Italy, assume that the best buy strategy consists in purchasing always the cheapest ticket, also considering that the difference in price, between fares, is great. But does 'best buy' mean buying always the cheapest ticket? Considering that the cheapest air fare is the fixed one, not allowing changes nor refund, whether the traveler changes or renounces to the flight, the cheapest first issued ticket leads to a higher overall cost of the flight, than that allowed by a more expensive but more flexible fare.

While choosing the cheapest air ticket is easy, as the prices of different fares are known and comparable at the time of booking, getting the needed flight at the lowest possible cost is not that straightforward. In fact, the cost of business flights does not depend only on the price of the first booked ticket, but also, all other conditions (class, route, date, time, airline) equal, on the chosen fare and on the behavior of the business traveler, which in turn will determine which fare was the optimal one (at the time of booking). The more flexible the fare, the higher the price, the lower the cost of changes and renounces.

If, at the time of booking, a leisure traveler does not know for sure whether he will change date or destination, or will have to renounce to the journey, a corporate travel department is even more uncertain about the business traveler's behavior. In fact, a leisure traveler can guess quite reliably that he will (not) flow the initially purchased ticket, because he has a wide knowledge of his personal, familiar and professional situation (for example, if he books a flight for a holiday, but his mother-in-law is seriously sick and may need assistance, he will opt for a refundable fare).

Contrarily, in a business context, the traveler undergoes the events, which are often unforeseen. For example, he could have to change the date of a flight, because a client needs to postpone or antedate a meeting, or he may have to renounce, because of an epidemic in the destination. Such a broad uncertainty implies that the cost of flight is unknown at the time of booking, so that the corporate travel department cannot minimize it directly. Therefore, 'best buy' means choosing the optimal fare minimizing the value of the cost of flight, as expected based on the available information, able to help predicting how the traveler will behave.

Therefore, this business problem requires a statistical solution, because of the uncertainty about the travelers' behavior, which can be modeled as a nominal random variable. The objective of the present work is to provide such a solution. Appearently, the one at hand, is a simple decision-making problem, under minimization constraints. The action space is constituted by the set of fares among which the purchasing agent can choose. The constraints consists in deciding for the fare minimizing the expected value of the cost of flights.

Indeed, the task is made especially difficult by the poor information available. In fact, the corporate dataset, provided by the TMC, was collected for purposes different from that of the investigation. Therefore crucial data are missing, beginning with the fares of observed tickets, their price and the penalties for changes. This lack of necessary information makes impossible to define the action space and to compute the expected value of the cost of flights.

Moreover, observations cover a too limited time span, in particular the estimation sample spans just seven months, so that no eventual time component can be detected. In addition, the hierarchical structure of data (tickets bought for a traveler, travelers working for a company, company belonging to a certain macro−category) cannot be modeled, because the dataset is greatly unbalanced, due to the 80/20 rule of sales, and the identifiers of groups are missing in too many cases and it is not even sure that non-identifiable tickets do not actually belong to an identified group.

Finally, most of the available data refer to the characteristics of the ticket and the flight, that are not well correlated with the behavior of the business traveler. Information about the flyer's professional status, personal and family situation, and its company's business would have been very useful to predict his behavior and give an indication about which fare to choose. Their lack causes serious forecasting problems, also considering that it is not possible to perform many elaborations on nominal data, due to their qualitative and non-ordinal nature.

As a consequence, the present work aims at developing alternative solutions, useful for predicting nominal data in border-line situations of this kind, where the available information is so poor that the traditional statistical techniques need to be integrated. Thus, the original contributions of this thesis, to the prediction of nominal data in presence of poor information, are three.

First, proposing an easy method to incorporate a guess about a non−estimable effect, of a non−observable variable, directly into an estimated model. Second, developing a classification algorithm, able to extract from the whole matrix of predicted probabilities, both the latitudinal and longitudinal information, able to correctly classify at least some of the most economically relevant outcomes, that are also the greatly less frequently observed ones. Last, propounding a new measure of forecasting accuracy to select the best predictor among a set of models appearently with identical predictive performance, as assessed through the extant statistical methods.

While the importance of the present work, from the phenomenic perspective, consists in providing a statistical tool suggesting the TMC which one is the best fare to buy. But also in providing a formal approach for helping corporate travel management departments minimizing business travel costs. More generally, in showing that 'best-buy' does not always mean choosing the cheapest fare.

This thesis is articulated as follows. The context of the study is first described, providing a digression about the birth and the development of companies specialized in business travel management, an illustration of the characteristics of the specific TMC commissioning this work, a formalization of the main research problem and a review of the little extant literature. A qualitative study of the business problem follows. Some methodological notes are premised, then the semi-structured interviews and the panel of participants are described, finally the findings and the obtained guidelines for the subsequent quantitative study are presented.

The dissertation continues illustrating the corporate database, its composition, the specification of the variables, the problems of missing data and some sample statistics. Chapter four deals with modeling the phenomenon under investigation: first the business problem is simplified, then a functional form is chosen, some notes on the important assumption of independence from irrelevant alternatives are discussed, explanatory variables are selected and estimation outputs are commented.

Chapter five is the most innovative one: after detecting the causes of the emerged prediction problems, some alternative solutions are developed. First an easy method for vague guess-based prediction, then a business-specific loss function computable with the few information at disposal, employable for comparing the predictors' economic performance. In addition, a new classification algorithm and a measure of forecasting capability considering the estimated probabilities. Finally results are presented and the thesis is concluded.

Chapter 1

Context of the study

1.1 Travel Management Companies

The present Ph.D. thesis was commissioned by Seneca, the Travel Management Company (TMC) which funded my Ph.D. scholarship. Contextualizing the research problems, presented in the following chapter, allows to understand the implications and importance of the study, because it is very operational in nature and tightly related the business activity of the TMC. As TMCs are a relatively new form of intermediaries in the travel market, it is worth spending some words to explain what they exactly are, which is the commercial room on which they rose and how they differ from the well known travel agencies.

A TMC can be seen as an evolution of the traditional travel agency, in the sense that it is specialized in business travel and, besides the usual intermediation activity, offers a wide set of travel management services. TMCs arose from the change in the relationship between airline companies and travel agents, begun in the USA in the second half of the nineties, when the economic recession induced the airlines to reduce costs (Levere, 2000). In fact, about the 17% of airlines' total operating costs in 2000 was related to distribution, the third largest cost, after labor and fuel (International Air Transport Association, 2000). Recently these costs are steadily decreasing (see: Air Transport Association, 2014).

As the progress of information technology (IT) made search and booking procedures much easier and cheaper, travel agencies' commissions, accounting for more than the 10% of the air companies' distribution cost in 1993 (currently about the 8%), were questioned and slashed by the airlines. Following the American companies, soon also European carriers reduced agents' commission fees, starting from the British Airways, in 1998 (Alamdari, 2002).

Gradually, the contraction of travel agents'revenues was exacerbated by the increase in the bargaining power of customers. In fact, more and more individuals and small companies prefer direct contact with airlines, or online booking sites, automatically comparing flights and prices. While big enterprises, became aware of the importance of managing travel expenses, started adopting self-tailored travel policies and buying directly from the airlines, to get volume discounts. Therefore travel agencies reshaped their business and often specialized in leisure tourism, focusing on the more remunerative supply of vacation packages, or in business travel, offering consultancy services to help companies developing and enforcing travel management policies, in order to optimize their travel spending.

The second choice characterizes TMCs, which mainly provide: up-to-the minute reports on travel patterns of employees, reports on the effectiveness of travel policies, advice on complicated itineraries, consultancy on travel data management, day-today operations of the corporate travel program, advice for planning and budgeting, traveler safety and security, and credit-card management. Thus, currently TMCs' revenues are mainly generated by the commissions charged for such services, on the supply of which, rather than on intermediation, they compete against their rivals worldwide.

However, some companies build their competitive advantage also on the offer of innovative intermediation contracts, adding, to the traditional distributional activity, conditions increasing the value for the clients. Whence the importance, for TMCs, of research and innovation, both of produced services and of production processes, as it is the objective of the present work.

1.2 The commissioning TMC

Seneca is an experienced TMC, for over 20 years one of the top business travel agencies in Italy, characterized by a constant commitment to research and development of both new services, in order to anticipate the market changes and widen its offer, and new production processes, for increasing its operational efficiency. Seneca'offer mainly consists of:

- Business travel services,
- Business hotel services,
- Travel Management services,
- Hotel representation,
- IT systems,
- A new Global Distribution System for hotels,
- Target Buy purchasing pattern.

The Target Buy contract is one of the most appealing Seneca' s innovations. It provides that the client and the TMC bargain the target price (for a single or set of air routes, for a category of hotel room, etc.) at the signing of the contract.Then, the client will always pay that same target price for that product, discharging the risk of price variations on the TMC, which will buy at the spot price from the providers. Thus, the client has the advantage of purchasing at stuck prices, hedging the risk of fluctuations in the travel expenses, and of knowing in advance the amount of the cost, making the travel budgeting more certain and its management easier. Moreover, if the customer renounces to the flight, or to the hotel stay, at any time prior to departure, the TMC refunds the full target price. The client is also allowed to change the time and/or place of flight/overnight stay, completely free of charge, or paying a fixed penalty, if provided in the contract. Therefore, this contract is very risky for the TMC, that has to be very careful in setting the target price and try to manage the risk of change and waiver.

Developing an innovation in the purchasing process of air tickets, able to minimize the economic losses due to the risk of change and waiver of air tickets, is the aim that Seneca set for the present thesis. The usefulness of such an innovation is not only circumscribed to the Target Buy contracts, but extends to the whole activity of air tickets intermediation. In fact, as it is explained in the following section, for the same route, date, advance booking and class (determined by the client), it is possible to buy various tickets, differing in the degree of flexibility and price. Thus, even for non-Target Buy clients, Seneca can choose which one to buy and the economic result of each transaction depends on the kind of purchased ticket and the risk of change and waiver.

1.3 Research Problems

Given the context of the present thesis and the requirement of the commissioning TMC, it is clear that this work develops around the solution of a very concrete and firm-specific problem, which poses itself operational and research problems, from which more general methodological issues, about applying Statistics to business problem solving, stem.

The research problem, that the TMC explicitly posed, is to find a method for minimizing the cost of flights. While choosing the cheapest air ticket is easy, as the prices of different fares are known and comparable at the time of booking, getting the needed flight at the lowest possible cost is not that straightforward. In fact, ceteris paribus (class, route, date, time), the cost of a flight depends on both the fare and the behavior of the business traveler (TB). The more flexible the fare, the higher the price, the lower the cost of changes and renounces.

The cost of a flight (C) is given by the price (Pr) of the chosen fare at the time of the first reservation (F), plus, in case the business traveler changes (Ch) the ticket and the fare is not flexible, the sum of the cost of each change (Cch_c , $c = 1, ..., Nc$), which in turn depends on the fare, minus, if the traveler renounces (Nf) to the flight and the fare is flexible, the amount of the refund (Ref):

$$C = Pr_F + 1_{\{TB=Ch,F\neq flex\}} \sum_{c=1}^{Nch_{TB}} Cch_{c,F} - 1_{\{TB=Nf,F=flex\}} Ref_F \tag{1.1}$$

where $1_{\{TB=Ch,F\neq flex\}}$ is an indicator function, equal to 1 if the business traveler changes the first issued ticket and the fare is not flexible, 0 otherwise. $1_{\{TB=Nf,F=flex\}}$ is an indicator function, equal to 1 if the business traveler renounces to the flight and the fare is flexible, 0 otherwise.

If, at the time of booking, a leisure traveler does not know for sure whether he will change date or destination, or will have to renounce to the journey, a TMC (or, within a 'client' company, is even more uncertain about the business traveler' s behavior. In fact, a leisure traveler can guess quite reliably that he will (not) flow the initially purchased ticket, because he has a wide knowledge of his personal, familiar and professional situation (for example, if he books a flight for a holiday, but his mother-in-law is seriously sick and may need assistance, he will opt for a refundable fare).

Contrarily, in a business context, the traveler undergoes the events, which are often unforeseen. For example, he could have to change the date of a flight, because a client needs to postpone or antedate a meeting, or he may have to renounce, because of an epidemic in the destination.

Such a broad uncertainty implies that the cost of flight is unknown at the time of booking, so that the TMC (or the corporate travel department) cannot minimize it directly. Therefore, 'best buy' means choosing the optimal fare ($F*$) minimizing the value of the cost of flight, as expected based on the available information (X), able to help predicting how the traveler will behave:

$$F* : E_{TB}[C_{F*} \mid X] = min\{E_{TB}[C_F \mid X]\} \tag{1.2}$$

$$E_{TB}[C_F \mid X] = Pr_F + 1_{\{F \neq flex\}} \sum_{c=1}^{Nch} P[TB = Ch_c \mid X] Cch_{c,F} - 1_{\{F = flex\}} P[TB = Nf \mid X] Ref_F \tag{1.3}$$

where $P[TB]$ is the probability of business traveler's behavior, or, seen from the TMC's perspective, the probability of tickets' outcomes.

Thus, for making the optimal purchasing decision, it is necessary to estimate $P[TB \mid X]$. Clearly, the optimal solution would be obtained if X were able to reduce the uncertainty about TB to the mere statistical error. Therefore, modelling and predicting the 'risk of non-fly' is the first research problem addressed in this thesis.

The other research problems derive from the poorness of the information available to perform this, otherwise easy, task. A first issue concerns the specification of variables. Given that a few client companies purchase most of the flights, that are nearly identical (same X values) and, among numerous tickes, just very few are not flown; that most of the available data about independent variables is categorical, with numerous classes for most of them and a highly concentrated distribution; and that the literature on this topic is neraly null, there is no guidance about how to specify explanatory variables and to aggregate the too many levels.

A specification problem emerges also with reference to the dependent variable. In fact, the number of possible changes is virtually infinite and they can be either total or partial. So a simplification of the events space must be done. However, the possible tickets' outcomes are still too many and, if specified as levels of a single dependent variable, would make any model unidentifiable.

In addition, most of the available variables are not well correlated with the traveler's behavior, so just a very small proportion of its variability can be explained. This problem is due to the fact that the database has been collected for purposes different from that of the investigation, but also to cost and privacy constraints, often found in business contexts. Furthermore, the estimation sample covers only seven month, so that no eventual time component can be modeled. Moreover, the mentioned high unbalance of the panel structure of the dataset does not allow to specify hierarchical models.

As a consequence, the discriminatory power of available explanatory variables is extremely low and the forecasting performance of estimable models is very disappointing. Thus, a further research problem, addressed in this thesis, refers to the development of alternative solutions to improve the predictive capability of the models.

In addition, for choosing the optimal fare, minimizing the cost of flights, it would be necessary to compute the expected value of such a cost. This is possible if the price, the penalty for changes and the fare of observed tickets are known, but in the available dataset this information is missing in too many cases, so that this traditional approach cannot be adopted.

Therefore, after estimating the probability of tickets' outcomes, it is necessary to define an optimal decision rule, as a function of such probability, ready-to-use within the everyday working practice of the TMC. Where the 'optimal' rule is the one maximizing the economic result of intermediation operations for Seneca. Thus, a method to compare the economic performance of candidate predictor models with that of the current business practice, in absence of sufficient economic information, must be proposed. In fact, in business environments a probability value acquires a meaning in relation to the economic output of a decision based on such value, rather than in comparison with other values.

For example, whether it is found that the probability of changing the ticket for business travelers, making trips with certain characteristics, is 60% and that of flying is 40%, should the purchasing clerk buy a, more expensive, flexible ticket, instead of a fixed one, because the probability of change is (slightly) higher than that on non-change? The issue is: will the higher cost be worth that 10% of extra probability? Thus, formulating the optimal decision rule, operationalizing the estimates of the risk of non-fly, obtained through different models (as no literature suggests how to model the phenomenon of interest), and selecting the best performing one, in terms of economic utility for the TMC, is the last research problem, addressed in this thesis.

1.4 Literature Review

The management of business travel costs is becoming an important issue for more and more corporations, in increasingly globalized business environments. While in companies that do not require numerous business flights the ticket is often booked by the traveler himself (often in view of his comfort and frequent flyer program), corporations needing many flights tend to formulate a corporate travel policy, aiming at maximizing travel security end efficiency, while minimizing the cost of business travel.

With reference to the air tickets, the capability of minimizing the cost of flights largely depends on the choice of the optimal fare. Many studies (Turismo d'affari, 2014) highlight that, in this prolonged period of economic difficulty for Italian enterprises, "best" buy become synonym of "cheapest ticket". In fact, many corporate travel policies prescribe to purchase always the cheapest fare available at the time of booking, also considering that the difference in price, between fares, is great. This evidence is also supported by the increase of business flights operated by low cost airlines (see: Amex investigations).

But Proussaloglou and Koppelman (1999) found evidence that appears to contrast with this hypothesis. Their study about how the air travelers choose among different carriers, flights, and fare classes, shows that both business and leisure travelers are willing to pay a premium for the flexibility offered by the most expensive fares. Fourie and Lubbe (2006) investigated the determinants of the business travelers' choice between low-fare and full-service air companies, in South Africa. They found evidence that, except for the price, the decision is led by considerations about comforts and service level, that are surely important for the traveler, but, for corporate travel departments, not as crucial as the cost management.

Hence, the question, whether best buy means buying always the cheapest ticket, is open and crucial. The present study suggest to face the choice of the optimal fare considering that the overall cost of a business flight is composed by the price of the first issuance, plus the cost of the required flexibility. The first component can be managed by bargaining volume discounts with the airlines, although the progressively more widespread diffusion of revenue management systems and the global agreements between air companies lowered the bargaining power of client companies, leaving the business travel management the only opportunity to contract a percentage reduction on the (expensive) business fare, unilaterally defined by the airline.The second component can be controlled by forecasting the travel behavior.

In the extant literature, this issue is addressed from the airlines' perspective (Gorin et al, 2012; Bartke et al, 2012), aiming at developing pricing methods minimizing the negative impact, on their margins, of cancellations (for unrestricted fares) and rebooking (for restricted fares) close to the departure. To the best of our knowledge, no study considers the prediction of the outcome of booking (change, no-show, cancellation), in the definition of 'best buy', from the perspective of corporate travel departments (or TMCs).

Indeed, literature deals with sunk costs (Park and Jang, 2014), or focuses on changes and waivers as determinants of the choice between low-fare and full-service air companies (Fourie and Lubbe 2006). Further papers concern the choice of the airline (Nako, 2992), the influence of time and service quality on air travel demand (Anderson, and Kraus, 1981), factors influencing price elasticity (see: Brons et al, 2002).

More in general, literature about the market of air tickets mainly deals with the mounting use of the internet for online booking. It analyses the new services, pulled by this phenomenon, offered by the air companies (Law and Leung, 2000), the distributors (Bitner and Booms, 1982) and accessory services providers (Law and Chang, 2007); the strategic and economic consequences of online booking for the airlines (Yoon et al, 2006); the consumers' perception of the risk of employing the new media (Kim et al, 2009) and, conversely, the determinants of their trust (Alam and Yasin, 2010).

The prices of air tickets have been considered in a few works, investigating, from the perspective of the airline, the influences of the compaies' financial situation (Borenstein and Rose, 1995), services provision costs and demand level (Botimer and Belobaba, 1999), and price wars (Brady and Cunningham, 2001) on pricing, but mainly on the theoretical level.

But the most interesting researches, for the aim of the present thesis, are those analysing air tickets' prices through hedonic-price and regression models, because they highlight the attributes of the ticket/flight, which play a major role in determining the fares. In particular, for air tickets distributed through 'traditional' channels, advance purchase, airline, destination and saturday night stay have been found to crucially influence prices (Vowles, 2000; Wallenberg, 2000).

Along with these characteristics, also online travel agent, time windows for departing and arrivals (Clemons et al, 2002), kind of connection, fuel price, peak hour of departure, seasonal dynamic, seat class (Chen, 2002), and recently maximum stay, refundability, restrictions on flights (Lin et al, 2009) were investigated, resulting important for price determination. However, these analysis are limited to tickets homogeneous with respect to the flight route and class. Moreover the different ticket conditions, especially with reference to the degree of flexibility, within the same class, are aggregated.

Indeed, the lack of more complete empirical studies on the market of air tickets is not that surprising, given the complexity of obtainable data. In fact, it has been proved that identical tickets can be bought at very different prices from different online distributors, evidences pointing at relevant imperfections in the online air ticket market (Clay et al, 2001; Lin et al, 2009).

Another difficulty, found in the present work and implicitly shared by the quoted studies, consists in classifying tickets based on their degree of flexibility, which is of special importance to business travelers (Mason and Gray, 1999), because different airlines offer different sets of alternatives, which are not fully comparable across operators. Moreover, often these conditions are affected by the uncertainty deriving from clauses providing that the penalties will be computed (who knows how) at the moment of ticket change/waiver.

In brief, the rare literature of interest for the present work confirms what is observed in the database at hand (see chapter 4): that the tickets' contractual conditions are extremely various, because each single airline defines its own typologies of fares and on-board booking classes, which jointly determine the degree of flexibility of the ticket and the differences in cost and services between them.

Each airline company has its own pricing policy, which can vary in time and space, and often does not offer all the listed conditions on every flight, while sometimes those conditions are available through insurance bills, managed by third-party insurance companies. This situation is made more complicated by the different prices applied to identical tickets by online travel agents. Although generally, a fully refundable ticket (which can be entirely refunded at any time by the airlines) costs about 50% more than a non-refundable one (McAfee and Velde, 2006), the differences in prices due to the different degrees of tickets flexibility can vary a lot, even within the same airline. These are the main reasons why the market of air tickets seems rather a 'jungle'.

Chapter 2

Qualitative investigation of the phenomenon

2.1 Methodological notes

In Social Sciences and Medicine qualitative research is widely spreading (e.g. Wilfried and Tarnai, 1999; Fossey et al, 2002), but it can be very useful also in other fields of study, especially when working on a problem for the solution of which no literature is available, as in the present case. The qualitative approach aims at analyzing a problem in depth (rather than ’in width’ as it is the case of the quantitative approach) involving a limited number of subjects, belonging to a certain group, defined on the basis of variables likely to be associated with the phenomenon of interest (Wunsch et al, 2014), contrarily to what happens in quantitative statistical sampling design.

Once overcome the long lasting dichotomy between ’comprehension’ and ’explanation’ of phenomena (Dilthey, 1883), qualitative and quantitative methods are finally being recognized as complementary (Malterud, 2001; Wunsch et al, 2014). The main purposes of qualitative research are (Fossey et al, 2002):

- to improve the understanding of the object of investigation from the perspective of the involved subjects;

- to explore the meanings of phenomena as directly experienced by individual themselves, within the context of their life and social environment;

- to lead designing of further, eventually quantitative, research.

As highlighted by Gordon and Smith (2004), qualitative investigation is an especially serviceable tool when addressing phenomena characterized by multiple causal mechanisms, in which different causes can produce the same outcome and it is not possible to observe which mechanism generated each effect in the available sample.

This is the case of the present work: different unobserved motivations are likely to lead business travelers to fly or not to fly. The available quantitative data allow to analyze the behavior of a (relatively) great number of travelers, thus promising to find constant patterns, which can be generalized to the whole population of business travelers, to forecast their choices. But no information about the actual motivations of changes, refunds, flights, non-flights is available. These motivations are (as shown in the following) the actual causes of tickets' outcomes and can be (hopefully) related to some of the candidate explanatory variables at hand, but from the dataset it is impossible to find out which variables are related in which way to the 'real causes' of the events of interests. While it is such a knowledge that should lead the modelling of the probability of non-fly. Whence the opportunity to realize a deeper qualitative study on just a few individuals, which can shed light on the causal mechanisms hidden behind the business travelers' behavior.

The first crucial issue, in qualitative studies, is the definition of the method to employ. The main alternatives are: observation, written questionnaire, oral/written interviews. In the present case, the first alternative is unviable, because it is not possible to follow business travelers at work. As the primary purpose of this qualitative research is to explore hypotheses on variables, for planning the subsequent quantitative study and implement a sound statistical strategy, a semistructured interviews appear to be the optimal choice (Malterud, 2001). In fact, the less pre-determined is the interview, the less the researcher's subjectivity, prejudices and a priori opinions affect the study. The idea is to let participants to express the 'lived' meaning of their own experience of the actual context of 'business flying', as freely as possible.

As an appropriate sampling methodology is fundamental for quantitative studies to be reliable, the choice of the units of analysis is possibly even more crucial in qualitative research, because just a few individuals are selected. Dealing with sampling, it is to be noted that in the present case there is no sampling, but census of all the tickets intermediated by Seneca since the data collection became possible, as the reference population is composed by all the tickets purchased by the TMC and, just by extension, to that of business travelers, of the behavior of which the tickets' outcomes are 'objective correlates'.

However, turning to qualitative analysis, Graneheimand Lundmanl (2004) suggested that the selection of participants living different experiences increases the potential of widen the understanding of the research question to a variety of aspects. Moreover, in order to enhance the transferability of findings, it is worth of providing an exaustive description of participants' cultural background, context and subjective characteristics. Thus, the present work proceeds as suggested, to the extent to which it is possible, because it is not easy to realize depth interviews with different frequent business flyers.

In fact, business travelers are not even minimally remunerated for the participation to the interview, which is motivated solely by friendship, therefore individuals have been chosen among my friends, based on the frequency of flights they do for work. Another difficulty consists in the fact that frequent flyers spend a lot of both spare and working time for traveling, thus they have very few time left for other purposes, like answering interviews.

Once chosen the method and selected the participants to the study, the issue of how to use the obtained information arises. As highlighted by Malterud (2001), coding such information as numerical variables and processing it through statistical techniques for qualitative data is not the most appropriated method. In fact, the scientific logic, on which statistical techniques rely (especially the requirement of independence of the variables used for the selection on the model's dependent one) are incompatible with the 'non-representativity' of the interviewed group and, as the interview is semistructured and mainly informal, with the circumstance that questions are not asked nor answered in standardized way (thus not liable to form homogeneous categories). Thus, retrieved information is 'qualitatively' employed, without formal elaboration.

It is to be noted that this qualitative analysis is not to be employed only before the quantitative study of the phenomenon under investigation, but it is also useful after that, to better interpreting the meaning and implications of findings. Thus, the actual pattern of integrating qualitative and quantitative analysis, in the present work, can be named 'triangulation': "The aim of triangulation is to increase the understanding of complex phenomena, not criteria-based validation, in which agreement among different sources confirms validity." (Malterud, 2001, p. 487). In particular, this approach refers to the practice of non-mixing qualitative and quantitative data in modelling the phenomenon, but respecting the difference in nature and in the statistical characteristics of the two type of informations. This imply integrating them in a meta-analysis (or secondary analysis), leading to mutual validation of the results of both.

2.2 The semi-structured interviews

The interviews were realized mainly through a popular social network, because participating business travelers were in different counties, at the time of the interview. Only an interview was realized in presence, orally. The interviews were informal, all followed a common draft, but in a very flexible way, as the questions were formulated as a function of previous answers, consistently with the business traveler profile which was gradually being outlined. The draft is the following:

1. *How many flights do you do for business, on average, per month?*
2. *Is there, in your company/institution, a (or more) person assigned to the task of buying or reserving flights? Or do you do it by yourself?*
3. *If it happens that you waiver a flight, for which you already bought the ticket, which are the motivations of the renounce?*
4. *If it happens that you waiver a flight, for which you already bought the ticket and are entitled to refund, but you (or the person assigned to this task) do not ask the refund, which are the motivations of the renounce to refund?*
5. *If it happens that you completely change the date or the routes (or both) of a flight, which are the motivations of the change?*
6. *When you hold an air ticket including different routes and/or a round trip, if it happens that you change just a part of the ticket (e.g. you fly only to one or more destinations, but change the other/s, or you return earlier, or change one or more, but not all the dates) which are the motivations of the change?*
7. *Which of the following factors:*

- *the destination of the flight,*
- *the weekday of departure,*
- *the month of departure,*
- *your professional status within the company (institution)'s hierarchy,*
- *the reserved on-board class,*

influence the decision of:

- *Completely change the air ticket?*
- *Partially change the air ticket?*
- *Completely waiver the flight?*
- *Partially waiver the flight?*

The anonimity was, of course, explicitly guaranteed to all the participants, along with the guarantee that none of the provided information will be disclosed to anyone related to their company/institution.

The time of the interviewed varied a lot, from case to case, depending on how many questions were excluded for consistency with previous answers (e.g. if a business traveler stated that he never changes ticket, all the questions about changes were skipped). However, the time of the interview never exceeded 20 minutes, as all the respondents' professional status, reference company/institution and socio-demographic information were already known.

2.3 Description of participants and answers

Seven business travelers were selected to participate to the interview, based on the high number of times they fly for work, which is a characteristic surely related to the variables of interest (the more times a business traveler flies, the higher, theoretically, may be the probability that, sometimes, he changes or waivers or asks refund of the ticket). The participants have various job positions in different companies and institutions, with headquarters in various countries. Unfortunately, they are all males, as it is rare to know women frequently flying for work (maybe due to familiar reasons), and their age is rather homogeneous, as it is usual in friendship environments. Moreover, most of them are single, as it is often the case of people traveling a great part of their time. For sake of easiness of exposition, each participant is indicated with a letter.

Participant A is a 30 years old entrepreneur, with a degree in History. His company is located in Brazil and deals with export and sale of made-in-Italy clothing. He is single and lives in Italy. He flies to Brazil, to stay there at least 1 week and 3 weeks maximum, and comes back to Italy once a month, on average. He purchases air tickets by himself, or together with his associate. As the company is still a start-up, participant A is very careful in controlling expenses, so he prefers buying last-minutes offers, or low cost flights in low demand (thus lower cost) periods of the year.

He can fly whenever the ticket is cheaper because, being the owner of the company, he is absolutely autonomous in his working decision and travelling choices. He has never changed an air ticket, neither totally, nor partially, since he started the business activity, to avoid increases of travel expenses. Nonetheless, he said that, whether he made a change, it would be surely due to serious familiar reasons and the very few times, when he waivers a flight, it is because of motivations related to his origin family. Consistently, the destination of the flight, which is always the same for him, the month of departure, which is indifferent to him, the reserved on-board class, which is always the cheapest, do not influence his travelling decisions.

In fact, participant A declared that only the weekday of departure influences the probability that he partially changes the flight, because, within his family, familiar meetings and events are organized in the weekend and he does not want to miss them. Thus, if the cheapest round trip ticket provided a date of return in the weekend, he would change such a date, in order to attend to his family's event. When asked explicitly, participant A denied that his professional status influences his travelling choices.

Participant B is a 37 years old professional, with a degree in Tourism Economics. In particular, he is marketing and management advisor for the hotel industry. He is single, lives and works in Italy, he flies 0.5 times per month, on average, staying at the destination a variable length of time, from a few days to various weeks. He purchases air tickets by himself, mainly for short-haul national flights. When he changes ticket, he changes it totally, because of his clients' modified schedules. But when he waivers a flight, it can be due to various motivations: origin family needs, his clients' requests, hitches occurred during his activity. In fact, his work must be approved by hotels' owners and managers, thus an activity which he planned to accomplish in a certain time, often requires more days, to satisfy the client, therefore planned journeys must be postponed.

Often it happens that participant B renounces to ask a refund to which he is entitled, because he thinks that the procedures to obtain it require too long time, compared to the small amount of money involved. As he does not fly very often, he does not care too much for the travelling expenses, thus the on-board class has no influence on his travelling decisions. He said that also the month of departure is indifferent, as changes and waivers are determined by unpredictable snags. He thinks that also his job position is ininfluent on the probability of changing or renouncing to flights, but eventually it is not on his tendency not to ask refund, as the time he spend in consulting is so remunerative that he is not willing to divert it on the application procedure for refund, while, if he did a less remunerative job, he could evaluate such an aspect differently.

The weekday of departure and the destination of the flight are important factors in determining whether he changes ticket or not, in fact no hitch occurs to him during the weekend, although his clients can work, as it is usual in the tourism field, because he does not work in the weekend. The destination is influential because he has different clients in different cities, thus for him the city of arrival 'represents' the corresponding client.

Participant C is a 33 years old manager of a direct distribution channel for a multinational corporation, producing and selling fitness machines, with headquarters in Italy and branches in many different countries. He holds a master degree in Marketing and Communication, lives in Italy and is single. On average, he flies twice per month, primarily to develop effective sales practices and manage human resources in the foreign branches of the corporation. So he does international and intercontinental flights, departing from Italy, landing in capital cities where he usually stays for up to a week, than leaving for another country, where he stays up to another week, and so forth until the returns to Italy. In the corporation, where he works, there is an employee who purchases flights for all the (numerous) workers.

He often changes the air ticket completely either when some problem occurs anywhere along the channel, changing his priority, or when his superior assigns him a new duty, incompatible with the planned journeys, or when the sales manager he had to meet are no longer available at the prefixed date. The latter is also the motivation of partial changes: as he often plans to reach various destinations, if a manager in one of these destination is no longer available for the meeting, then he changes the ticket only for that date/destination. Participant C has never renounced to a flight, but he thinks that the only motivations which could lead him to give up a journey are health problems and very serious familiar reasons. If it happens, but it does generally not, that a refund, to which the corporation is entitled, is not asked, it is due to an oversight of the employee assigned to this task, as he has such a huge number of flights to manage for all the business traveler of the corporation, that may losing sight of some one.

Participant C declared that the class on board has no influence on his tendency to change or (non) waiver a flight. This answer is very interesting, because the travel management policy of the corporation is to purchase seats in economy class for all the flights lasting less than 10 hours, thus the duration of the flight and, consequently, the covered distance, is not important, at least in this case, in determining tickets' changes and refunds. On the contrary, his professional status, within the corporation's hierarchy, is very influential on the probability that he flies, changes the ticket or gets the refund, because he is not autonomous, as his superior can change his programs, and he must respect foreign managers' schedules and appointments, while, being himself a manager at a 'super-national' level, he must travel very often. The month of departure influences, in the experience of participant C, the tendency to change, both totally and partially, a ticket, but it is not relevant for waiving the flight. This fact depends especially on the seasonal dynamic of sales and it is true also for the weekday, that is said to determines the probability of partial change, but not that of total change, because the air tickets bought for participant C are usually multi-routes and he stays far from home long time, thus if he needs to change a date, it is only a partial change. For the same reason, the destination of the flight influences the tendency to make partial changes and ask partial refunds, but not total ones.

Participant D is a 38 years old parliamentary politician, working as responsible for international relations within his reference party. He holds a master's degree in Electronic Engineering, is single and lives in San Marino. On average, he flies 16 times per year, to a single destination, where he stays for a few days, mainly to attend to political conventions, to meet officials of international organizations and in general for institutional reasons. Thus, the air tickets he uses are purchased by the secretary of the involved institution, which belongs to the public administration.

He has never changed an air ticket, because the public institution schedules, and especially those related to political activities, are much more invariant, compared to those of private corporations. In fact, generally business travellers in the private sector are contemporarily involved in different working activities, with various clients and superiors changing programs, while political and institutional meetings, involving many people full time dedicated to that specific activity, hardly change. In fact, he waived a flight only once, because of adverse meteorological conditions. So he has no direct experience with refunds, but he said that the institutional secretaries purchase non-refundable tickets, in order to save money in times of spending review.

Participant E is a 33 years old area manager, in a multinational corporation, with headquarters in Italy, producing and selling tools and products for beverages. He has a degree in International Relations, is married, father of 3 little children and lives in San Marino. On average, he flies twice per month, mainly to Russia, but also to other European and north-eastern countries, to meet clients and agents, to prepare and finalize sales contracts. Generally he buys flights for multiple destinations, where he stays for some days.

He purchases air tickets by himself, through either online applications or a travel agency. He sometimes changes ticket, both partially and totally, because of modifications of his clients' schedules and snags, or due to unavailability of trade agents for the prefixed date. While at the beginning of his career, participant E made great efforts to travel as much as possible, to reach sales targets, obtain performance rewards and also because transfer periods are paid nearly the double, now, that he became rich and has a numerous family, he often renounce to flights, to spend more time with his wife and babies.

Like participant B, he rarely asks refunds to which he is entitled, because the time, required to apply for the refund, is too long, compared to the small amount of money involved, especially for economy class tickets. So he prefers to employ such time with his family or working for an higher amount of money.

Participant F is a 37 years old key account managers coordinator, in a multinational corporation producing packagings. He holds a master's degree in Electronic Engineering, is single and lives in Dubai, but is Italian. On average, he flies 8 times per month, mainly to meet and coordinate the key account managers, which work in the branches of the corporations, located worldwide, but also to deal with the most important clients and to solve emergencies. So he does flights with multiple routes, both international and, more often, intercontinental, as the headquarters of the corporation is in Austria. Usually he stays from a few days up to 3 weeks in each destination, before returning to the Emirates.

He buys the air tickets by himself, usually through a travel agency, sometimes directly online, in case of emergency, when he is abroad and has an urgency at home or if he finds low cost flights, which are not intermediated by the agency. He often changes ticket, both totally and partially, or renounce to the flight, due to modifications in clients' programs, to temporary unavailability of the key account manager he had to meet or to a sudden emergency, occurred in another country. He is not involved in the choice to ask or not refunds, as his assistant deals with this task and he does not control her, as the money is of the corporation and he does not gain nor loose anything anyhow.

Participant F said that the class on board and the month of the year are ininfluent on his travel behavior. With reference to the class of the seat, this answer implies that the covered distance and the duration of the flight too are not important, because the company for which he travels has a travel management policy similar to that of the corporation for which participant C works. It may be curious that also participant F, like participants A, declared that his professional status is irrelevant to explain his travelling choices. In fact, whether he was not in the position of being responsible for emergency and for coordinating managers, which are themselves frequent business travelers, thus subject to programs changes by clients, it could appear that he would have much less motivation for changing and renouncing to flights. However this answer is due to the fact that, also in other job positions he previously

held, he has always flown very frequently for business and with a very similar travel behavior.
Both the weekday of departure and the destination of the flights are judged unimportant in determining his tendency to change the ticket. In fact, he is used to fly for business during the weekend too, also because he often does flights lasting more than a day. Moreover, he travels more often to meet managers working for the same corporations, rather than to visit clients, so the way of working and scheduling, within different branches of the same company, is very similar, independently of their location. On the contrary, the destination and the weekday of departure influence, in the opinion of participant F, the probability that he renonces, both completely and partially, to the flight. But, in this case, the influence of the destination is due to local factors, like terrorist activities, epidemics and wars. While the effect of the weekday on his choice not to fly is due to the fact that, when he is abroad and should go back home for the weekend, but suddenly he has to plan a further flight, departing immediately after the weekend, for a destination not far from where he currently is, he prefers not to return home.

Participant G is a 38 years old professor of Bioengineering, he is single and lives in Belgium. On average, he flies 10 times per year, to a single destination, when it is not far, or to multiple destinations, when they are far from the departure place, but close one another. He usually stays at the destination for 1 or 2 weeks, to attend to scientific conferences, to hold lectures in other universities or to apply the results of his applied research on behalf of other countries, requiring it.
The air tickets that participant G uses are purchased and managed by the secretary of the public institution, commissioning his applications, or by the university. Thus, he has practically no room for changes and waives. In fact, he has never changed an air ticket, also because the public institutions and universities' schedules are practically invariant, as they normally involve scholars and experts coming from different parts of the world, so that chang-

ing plans would create many organizational problems. However, participant G said that he could waiver a flight in case of serious health problems. The institutional or university secretaries generally purchase non-refundable tickets for him, as they are cheaper.

2.4 Findings and guidelines for the quantitative study

The qualitative study described above yields a lot of information on the context of the present study. However these findings must be considered very cautiously, because of the methodological limits discussed in the first section. The main results of the study can be summarized as follows.

- The dimension, the level of development of the company, for which the business traveler works, and its economic condition, appear to greatly influence the tendency to both renounce to the flight and change ticket. In fact, the bigger and the more internationalized the corporation, the more developed its business, the more abundant its economic resources, the higher the likelihood that more expensive and flexible, so changeable and refundable, air tickets are purchased and that the business traveler often changes or renounces to the flight. Even in case that big corporations buy cheap and fixed tickets, it is more likely that changes and renounces happen, compared to small companies, at an early stage of development, with scarce economic resources. The latter are much more careful in controlling and limiting the travel expenses to the minimum, thus they buy low cost, non refundable and non changeable tickets, then strictly avoid changes and waivers.

- It emerged that the professional status of the business traveler is ininfluent on his 'flying behavior', in the sense that, within the set of business travelers which often take the plane, because the corporation where they work is multinational, or serves clients worldwide, or has providers in different countries, or the nature of its activity itself (e.g. export) makes frequent travelling necessary, the position of the traveler, within the company's hierarchy, does not make any difference in the likelihood that he changes or waivers flights.

- The reasons of changes and renounces to flights, which most frequently occurr in the answers of the participants to the interview, are related either to the family, or to working snags. The family seems the priority for business travelers with children or living still with their origin family. Familiar reasons appear to be the reason why, for some interviewed, the weekday of the departure is important, as they want to spend the weekend together with their family. Working motivations seem to prevail in single and independent workers. However, the familiar situation of the travelers appears not to be related to their age, at least within the very narrow age interval represented in this study. Moreover, familiar reasons seem more influential on the choice to renounce to the flight, rather than on that to change the ticket. Conversely, working motivations, primarily modification in programs due to the clients, the superior or the colleagues' changed schedules, appear to be more important in determining tickets changes, and especially partial changes for workers flying multiple routes, rather than in causing renounce to the flight.

- Another finding concerns the scarce convenience of asking refunds. Even in case the purchased ticket is refundable, thus more expensive than non refundable ones, it is rare that refund is asked, because the procedures to obtain it require too long time, compared to the small amount of money involved. Thus, this is a clear example of two different causal mechanisms, producing the same effect: small companies, with limited resources, do not ask the refund, because they buy low cost non refundable tickets; big and rich corporations, prefer employing their workers' time in more profitable occupations, rather than in applying for refunds. Therefore, the renounce to refunds may be consequence of both the scarcity and the abundance of company's resources.

- From the participants' answers, it also emerged that within the public institutions schedules are much less variable and snags much more rare, than in private companies. Thus, it may be more rare that a business traveler belonging to the public administration changes or waivers the air ticket, compared to who works for a private corporation.

- The fact that the class reserved on board resulted unimportant in determining the travel behavior for all of the participants, is especially interesting, considering that the travel management policy of some companies assigns the class based on the length of the flight, thus, substantially, on the covered distance. Therefore, it seems that the distance covered by the flight and its duration are not relevant. Moreover, this finding suggests that the importance of the destination city, in explaining the tendency to renounce to the flight or change ticket, is not due its distance from the departure city.

- Indeed, the found relevance of the city of arrival, for explaining both changes and renounces to flights, appears to be rather related to the purpose of the travel. Whether the worker makes the journey to meet clients, then the influence of the destination represents indeed the influence of the client located in (or near) that city. For example, if a client in Milan frequently changes programs or is especially subject to hitches, then the tendency of the business traveler, who flies to reach that client, to change the ticket will be higher for flights landing in Milan. In case the business traveler makes the journey to meet his collegues, or to attend to a convention or a congress, then the relevance of the destination corresponds rather to the impact, on the 'flying behavior', of location-specific events, like epidemics or extreme meteorological conditions. In fact, congresses and conventions' schedules hardly change, and the way of facing snaps and of following programs is rather homogeneous between branches of the same corporation.

- Finally, it resulted that the weekday of departure is much more important than the month of departure, for explaining the choice to change the ticket or renounce to the flight. This seems to be mainly due to family reasons and the crucial difference appears to be that between weekend and other days, as business travelers prefer to stay or return at home for the weekend, or simply relax without travelling.

Although, due to the methodological reasons discussed in section one, these findings cannot be directly translated into instructions for modelling the probability of non fly, and especially for choosing the explanatory variables among the available candidates, they can yield some hints, but above all they will help the interpretation of the subsequent estimates.

Maybe the most important suggestion, retrievable from this qualitative study, is that most of the available variables cannot be acritically taken for exogenous, as they are indeed 'generated' by unobserved variables, from which the dependent variables may be not independent, conditionally to the former. In particular:

1. the probability of non-fly seems to be influenced by the unobserved familiar situation of the traveler, which also affects the choice of the weekday of departure and of return;

2. the probability of changing ticket appears to be affected by the unobserved working purpose of the journey, which may be only partly attributable to the type of institution for which the business traveler works (public administration/private corporation);

3. besides the previously mentioned probability, it appears that also the destination city depends on the unobserved purpose of the journey;

4. the unobserved dimension, economic situation and stage of development and internationalization of the corporation (indeed observed only for a few tickets and affected by a great selection bias) seems crucial in explaining both the probability of change and that of renounce, as it appears to determine the affordable cost of the ticket, thus the on board class, whether the flight is low cost, whether round trip (cheaper than two one way tickets) and also the date of departure (flights in peak periods are more expensive);

5. the same unobserved variable may also influence the purchasing pattern: if it can be assumed that the companies more accustomed to recourse to a travel agency are more interested in innovative contracts, as they make more frequent use of intermediation and developed trust in the agency, it may also be concluded, in the light of what emerged from the interviews, that bigger client companies of Seneca should be more likely to subscribe the Target Buy;

6. the probability of refund seems caused by two different mechanisms, summarized in two opposite categories (abundant/scarce resources available to the company) of the same unobserved variable, leading to the same choice of non refund.

These hints must be integrated with the analysis of data, in order to find the most suitable statistical methodology to model the phenomenon of interest. Then, they must be used for a meta analysis of the results.

Chapter 3

Description of the dataset

3.1 The corporate database

The phenomenon analyzed in the present work is specifically related to the commercial activity of Seneca TMC, thus there is no extant literature guiding the choice of the methodology, nor any theory providing a conceptual framework. Therefore this thesis is necessarily data-driven: the data availability circumscribes the methodological possibilities, which can be usefully developed and applied, and the difficulties, presented by the dataset itself, require appropriate elaboration procedures. The database is corporate and is provided by Seneca.

The first problem, concerning the data, is the timing of the collection. In fact, in order to gather information appropriate to develop a solution to the business problem faced by Seneca, the company' s IT system needed to be upgraded and re-organized. This process required long time and allowed to collect suitable information only since the 1 January 2014. As also the elaboration of this thesis requires time and there are deadlines to be met, it was not possible to postpone the beginning of this work beyond August 2014. Thus, it was not possible to wait for at least a whole year observation period to become available. As a consequence, the sample at disposal for estimation purposes is relatively small, especially considering that most of the potential explanatory variables are categorical, so coded as dummy and naturally tending to erode degrees of freedom, and that it is not known how many of them should be included in the model(s).

Another consequence of not observing the phenomenon for a whole year, is that it is not possible to investigate the dynamic of the phenomenon, nor to detect eventual seasonal components (not to speak of trend and cycle). Indeed, some authors wrote that business travel is theoretically non-seasonal (Ritchie and Beliveau, 1974) or less affected by seasonality than leisure and visiting-friends-and-relatives tourism (Kulendran and Wilson, 2000). Nonetheless, in the practice, modelling seasonality can be useful for forecasting purposes (Kulendran and Witt, 2003), as business travel appear to actually follow a seasonal pattern, where the seasons are longer, equal to non-holiday periods (Swarbrooke and Horner, 2001). However, Swarbrooke and Horner (2001) highlighted that, contrarily to leisure tourism, business travel exhibits also a weekly dynamic, as it is likely that workers do not travel during the weekend. At least, this last aspect can be studied with the available data.

However, the data collection continues every day, at Seneca, thus in some months new data will be available. Unfortunately, when they will arrive, it will be too late for studying the dynamic pattern of business travel flights and increasing the degrees of freedom for the models estimation. Nonetheless, the complete nescience of the second part of the dataset, implied by this timing of the data collection, makes the forecasting experiment, which is properly the main focus of the present work, completely realistic. In fact, the second part of the database will be employed as forecasting sample for the out-of-sample evaluation of the forecasting performances of competing models.

Another relavant problem posed by the database derives from the fact that data strictly concerning the flights (namely: output of the ticket, departure date, date of issuance, advance booking, route, class, whether low cost,

airline, city of arrival, city of departure, number of routes, type of ticket) are automatically collected by the IT-system of Seneca, thus never missing, while the other data are collected manually, so often missing.
As this study is data-driven, it have to necessarily begin with an extensive description of data.

3.2 The composition of the dataset

The (first) database originally provided by Seneca is composed by observations on 48.085 air tickets, issued since the 18 April 2012 to the 21 August 2014, in fact some data, referring to the 2 years preceding the upgrade of the corporate IT system, have been manually reshaped and inserted within the dataset. The observed tickets, of which the 37% are one way and the 62% round trip (for the remaining this information is missing), concern 39.602 flight itineraries. These tickets, the 10% of which are low cost, were purchased by Seneca from 159 different airlines, on behalf of 146 different clients, including corporations, public administration and banks. The tickets were issued in the name of 15.095 different business travelers, for most of which, unfortunately, no information on sex, age and professional status is available. Retrieving this information has been very difficult, as the master data were collected in and until a period preceding the upgrade of the corporate IT system, when the data collection ended. Thus, master information on travelers not included in the main database is available, but most of the information on travelers observed in the latter is not.

The observed flight itineraries cover 6.398 different air routes, departing from 360 different cities, located in 119 countries, and arriving in 472 destinations, located in 134 countries worldwide. The difference in number between origin and destination places is due to both missing data and to the so called 'cross routes', which complicate a lot the analysis. In fact, the term refers to the practice, subject to penalties by the airlines, of buying various one way tickets, including multiple routes, of which only some are flown. For the aim of the present work, it is convenient to consider the non-flown routes as partial changes. The 55% of the tickets are national (concerning routes within Italy), the 35% are international (referring to flights within Europe) and the remaining intercontinental. Each one includes from 1 to 8 routes, covering a total distance varying from 34 km to 37.678 km.

From the original database, a dataset in wide format is constructed. Each row corresponds to an itinerary, which can be composed by multiple tickets, including changes and refunds. Itineraries for which the exit is unknown at the time of (the first) data delivery are excluded, but will be included in the out-of-sample dataset. Also those tickets lacking the itinerary key, which are impossible to connect to the corresponding itinerary, are deleted. Thus, this dataset is constituted by 39.389 itineraries. These data constitute the estimation sample for modelling the behavior of clients, or 'trajectories' of tickets related to the same itinerary.

3.3 The variables

The dataset structured for itineraries includes the dependent variable(s), constituted by the outcomes of air tickets. Possibly the most complex and crucial tasks, in the present work, is finding a correct specification for such a dependent variable, consistent with the dinamic and logical structure of business travelers' flying decision-making process, appropriate for meaningfully modelling the relationships between the phenomenon and the explanatory variables, as well as computationally feasible, not only in the modelling phase, but also in the implementation of the resulting solution within the everyday business practice of the TMC.

The first step in the search for the optimal specification of the dependent variable, which is described in the following chapter, is analizing data, but this analysis itself requires a provisional specification of this same variable. Thus, in the preliminary analysis, the following specifications are considered:

- For Y_1: $\Omega_{Y_1} = \{FLOWN, NONflown\}$; $R_{Y_1} = \{0; 1\}$.
- ForY_2: $\Omega_{Y_2} = \{FLOWN, CHANGED, NONflown\}$;
  $R_{Y_2} = \{1; 2; 3\}$.
- ForY_3: $\Omega_{Y_3} = \{FLOWN, CHANGED, NONflown, REFUNDED\}$;
  $R_{Y_3} = \{1; 2; 3; 4\}$.
- ForY_4:
  $\Omega_{Y_4} = \{FLOWN, PartCHANGED, NONflown, REFUNDED, TotCHANGED\}$;
  $R_{Y_4} = \{1; 2; 3; 4; 5\}$.
- ForY_5:
  $\Omega_{Y_5} = \{FLWN, PartCHNG, NONflwn, TotREFUND, TotCHNG, PartREFUND\}$;
  $R_{Y_5} = \{1; 2; 3; 4; 5; 6\}$.

Where "FLOWN", "NON-flown", "Partially CHANGED", "Totally REFUNDED", "Totally CHANGED" and "Partially REFUNDED" refer to the first issued ticket for the corresponding business travel. Indeed, the business traveler can change ticket more than once, for the same travel, and ask more than one refund, both partial and total.

The candidate explanatory variables for Y_f, of which some are continuous, some dichotomous and some nominal, are listed in table 1 below (along with their levels, in case of nominal variables).

Table 3.1: Explanatory Variables

Table 1: Explanatory Variables

Explanatory Variables					
NOMINAL			**N (Ntot= 38,889)**	**% Changes Marginal% = 10.358%**	**N Missing**
Variable	**Levels**	**Encoded as:**			
TYPE OF CLIENT COMPANY	Corporate	Dummy (= 1 if Corporate, = 0 otherwise)	27,940	10%	9
	Bank	Dummy (= 1 if Bank, = 0 otherwise)	515	65%	
	Public Administration	Dummy (= 1 if Public Administration, = 0 otherwise)	10,425	8%	
PROFESSIONAL STATUS OF TRAVELER	Entrepreneur	Dummy (= 1 if Entrepreneur, = 0 otherwise)	14	79%	36,350
	CEO/President	Dummy (= 1 if CEO/President, = 0 otherwise)	50	58%	
	Manager	Dummy (= 1 if Manager, = 0 otherwise)	1,073	69%	
	MidManager	Dummy (= 1 if MidManager, = 0 otherwise)	556	65%	
	Employee	Dummy (= 1 if Employee, = 0 otherwise)	700	72%	
	Factory Worker	Dummy (= 1 if Factory Worker, = 0 otherwise)	1	100%	
	Professional	Dummy (= 1 if Professional, = 0 otherwise)	9	89%	
	Professor	Dummy (= 1 if Professor, = 0 otherwise)	32	84%	
	Host	Dummy (= 1 if Host, = 0 otherwise)	13	69%	
	Magistrate	Dummy (= 1 if Magistrate, = 0 otherwise)	3	67%	
	Army General	Dummy (= 1 if Army General, = 0 otherwise)	19	63%	
	Army Graduate	Dummy (= 1 if Army Graduate, = 0 otherwise)	30	57%	
	Army Inferior Officer	Dummy (= 1 if Army Inferior Officer, = 0 otherwise)	19	68%	
	Army Superior Office	Dummy (= 1 if Army Superior Office, = 0 otherwise)	10	50%	
	Army NCO	Dummy (= 1 if Army NCO, = 0 otherwise)	34	47%	
TYPE OF AIR TICKET	National	Dummy (= 1 if National, = 0 otherwise)	21,273	11%	0
	International	Dummy (= 1 if International, = 0 otherwise)	14,013	10%	
	Intercontinental	Dummy (= 1 if Intercontinental, = 0 otherwise)	3,603	10%	
ROUTE	to be aggregated	Dummy (= 1 if equal to that level, = 0 otherwise)			0
NUMBER OF ROUTES	1	Dummy (= 1 if equal to 1, = 0 otherwise)	10,917	15%	

	2	Dummy (= 1 if equal to 2, = 0 otherwise)	23,165	9%	0
	3	Dummy (= 1 if equal to 3, = 0 otherwise)	729	7%	
	4	Dummy (= 1 if equal to 4, = 0 otherwise)	3,846	8%	
	5	Dummy (= 1 if equal to 5, = 0 otherwise)	114	16%	
	6	Dummy (= 1 if equal to 6, = 0 otherwise)	114	11%	
	7	Dummy (= 1 if equal to 7, = 0 otherwise)	3	33%	
	8	Dummy (= 1 if equal to 8, = 0 otherwise)	1	0%	
AIRLINE	ALITALIA_CAI_SPA	Dummy (= 1 if equal to that level, = 0 otherwise)	21,838	11%	0
(the 8 most numerous	LUFTHANSA		2,615	10%	
covering the 80% of the sample)	EASYJET		1,642	6%	
	RYANAIR		1,486	7%	
	BRUSSELS_AIRLINES		1,260	8%	
	AIR_FRANCE		939	11%	
	MERIDIANA_FLY		905	10%	
	BRITISH_AIRWAYS		601	11%	
	Other airlines		7,603	10%	
CLASS	First	Dummy (= 1 if First, = 0 otherwise)	853	12%	0
	Business	Dummy (= 1 if Business, = 0 otherwise)	2,180	20%	
	Economy	Dummy (= 1 if Economy, = 0 otherwise)	35,843	10%	
	Premium Economy	Dummy (= 1 if Premium Economy, = 0 otherwise)	12	25%	
SPATIAL	to be aggregated	Dummy (= 1 if equal to that level, = 0 otherwise)			0
DEPARTURE COUNTRY					
DEPARTURE CIYY					
ARRIVAL COUNTRY					
ARRIVAL CITY					
CONTINUOUS/count					
COVERED DISTANCE (KM)		continuous variable			300
TRAVELER'S AGE		count variable			30,254
ADVANCE BOOKING (DAYS)		count variable			0
(NON_FLOWN DISTANCE)		continuous variable			0
TEMPORAL					
DATE OF ISSUANCE	Mon	Dummy (= 1 for Monday, = 0 otherwise)	7,101	11%	
	Tue	Dummy (= 1 for Tuesday, = 0 otherwise)	7,790	12%	

	Wed	Dummy (= 1 for Wednesday, = 0 otherwise)	7,915	10%	0
	Thu	Dummy (= 1 for Thursday, = 0 otherwise)	7,488	10%	
	Fri	Dummy (= 1 for Friday, = 0 otherwise)	7,879	10%	
	Sat	Dummy (= 1 for Saturday, = 0 otherwise)	500	6%	
	Sun	Dummy (= 1 for Sunday, = 0 otherwise)	216	9%	
DATE OF DEPARTURE	Mon	Dummy (= 1 for Monday, = 0 otherwise)	8,590	9%	0
	Tue	Dummy (= 1 for Tuesday, = 0 otherwise)	7,965	10%	
	Wed	Dummy (= 1 for Wednesday, = 0 otherwise)	7,486	12%	
	Thu	Dummy (= 1 for Thursday, = 0 otherwise)	5,597	11%	
	Fri	Dummy (= 1 for Friday, = 0 otherwise)	4,426	11%	
	Sat	Dummy (= 1 for Saturday, = 0 otherwise)	1,505	9%	
	Sun	Dummy (= 1 for Sunday, = 0 otherwise)	3,320	8%	
DATE OF RETURN	Mon	Dummy (= 1 for Monday, = 0 otherwise)	2,876	8%	12,936 (OW)
	Tue	Dummy (= 1 for Tuesday, = 0 otherwise)	3,723	8%	
	Wed	Dummy (= 1 for Wednesday, = 0 otherwise)	4,984	7%	
	Thu	Dummy (= 1 for Thursday, = 0 otherwise)	5,511	8%	
	Fri	Dummy (= 1 for Friday, = 0 otherwise)	5,908	9%	
	Sat	Dummy (= 1 for Saturday, = 0 otherwise)	1,417	9%	
	Sun	Dummy (= 1 for Sunday, = 0 otherwise)	1,551	12%	
BINARY					
TRAVELER' SEX	Male	Dummy (= 1 if Male, = 0 otherwise)	28,073	13%	10,029
	Female		788	63%	
ONE WAY /ROUND TRIP	One Way	Dummy (= 1 if Round Trip, = 0 otherwise)	13,051	14%	4
	Round Trip		25,834	8%	
LOW COST FLIGHT	Low cost	Dummy (= 1 if Low Cost, = 0 otherwise)	4,123	8%	0
	Non Low cost		34,766	11%	
TARGET BUY CONTRACT	Target Buy		5,603	9%	0
	Non Targer Buy	Dummy (= 1 if Target Buy, = 0 otherwise)	33,286	11%	

Though the potential explanatory variables are numerous, it is likely that some of them encompass at least part of the information carried by others. In particular:

- some levels of the variable ’professional status’ are observable only in certain type of client companies (e.g. professional status in the army and magistrate are only observable in public administration);
- the variable ’route’ includes ’departure city’ and ’arrival city’, which in turn implicitly include ’departure country’ and ’arrival country’ respectively. Thus, it partially encompasses also the ’type of air ticket’ and ’one way/round trip’;
- the variable ’covered distance’ is strictly related to ’route’, ’number of routes’, ’one way/round trip’ and ’type of air ticket’;
- the ’traveler’s age’ can be related to the ’professional status’ of the business traveler;
- if considered jointly, ’date of issuance’ and ’date of departure’ encompass the information carried by ’advance booking’, which equals to the number of days elapsed between them.

Therefore, it is necessary to choose the best predictor between ’equivalent’ variables or to extract common factors from variables encompassing one another. The second solution would make results very hardly interpretable and require more complicated elaboration for implementation in the business activity of the TMC, therefore it is excluded. Next chapter explains how explanatory variables are chosen, when dealing with the models specification.

3.4 Missing data and selection bias

As beckoned above, no missing data is present in automatically collected variables, namely:

- output of the ticket,
- date of departure,
- date of issuance,
- advance booking,
- route,
- class,
- whether low cost,
- airline,
- city of departure,
- country of departure,
- city of arrival,
- country of arrival,
- number of routes,
- type of ticket.

While the number and percentage of missing data in each manually collected candidate explanatory variable are listed in table 2 below.

Table 3.2: Missing data in explanatory variables

ES (1/1/2014 – 7/31/2014)	*N^o*	%	Chi-Square
Date of Return	$12,936$	33%	0.261
Coveblue Distance (KM)	300	0.77%	3.573
Age	$30,254$	78%	0.074
Sex	$10,029$	26%	0.999
Professional Status	$36,350$	93%	3.909
Traveler	4	34%	0.741
Type of client	9	0.02%	17.04
38,782 flights			
FS (8/1/2014 – 1/10/2015)			
Ticket price	$19,974$	68%	
Ticket fare	$29,144$	99.6%	
29,252 flights			

$Chi - square$ critical value at 1% = 29.14

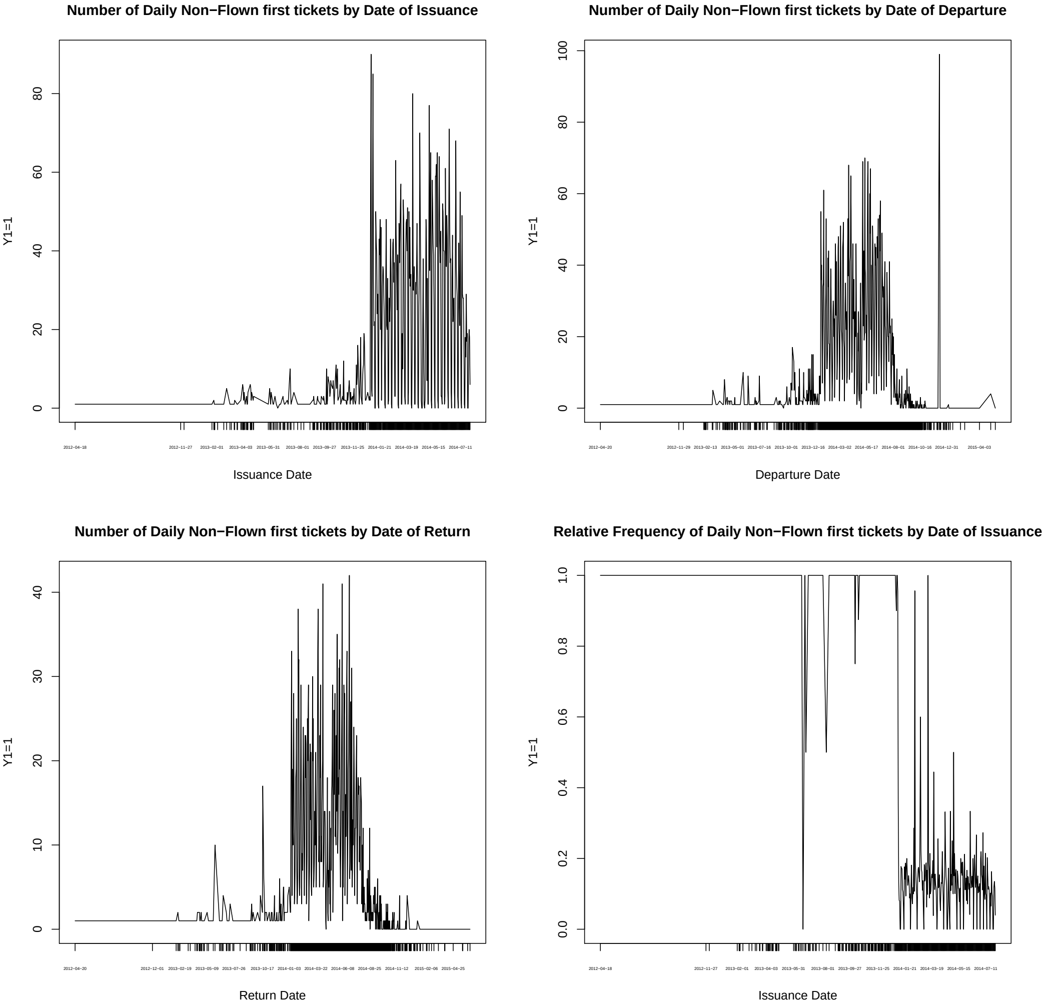
With reference to the estiation sample (ES), while the number of missing data in variables: Target Buy, Covered Distance, OW/RT and Type of client appear to be negligible, the number of non available data in the remaining variables is worrying, especially for the Professional Status and Age of the business traveler. The percentages, reported in table 2, concerning the latter, suggest to exclude them from the analysis, but theoretically these variables should have a great influence on the probability of non-departure.

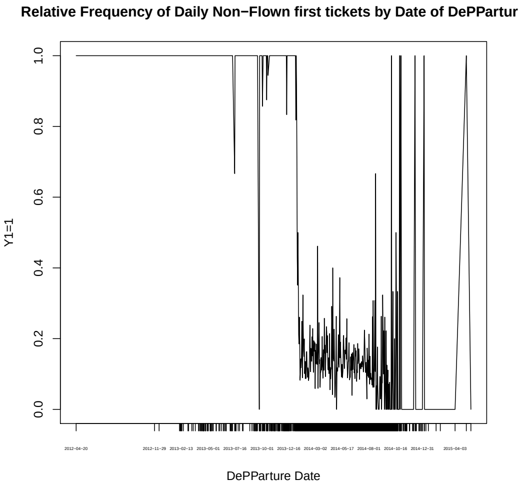
In fact, it can be hypothesized that the lower the professional status of the traveler, in his company's hierarchy, the lower the probability that he can postpone or avoid a journey, which is likely to have been imposed by his boss, and vice versa. Moreover, it may be that the age of the traveler can approximate his professional status, as it seems probable that the older the worker, the more career he did, the higher his status in the hierarchy, the higher the probability that he can waive or change the journey. The number of observed tickets for which the information on both the professional status and the age are missing is 28,602 (the 73%).

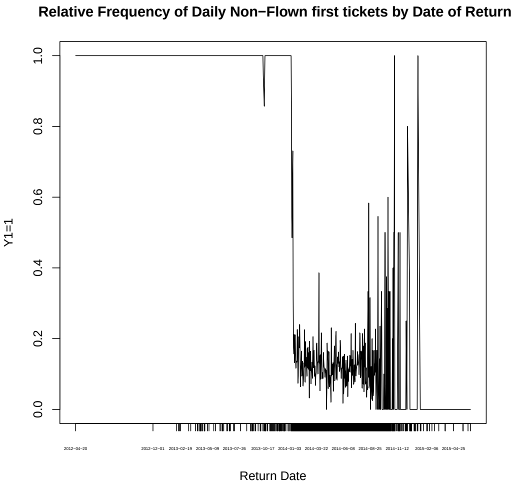
Therefore, in order to verify if it is meaningful to perform a separate variables selection (see next chapter), including only rows for which these appearently importan variables are observed, it is useful to look at the chi-squared test on the differences in relative frequencies of the dependent variable, between missing and non-missing data in exogenous variables, to detect the presence of selection bias (as e.g. in: Berger and Exner, 1999; McNichols and O'Brien, 1997) .
The values of chi-squared statistics, reported in table 2, allow to accept the null hypothesis of no selection bias for all of the variables, with a P-value lower than 1%. Thus, these variables can be used to calculate the conditional probability of non-flights, at least for tickets for which they are observed. But, to avoid favoring the significance of variables with less missing data (allowing to estimate models with more degrees of freedom), a separate selection on complete observations it is made.

As previously mentioned, most of the data are automatically collected by the IT system of the TMC since the 1st January 2014, but some data on the two previous years were manually inserted in the database. Thus, it is possible to visualize a relatively long historical dynamic of Y_f, based on either the date of issuance of the air ticket, or that of departure, or that of return. The dynamic pattern of the dependent variable clearly differs, as a function of the chosen type of date (as shown in the following plots). However, whether data anterior to 2014 are exploitable, is doubtful, as it appears from the plots below.

Figure 3.1: Dynamic of Non-Flown first tickets







Although looking at the plots of the daily number of non-flown first tickets it is not that obvious, from those of their relative frequency it is clear that data collected before the 1st January 2014 are affected by a serious selection bias. In fact, the 99% of tickets issued before 2014, namely 470 over 475, are not flown. Considering that only the 15% of tickets issued since the 1st January 2014 to the 21th August 2014 are not flown, namely 5,787 over 38,914, it appears necessary to drop observations on the previously issued tickets. Excluding this 1.2% of observations deprives the possibility of following the seasonal dynamic of the phenomenon, an information that should have been important. Nonetheless, the bias caused by the inclusion of these observations is expected to be definitely greater than that which could be caused by the lack of modelled seasonality.

Unfortunately, a similar selection bias affects also the variable 'client company'. In 34,136 cases over 39,389, equal to the 87%, the field concerning the client company is empty. In 33,973 cases, the 86%, the business traveler is unidentifiable. Furthermore, only for the 6% of the tickets issued before 2014 the information about the client company is missing (and for the 57%

of the tickets issued in 2014). Of the remaining 94%, the 99% are not flown. While, in the whole database, 5,006 first tickets over 5,253, for which the client company is known, equal to the 95% are not flown. Limiting the calculus to 2014, this percentage stays at 95%. Only the 4% of tickets issued before 2014 miss the identification of the corresponing business traveler. Of the remaining 96%, the 71% are not flown.

These percentages suggest that a separate analysis should be conducted for tickets for which the client company is identifiable. Thus, a further variables selection is performed only on observations complete with reference to this information, using variables descrribing the past behavior of the company and the business traveler. The first are intended to be indicators of the company's business, stage of development, economic condition and internationalization degree. The second is meant to be a proxy for the traveler's private life, family situation and professional role.

Among these 'past−behavior' variables, only a few resulted significant (see next chapter) and are reported here:

- NflightsC: number of flights purchased by the company (during the period spanned by the estimation sample).

- OccTravC: dummy variable, equal to 1 if the company requres business flights rarely (with a frequency less than the tird quartile), 0 otherwise.

- OccCh: dummy variable, equal to 1 if the traveler changed tickets rarely (with a frequency less than the tird quartile), 0 otherwise.

- OccTrav: dummy variable, equal to 1 if the worker flies rarely (with a frequency less than the tird quartile), 0 otherwise.

- AverTrav: : dummy variable, equal to 1 if the frequency of flights, booked in the name of the traveler, is included in the second or tird quartile), 0 otherwise.

- FreqWai: dummy variable, equal to 1 if the traveler cancels the flight often (with a frequency included in the first quartile), 0 otherwise.

Indeed, variables OccCh and FreqWai are tautological for descriptive purposes, but crucial in prediction, as shown in the final chapter.

With reference to the forecasting sample (FS in table 1.2), the too many missing data about the cost of tickets and their fare make impossible to estimate the expected cost of flights, to provide an actual (not based on approximation) economic evaluation of the predictors' performance and requires the development of a different method to assess the economic performance of what will be the proposed solution.

3.5 Descriptive statistics

Once removed data collected prior to 2014 and those presenting mistakes, the sample includes 38,782 observations. The sample temporal distribution of air tickets purchases can be viewed as a function of the date of issuance, date of departure and date of return. It is shown in figure 2.

Figure 3.2: Sample temporal distribution of air tickets purchases

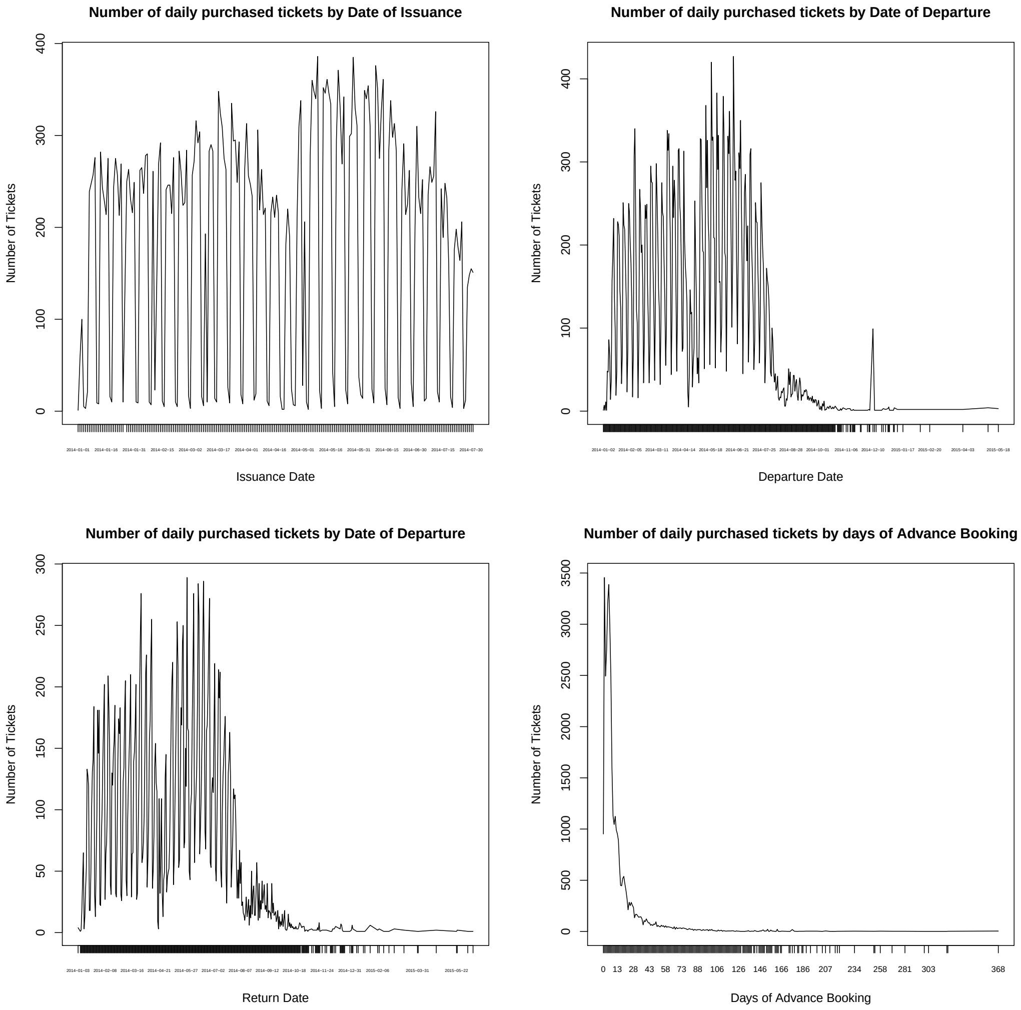


Table 3.3: Sample temporal distribution of air tickets purchases

of Issuance	n	rel. freq.
Fri	7,879	0.20
Tues	7,790	0.20
Wed	7,915	0.20
Thu	7,488	0.19
Mon	7,101	0.18
Sat	500	0.01
Sun	216	0.005

Month of Issuance	n	rel. freq.
May	7,073	0.18
Jun	6,143	0.16
Mar	5,874	0.15
Jul	5,016	0.13
Jan	4,970	0.13
Apr	4,940	0.13
Feb	4,873	0.12

of Departure	n	rel. freq.
Mon	8,590	0.22
Tues	7,965	0.20
Wed	7,486	0.19
Thu	5,597	0.14
Fri	4,426	0.11
Sun	3,320	0.08
Sat	1,505	0.04

Month of Departure	n	rel. freq.
Jun	7,040	0.18
May	6,460	0.17
Mar	5,713	0.15
Jul	5,252	0.13
Apr	4,685	0.12
Feb	4,571	0.12
Jan	3,243	0.08
Aug	1,046	0.03
Sep	615	0.02
Dec	120	0.003
Oct	124	0.003
Nov	20	0.0005

of Return	n	rel. freq.
Fri	5,908	0.15
Thu	5,511	0.14
Wed	4,984	0.13
Tues	3,723	0.10
Mon	2,876	0.07
Sat	1,417	0.04
Sun	1,551	0.04

Month of Return	n	rel. freq.
Jun	4,621	0.12
May	4,143	0.11
Mar	3,563	0.09
Jul	3,593	0.09
Apr	3,104	0.08
Feb	3,167	0.08
Jan	1,981	0.05
Aug	943	0.02
Sep	600	0.01
Oct	177	0.004
Dec	36	0.001
Nov	42	0.001

The plot of the number of daily purchased tickets by the days of advance booking completes the frame. Although the distribution of purchases by date of issuance is quite homogeneous, the main peaks are observed in the months of May and June, while there seems to be no relation between the week day of issuance and the number of issued tickets, except for the fact that most of the tickets are issued during working days, as can be expected, considering the business purpose of all the flights.

The plot for the dates of departure is more jagged. The main peaks are observed in May and June, in fact most of the tickets are issued less than 10 days before the date of departure. It appears that most of the business travelers tend to depart at the beginning of the week, especially on Monday, Wednesday and Tuesday, while it is rare that they fly on Sunday, as can be expected. Most of the observed tickets provide return in May and June, again, and in the second half of the week, especially on Friday, Thursday and Wednesday, suggesting that most of the business sojourn lasts half a week. While it is rare that travelers come back on Saturday.

The spatial distribution of the sample, by country of departure and country of arrival is shown in figure 3.

Figure 3.3: Sample spacial distribution of air tickets purchases

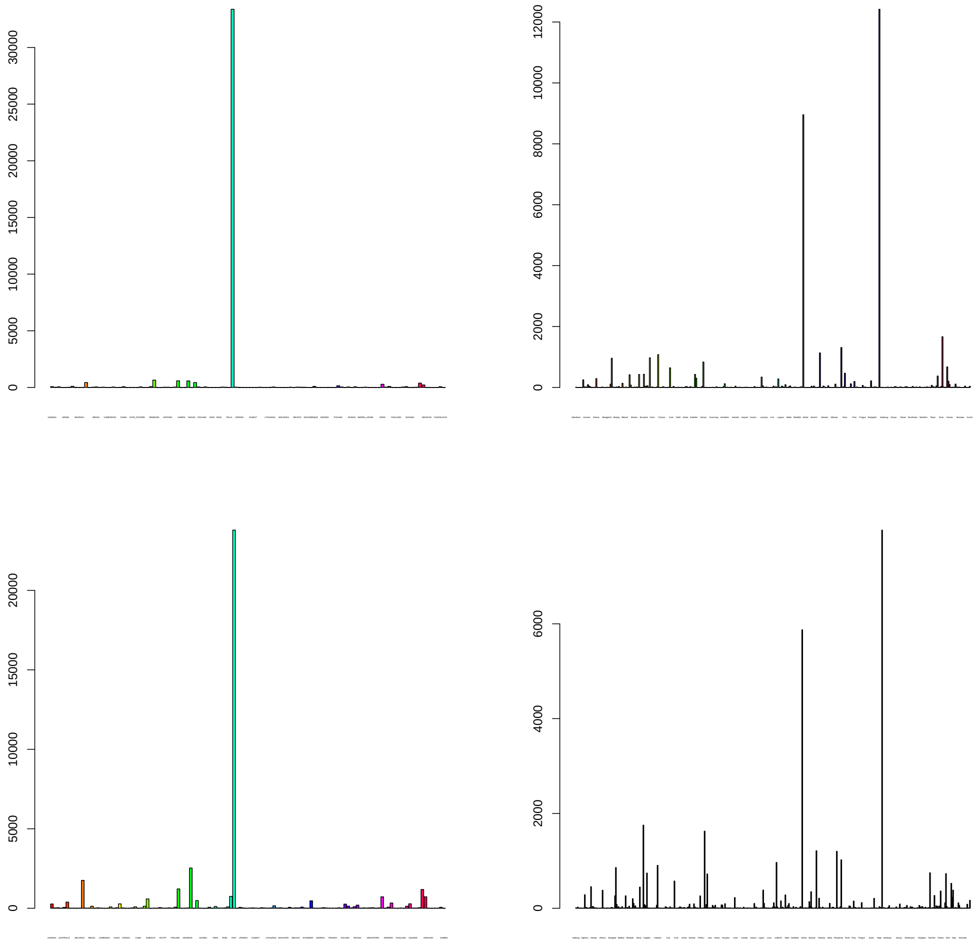


Table 3.4: Sample spacial distribution of air tickets purchases

Country of Departure (20 most freq.obs.)	n	rel. freq.
ITALY	33372	0.858
DENMARK	647	0.016
FRANCE	584	0.015
GERMANY	574	0.014
BELGIUM	433	0.011
GREECE	429	0.011
UNITED_KINGDOM	381	0.009
SPAIN	281	0.007
UNITED_STATES	221	0.005
POLAND	149	0.003
AUSTRIA	102	0.002
NETHERLANDS	93	0.002
SWITZERLAND	90	0.002
ALBANIA	70	0.001
TURKEY	69	0.001
CZECH_REPUBLIC	66	0.001
CHINA	65	0.001
YUGOSLAVIA	64	0.001
RUSSIAN_FEDERATION	57	0.001
ARGENTINA	54	0.001

City of Departure (20 most freq. obs.)	n	rel. freq.
Rome	12419	0.319
Milan	8956	0.23
Turin	1666	0.0428
Palermo	1311	0.0337
Naples	1134	0.0292
Catania	1078	0.0277
Cagliari	973	0.025
Bari	958	0.0246
Genoa	835	0.0215
Venice	675	0.0174
Copenhagen	645	0.0166
Paris	467	0.012
Brussels	432	0.0111
Florence	428	0.011
Brindisi	425	0.0109
Bologna	417	0.0107
Trieste	375	0.0096
Lamezia_Terme	340	0.0087
Frankfurt	305	0.0078
Athens	291	0.0074

Country of Arrival (20 most freq. obs.)	n	rel. freq.
ITALY	23791	0.612
GERMANY	2531	0.0651
BELGIUM	1756	0.0452
FRANCE	1218	0.0313
UNITED_KINGDOM	1182	0.0304
ISRAEL	749	0.0193
SPAIN	722	0.0186
UNITED_STATES	723	0.0186
DENMARK	586	0.0151
GREECE	483	0.0124
NETHERLANDS	470	0.0121
AUSTRIA	389	0.01
SWITZERLAND	332	0.0085
TURKEY	281	0.0072
CHINA	277	0.0071
ALBANIA	270	0.0069
POLAND	261	0.0067
RUSSIA	198	0.005
LUXEMBOURG	156	0.004
BRAZIL	130	0.0033

City of Arrival (20 most freq. obs.)	n	rel. freq.
Rome	7977	0.205
Milan	5874	0.151
Brussels	1754	0.0451
Frankfurt	1627	0.0418
Naples	1214	0.0312
Palermo	1201	0.0309
Paris	1022	0.0263
London	967	0.0249
Catania	906	0.0233
Bari	860	0.0221
Tel_Aviv_Yafo	749	0.0193
Cagliari	743	0.0191
Turin	731	0.0188
Genoa	724	0.0186
Copenhagen	575	0.0148
Venice	526	0.0135
Amsterdam	456	0.0117
Brindisi	447	0.0115
Lamezia_Terme	384	0.0098
Vienna	383	0.0098

As in one way tickets the arrival city/country for the way-forward equals the departure city/country for the way-back, considering that Seneca is located in Italy, thus many of its clients have headquarters in Italy, it is not surprising that Italy and italian cities are the most observed locations. Table 5 shows only places observed at least 4 times, the complete table is available upon request.

Table 6 shows the composition of the sample by route.

Figure 3.4: Sample composition by route

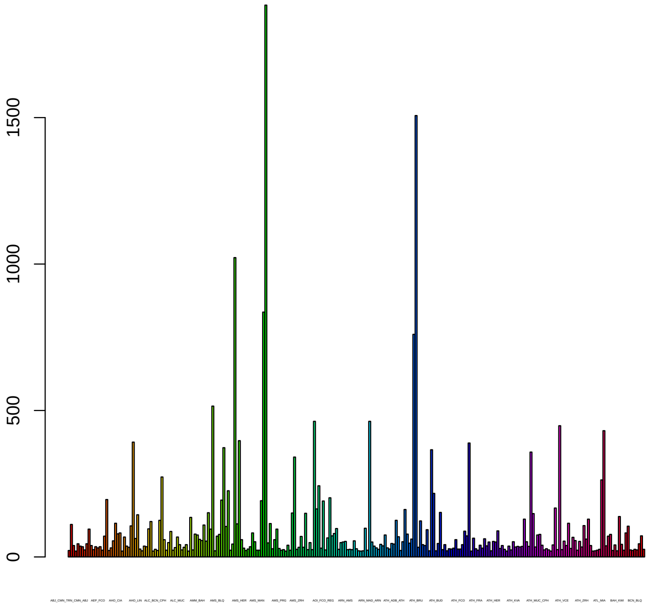


Table 3.5: Sample composition by route

Route (20 most freq. obs.)	n	rel. freq.
FCO/LIN/FCO	1884	0.0664
LIN/FCO/LIN	1507	0.0531
FCO/FRA/FCO	1022	0.036
FCO/LIN	836	0.0294
LIN/FCO	760	0.0268
FCO/BRU/FCO	515	0.0181
FCO/TLV/FCO	463	0.0163
GOA/FCO/GOA	463	0.0163
PMO/FCO/PMO	448	0.0158
TRN/FCO/TRN	431	0.0152
FCO/GOA/FCO	397	0.014
CAG/FCO/CAG	392	0.0138
MXP/BRU/MXP	389	0.0137
FCO/CDG/FCO	373	0.0131
LIN/NAP	366	0.0129
NAP/LIN	358	0.0126
FCO/PMO/FCO	341	0.012
CTA/FCO/CTA	273	0.0096
TRN/FCO	263	0.0092
FCO/TRN/FCO	243	0.0085

The variable 'route' is more heterogeneous than the variables of places of departure and arrival, because the same destination can be reached through different routes, from the same departure city, but also because of 'crossed tickets'. The above tables show only the 20 most frequently observed routes or places, to save space. The full tables are available upon request.

More synthetic variables, describing a characteristics of the path between the city of departure and that of arrival, are constituted by the number of routes and the covered distance.

Figure 3.5: Sample composition by number of routes

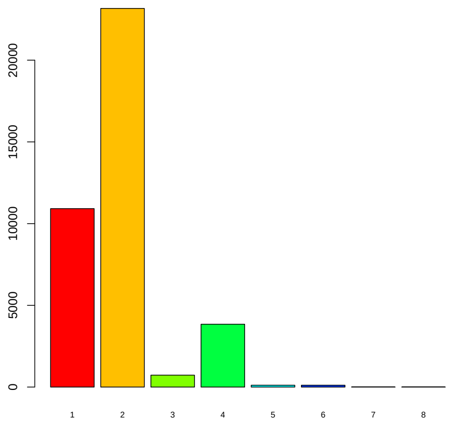


Figure 3.6: Sample composition by covered distance (Km)

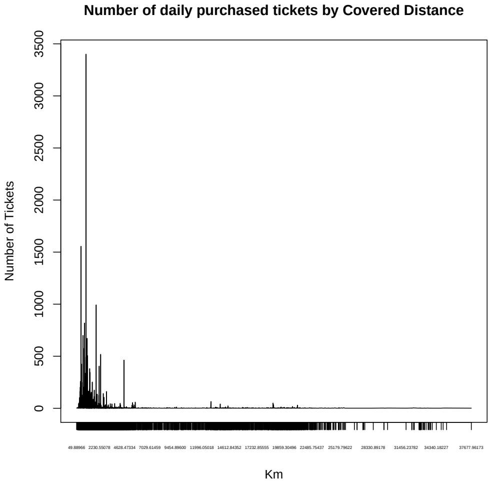


Table 3.6: Sample composition by number of routes (Km)

N. Routes	n	Rel.Freq
2	23,165	0.601
1	10,917	0.28
4	3,846	0.10
3	729	0.02
5	114	0.003
6	114	0.003
7	3	0.00008
8	1	0.00003

Table 3.7: Sample compositionby by covered distance (Km)

Covered Distance	Km
Max	37,700
Median	1,060
Mean	2,260
Min	49.9
Var.	14,800,000
Sd.	3,840
Var. Coeff.	1.70

Most of the observed flights include an only stopover, flights including more than 4 stopovers are very few, possibly related to 'crossed tickets'. Most of the flights cover a distance shorter than 5,000 Km. The 4 outstanding peaks in figure 6 correspond to flights covering: 940 Km, 470 Km, 1,918 Km and 798 Km of distance, in decreasing order of relative frequency.

Consistently with the fact that most of the flights depart from and arrive in an Italian city, most of the observed air tickets are national. Round trip itineraries are nearly the double of one-way ones, that can be explained by the lower cost of the first ones, ceteris paribus. These informations are shown in figures 5 and 6.

Figure 3.7: Sample composition by ticket type

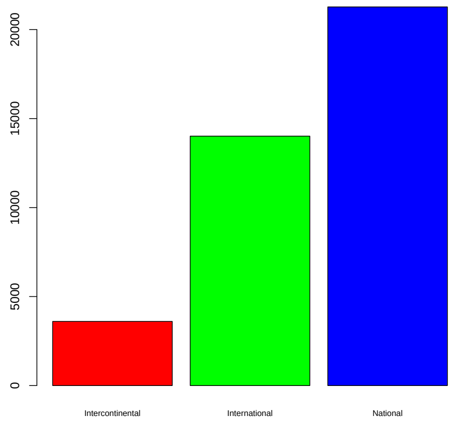


Figure 3.8: Sample composition by type of itinerary

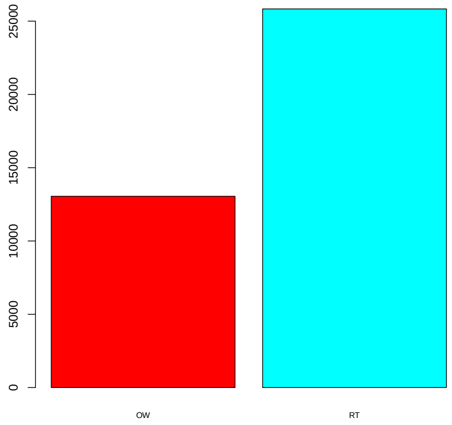


Table 3.8: Sample composition by ticket type

Type of ticket	n	Rel.Freq
Intercont	3, 603	0.093
Internat	1, 4013	0.360
National	21, 273	0.547

Table 3.9: Sample composition by type of itinerary

Type of itinerary	n	Rel.Freq
OW	13051	0.336
RT	25834	0.664

Figure 3.9: Sample composition by aircompany

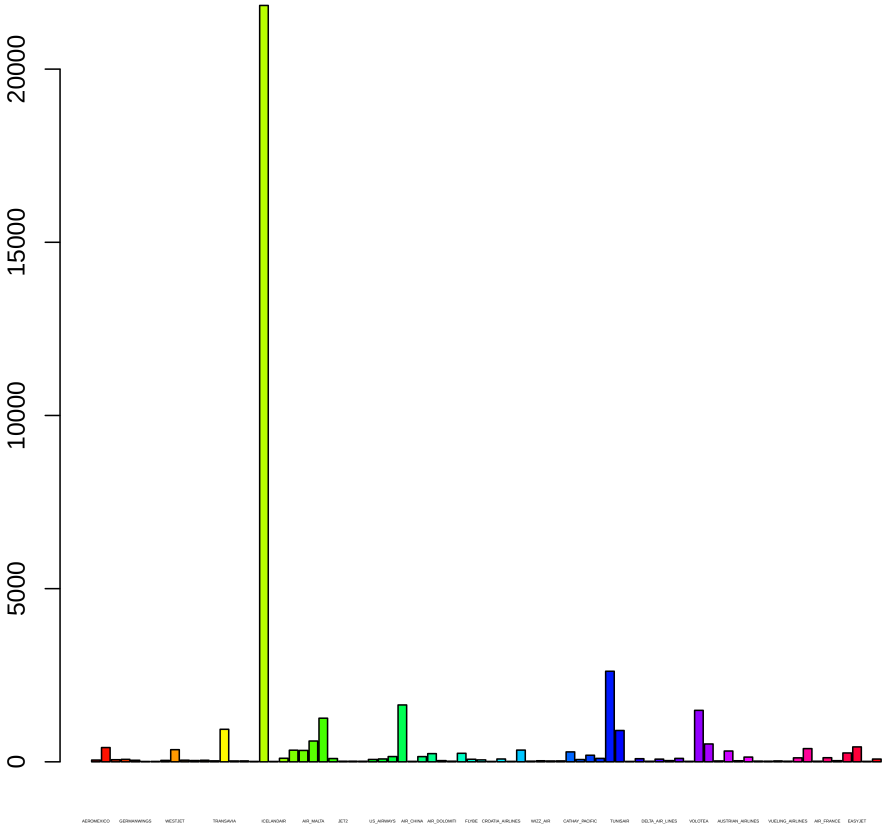


Table 3.10: Sample composition by aircompany

Airline (20 most freq. obs.)	n	Rel.Freq
ALITALIA_CAI_SPA	21, 838	0.5615
LUFTHANSA	2, 615	0.0672
EASYJET	1, 642	0.0422
RYANAIR	1, 486	0.0382
BRUSSELS_AIRLINES	1, 260	0.0324
AIR_FRANCE	939	0.0241
MERIDIANA_FLY	905	0.0232
BRITISH_AIRWAYS	601	0.0154
SCANDINAVIAN_AIRLINES	516	0.0132
VUELING_AIRLINES	431	0.011
AEGEAN_AIRLINE_S.A	412	0.0105
TURKISH_AIRLINES	382	0.0098
AIR_BERLIN	351	0.009
IBERIA	339	0.0087
AUSTRIAN_AIRLINES	336	0.0086
BLUE_PANORAMA_AIRLINES_S	329	0.0084
SWISS_INTERNATIONAL	312	0.008
KLM_ROYAL_DUTCH_AIRLINES	287	0.0073
VOLOTEA	256	0.0065
ETIHAD_REGIONAL	246	0.0063

Figure 3.10: Sample composition by Low Cost flights

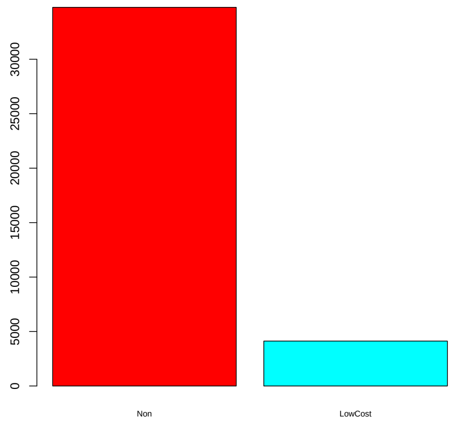


Table 3.11: Sample composition by Low Cost flights

	n	Rel.Freq
Non-low cost	34,766	0.894
Low Cost	4,123	0.106

Figure 3.11: Sample composition by class

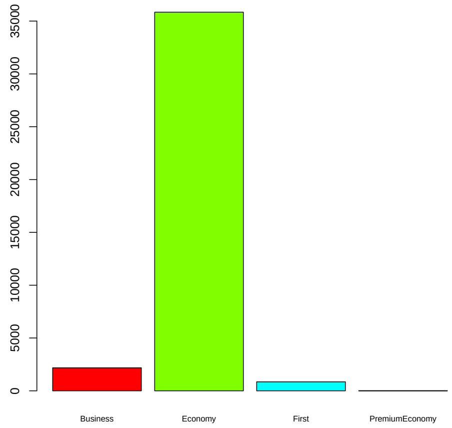


Table 3.12: Sample composition by class

on-board class	n	Rel.Freq
Economy	35,843	0.92
Business	2,180	0.06
First	853	0.02
PremiumEconomy	12	0.0003

Most of the observed client companies appear to have aimed at saving money, though not through low cost flights, but reserving a seat in economy class (possibly no low cost flight was available for the desired route). Tickets for seats in premium economy class are very rarely observed, maybe because often the possible saving, in comparison to the business class, is smaller than the actual difference in offered conditions.

With reference to the characteristics of the client companies, on behalf of which the tickets are purchased, unfortunately only the typology of client is available. Figure 10 shows how the sample is composed, based on it.

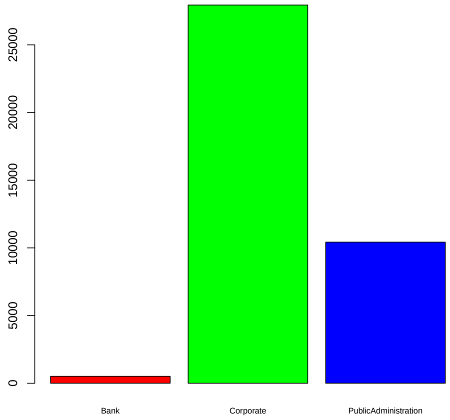


Figure 3.12: Sample composition by typology of client company

Table 3.13: Sample composition by typology of client company

on-board class	n	Rel.Freq
CORPORATE	27,940	0.72
PUBLIC ADM.	10,425	0.27
BANK	515	0.01

Most of the observed tickets were bought for private companies, more than a quarter for public institutions and just a few for banks, which may have less motivations to make their employee travelling. However, for the purpose of the present work, the most interesting variables should be those concerning the subject who actually flies (or does not fly).

In fact, the waiver to the flight may be due to personal issues or professional status related reasons, rather than to the characteristics of the itinerary. But unfortunately, as shown in table 2, except for 'sex', these are exactly the most problematic data, because of the number of not available cases.

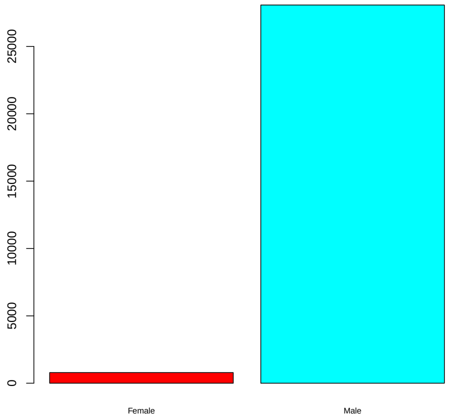


Figure 3.13: Sample composition by business traveler' sex

Table 3.14: Sample composition by business traveler' sex

Business traveler' sex	n	Rel.Freq
Female	788	0.02
Male	28,073	0.72

The clear preponderance of male business travelers in the sample is consistent with literature, which remarks that, though following an increasing trend, the number of females traveling for business purposes is still much lower than that of male (Collins and Tisdell, 2002; Tunstall, 1989). If this is true in general, that difference is also greater in a country like Italy (where most of Seneca's client companies have headquarters), where the share of working women is neatly smaller than the EU average (ISTAT, 2013).

Figure 3.14: Sample composition by business traveler' age

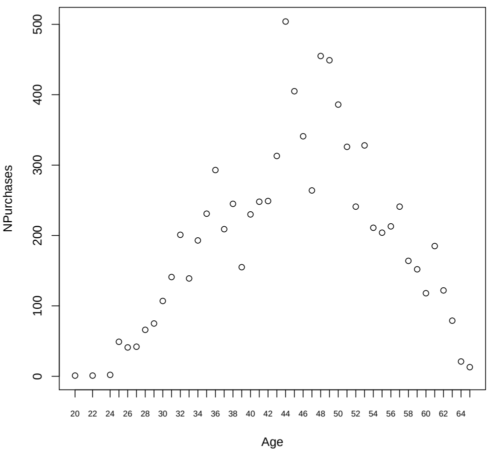


Table 3.15: Sample composition by business traveler' age

Business traveler's age	
Max	65
Median	46
Mean	45.73
Min	20
Var.	77.91
Sd.	8.83
Var. Coeff.	0.19

The observed frequencies are consistent with the evidence presented by the extant literature on business travel, which found that the modal age class of business travelers is that ranging from 45 to 54 (Collins and Tisdell, 2002 and 2000). The sample composition by professional status is represented in figure 12.

Figure 3.15: Sample composition by business traveler' job

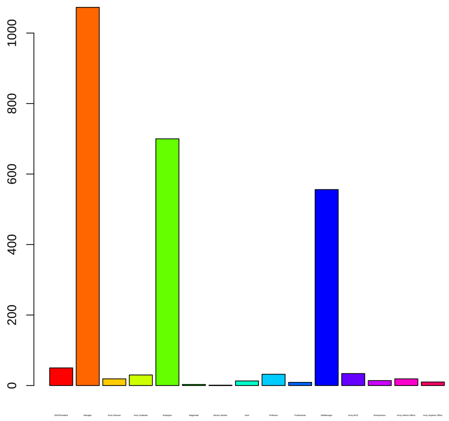


Table 3.16: Sample composition by business traveler' job

Business traveler's job	n	Rel.Freq
Manager	1, 073	0.0276
Employee	700	0.018
MidManager	556	0.0143
CEO/President	50	0.00129
Army NCO	34	0.00087
Professor	32	0.00082
Army Graduate	30	0.00077
Army General	19	0.00049
Army Inferior Officer	19	0.00049
Entrepreneur	14	0.00036
Host	13	0.00033
Army Superior Office	10	0.00026
Professional	9	0.00023
Magistrate	3	0.00008
Factory Worker	1	0.00003

The most frequent business fliers appear to be managers, followed by employees. Only an observed ticket was bought for a factory worker. Finally, a wider analysis is required for the variable of main interest to Seneca: the type of contract between the TMC and the client company, namely whether the the purchasing pattern is Target Buy.

Figure 3.16: Sample composition by type of contract

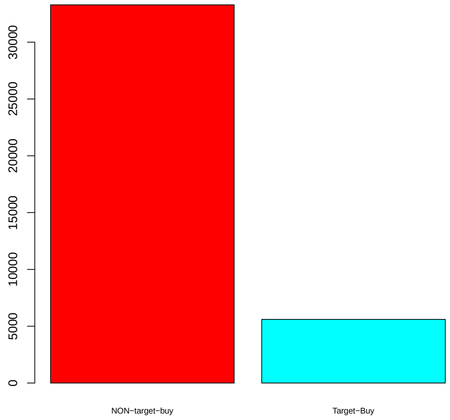


Table 3.17: Sample composition by type of contract

Type of Conctract	n	Rel.Freq
NON Targe- Buy	33, 286	0.856
Target-Buy	5, 603	0.144

About the 86% of observed air tickets were bought without the Target-Buy contract, maybe because this typology of purchasing pattern is relatively new and Seneca is the only TMC offering it, thus it is possible that it is still little known by firms. Therefore it can be interesting to look at which type of client companies subscribed the contract and which are the characteristics of flights purchased through it.

Figure 3.17: Distribution of Target-Buy by type of client company

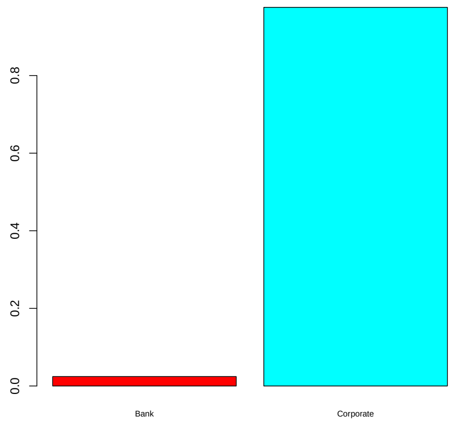


Table 3.18: Distribution of Target-Buy by type of client company

Type of client company	n	Rel.Freq
Corporate	5,467	0.98
Bank	136	0.02
Public Administration	0	0.00

Interestingly, no public institution subscribed the Target Buy contract, notwithstanding the need to make public spending more efficient and controlled, to which aim this contract can be helpful. While the 26% of client banks and the 20% of corporations employ Target Buy purchasing pattern.

Figure 3.18: Distribution of Target-Buy by ticket type

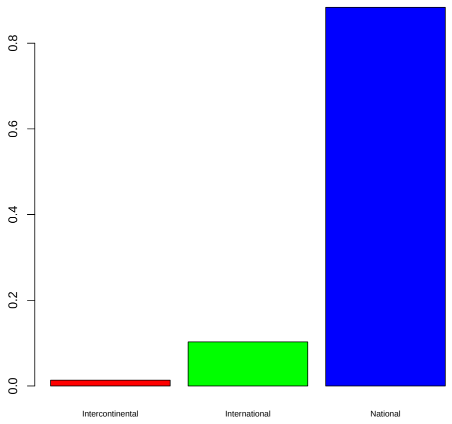


Table 3.19: Distribution of Target-Buy by ticket type

Type of ticket	n	Rel.Freq
National	4,951	0.884
Internat	576	0.103
Intercon	76	0.014

The 23% of tickets for national flights were bought through the Target-Buy contract, as just the 4% of those for international journeys and the 2% for intercontinental.

Figure 3.19: Distribution of Target-Buy by type of itinerary

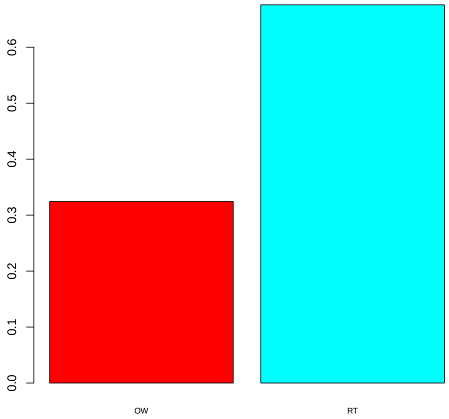


Table 3.20: Distribution of Target-Buy by type of itinerary

Type of itinerary	n	Rel.Freq
One Way	1,817	0.324
Round Trip	3,786	0.676

The 14% of one way and the 15% of round trip tickets were bought through the Target-Buy contract.

Figure 3.20: Distribution of Target-Buy by low cost flights

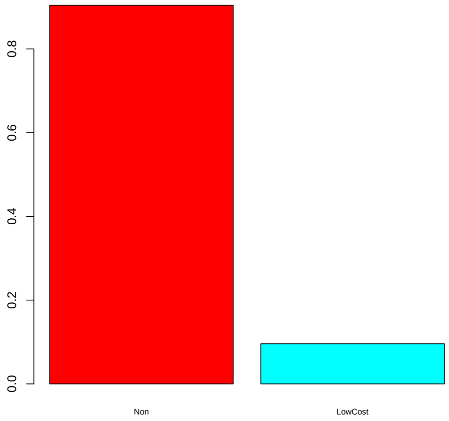


Table 3.21: Distribution of Target-Buy by low cost flights

	n	Rel.Freq
NON low cost	5,066	0.9
Low Cost	537	0.1

It is remarkable that the 90% of flights, purchased through Target Buy, are not low cost. However, among all the observed non low cost flights 'only' the 15% were bought through Target Buy (the 13% for low cost ones).

Figure 3.21: Distribution of Target-Buy by class

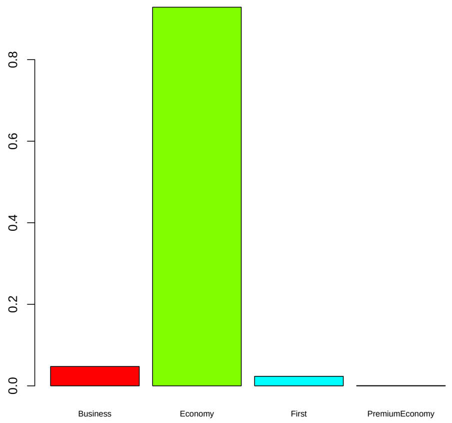


Table 3.22: Distribution of Target-Buy by class

On-board Class	n	Rel.Freq
ECO	5, 203	0.929
Bus	266	0.047
FIR	131	0.023
PREMIUMECO	3	0.0005

The 15% of tickets in economy class, the 12% in business, the 15% in first and the 25% in premium economy were purchased through Target-Buy.

The 97% of tickets, bought through Target-Buy, are for flights departing from Italy, the 1% from the United Kingdom and most of the remaining from European countries. Thus, it is likely that most of the Target-Buy clients are Italian. The 90% of tickets, purchased through Target-Buy, are for flights landing in Italy, the 2% in the United Kingdom and in the Netherlands, the 1% in France, Belgium and Spain.

Figure 3.22: Distribution of Target-Buy by advance booking

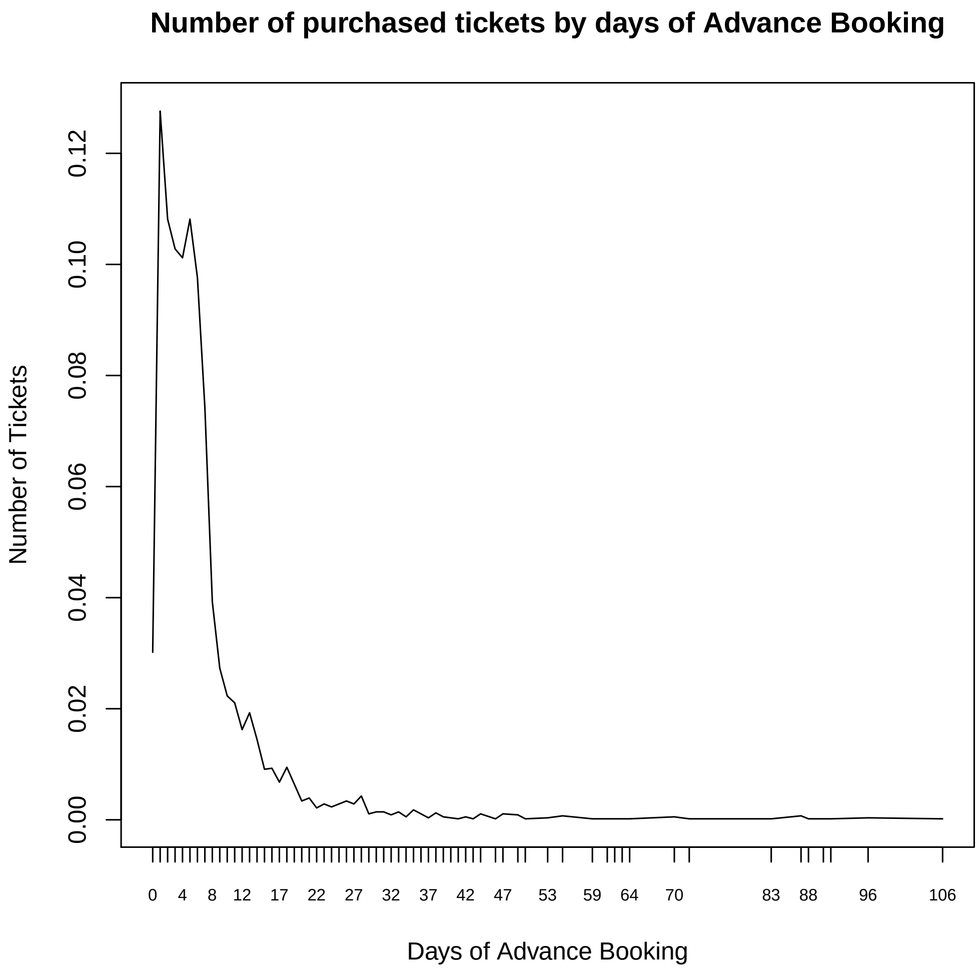

Number of purchased tickets by days of Advance Booking
Number of Tickets
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
4
8
12
17
22
27
32
37
42
47
53
59
64
70
83
88
96
106
Days of Advance Booking

Table 3.23: Distribution of Target-Buy by advance booking

Advance Booking	Days
Max	368
Median	7
Mean	13.3
Min	0
Var.	426.4
Sd.	20.6
Var. Coeff.	1.5

Table 23: Target-Buy contracts by advance booking

Advance Booking	Days
Max	106
Median	5
Mean	6.68
Min	0
Var.	63.77
Sd.	7.99
Var. Coeff.	1.9

Comparing the descriptive statistics of days of advance booking for the full sample and those for tickets purchased through Target-Buy it is clear that clients who subscribed the contract tend to make the reservations less days before the departure. This evidence may confirm that Target-Buy is the priviledged purchasing pattern for companies needing high flexibility.

The correlation (or association) coefficients between Y_f and each potential explanatory variable are listed in table 23. As 'age', 'advance booking' and 'number of routes' are count variables, they are scaled to approximate continuous variables, in order to get a complete frame of the relations between all the variables, although generally no correlation coefficient is to be computed for count variables.

Table 3.24: Correlation Coefficients

Y_1,	Phi Coefficients
Target-Buy	-0.014
Low Cost flight	-0.049
Type of Itinerary (RT)	-0.103
Sex (Male)	-0.332
Y_2,	Cramer's V
Sex (Male)	0.336
Type of Itinerary (RT)	0.104
Low Cost flight	0.051
Target-Buy	0.015
Y_3,	Cramer's V
Sex (Male)	0.335
Type of Itinerary (RT)	0.105
Low Cost flight	0.051
Target-Buy	0.015
Y_4,	Cramer's V
Sex (Male)	0.342
Type of Itinerary (RT)	0.107
Low Cost flight	0.063
Target-Buy	0.017
Y_5,	Cramer's V
Sex (Male)	0.343
Type of Itinerary (RT)	0.108
Low Cost flight	0.064
Target-Buy	0.025

Y_1,	Point-Biserial
Number of Routes	0.064
Age	0.015
Covered Distance	−0.011
Advance Booking	−0.04

Y_2,	Eta-squared
Number of Routes	0.0028
Advance Booking	0.0012
Covered Distance	0.0002
Age	0.0001

Y_3,	Eta-squared
Number of Routes	0.0018
Advance Booking	0.0011
Covered Distance	0.0003
Age	0.0000

Y_4,	Eta-squared
Number of Routes	0.0014
Advance Booking	0.0022
Covered Distance	0.0005
Age	0.0000

Y_5,	Eta-squared
Number of Routes	0.0010
Advance Booking	0.0018
Covered Distance	0.0006
Age	0.0001

Y_1,	Cramer’s V
Route	0.506
Type of Client	0.267
Arrival Country	0.195
Departure City	0.163
Professional Status	0.137
Arrival City	0.129
Airline	0.116
Class	0.114
Departure Country	0.081
Day of Departure	0.045
Day of Issuance	0.032
Day of Return	0.031
Type of Ticket	0.023

Y_2,	Cramer's V
Route	0.450
Type of Client	0.189
Arrival City	0.168
Departure City	0.142
Professional Status	0.117
Arrival Country	0.107
Airline	0.097
Class	0.082
Departure Country	0.074
Day of Departure	0.034
Day of Return	0.028
Day of Issuance	0.025
Type of Ticket	0.017

Y_3,	Cramer's V
Route	0.426
Arrival City	0.147
Departure City	0.126
Type of Client	0.12
Professional Status	0.107
Arrival Country	0.093
Airline	0.083
Class	0.067
Departure Country	0.064
Day of Departure	0.028
Day of Return	0.024
Day of Issuance	0.021
Type of Ticket	0.018

Y_4,	Cramer's V
Route	0.421
Type of Client	0.195
Arrival City	0.145
Departure City	0.123
Professional Status	0.113
Arrival Country	0.089
Airline	0.082
Class	0.07
Departure Country	0.064
Day of Departure	0.032
Day of Return	0.025
Type of Ticket	0.024
Day of Issuance	0.02

Y_5,	Cramer's V
Route	0.419
Type of Client	0.201
Arrival City	0.147
Departure City	0.124
Professional Status	0.109
Arrival Country	0.089
Airline	0.081
Class	0.071
Departure Country	0.064
Day of Departure	0.030
Day of Return	0.025
Type of Ticket	0.034
Day of Issuance	0.021

The interpretation of the above correlations is not straightforward, as different coefficients are calculated in different ways and take values on different ranges. Namely:

- Cramer's V and eta-squared coefficient range between 0 and 1;
- the range of the Phi coefficient depends on the probability distribution of the considered variables;
- the Point-Biserial coefficient ranges between -1 and 1.

Nonetheless, the reported values can be useful to compare values of the same coefficient among different explanatory variables and, for the same explanatory variable, between different specification of the dependent variable.

In particular, the type of itinerary appears to be relatively highly correlated with all the f specification of Y and preminently with Y_5. Thus, this variable is likely to particularly influence the rate of refund. This may be explained by the fact that it is more likely that a round trip is partially refunded, than an one way, as it is constituted by two ways and the business traveler can fly only one way, then change the other one or buy a completely new ticket and ask the refund of the non-flown way.

However, the sex of the business traveler seems to be much more correlated with the variable of interest. In particular, it appears that women are much more likely not to fly the first ticket, so that they tend to do more changes and ask refunds more frequently. On the contrary, low cost flights tend to be flown and Target-Buy clients seem to tend to depart.

The small, but negative correlation of advance booking with non-departure may be influenced by the fact that some tickets reserved with many days of advance were not still modified nor flown at the date of the data collection (right censoring). A similar relation seems to hold between the tendency not to fly the first ticket and the covered distance. This evidence may be explained hypothesizing that the longer the journey the less easily it can be re-organized, for another date, for example. Indeed, this consideration may be in contrast with the values of the correlation coefficients for the number of routes. But it is not to be taken for granted that a higher number of routes implies a wider covered distance.

In general, it seems that the variable "route" is the most related with the risk of non-fly. The type of client company can also be an important predictor for such a risk. The places of arrival appear to be more related with the behavior of the business traveler than the places of departure. Thus, maybe the characteristics of the journey influence the risk of ticket change/refund more than the client company (the nationality of which is rather related with the departure place).

However, a measure of association comparable between different specifications of the dependent variable and between all of the candidate regressors is more useful, especially to lead the aggregation of categorical variables' levels and the definitive specification of the dependent variable. Thus, the P-values of chi-squared tests for independence are reported below.

Table 3.25: P-values of chi-squared tests

Table 25: Chi-square test of Independence: P-values

Variables	FLOWN	NONflown	PartREFUND	TotREFUND	1PartCHANGEandFlown
ETA	0.135	0.483	0.859	0.247	0.020
AIRLINE	0.000	0.001	0.001	0.000	0.001
Army	0.000	0.455	0.099	0.033	1.000
classe	0.366	0.126	0.725	0.176	0.734
TipoCliente	0.000	0.000	0.000	0.000	0.000
Nazionale	0.000	0.004	0.001	0.000	0.247
Nroutes	0.000	0.001	0.000	0.000	0.000
Weemiss	0.012	0.380	0.117	0.040	0.052
Wepart	0.000	0.015	0.915	0.004	0.000
Werit	0.013	0.448	0.252	0.004	0.000
Maschi	0.000	0.000	0.000	0.000	0.000
RT1_OW0	0.000	0.000	0.001	0.000	0.000
AnticipoPrenotazioneG	0.000	0.000	0.000	0.000	0.000
Cliente_Target_Buy	0.006	0.886	0.011	0.008	0.001
Tariffa_piena	0.000	0.000	0.000	0.000	0.000
SpesaBiglietto	0.000	1.000	0.000	0.551	0.000
CostoAlKM	0.000	0.000	0.000	0.000	0.000
Distanza_KM	0.000	0.000	0.000	0.000	0.000
TotVersato a Seneca	0.000	1.000	0.000	1.000	0.000

ETA	1TotChangeandFlown	2PartCHANGEandFlown	2TotChangeandFlown	1TotChange1PartChangeandFLOWN	2PartCHANGEandNOTFlown
AIRLINE	0.816	0.064	0.789	0.696	/
Army	0.000	0.078	1.000	0.847	1.000
classe	0.000	0.530	1.000	0.862	/
TipoCliente	0.618	0.097	0.607	0.066	0.607
Nazionale	0.000	0.000	0.000	0.103	0.822
Nroutes	0.000	0.320	1.000	0.870	1.000
Weemiss	0.000	0.071	0.954	0.039	0.634
Wepart	0.053	1.000	1.000	1.000	1.000
Werit	0.093	0.127	1.000	1.000	1.000
Maschi	0.429	0.654	1.000	0.990	/
RT1_OW0	0.000	0.000	0.004	0.004	1.000
AnticipoPrenotazioneG	0.000	0.008	1.000	0.787	0.728
Cliente_Target_Buy	0.000	0.743	1.000	1.000	1.000
Tariffa_piena	1.000	0.907	1.000	1.000	1.000
SpesaBiglietto	0.000	0.988	1.000	0.000	1.000
CostoAlKM	0.000	1.000	1.000	0.000	1.000
Distanza_KM	0.000	0.000	1.000	0.940	1.000
TotVersato a Seneca	0.000	1.000	1.000	0.000	1.000
	0.000	1.000	1.000	0.000	1.000

[redacted]

	1TotChange1PartChangeandPartialREFUND	3PartialChangeand.FLOWN	1TotChange2PartChangeandFLOWN	1TotChange3PartChangeandFLOWN
ETA	/	0.854	/	/
AIRLINE	0.840	0.678	0.707	1.000
Army	/	1.000	1.000	/
classe	0.607	0.203	0.606	0.607
TipoCliente	0.255	0.372	0.756	0.822
Nazionale	1.000	0.657	1.000	1.000
Nroutes	0.954	0.350	0.420	0.634
Weemiss	1.000	1.000	1.000	1.000
Wepart	1.000	0.944	1.000	1.000
Werit	1.000	0.621	1.000	/
Maschi	1.000	0.000	0.053	1.000
RT1_OW0	1.000	0.040	0.798	0.728
AnticipoPrenotazioneG	1.000	1.000	1.000	0.000
Cliente_Target_Buy	1.000	1.000	1.000	1.000
Tariffa_piena	1.000	0.001	1.000	1.000
SpesaBiglietto	0.000	1.000	0.000	1.000
CostoAlKM	1.000	0.938	0.948	0.101
Distanza_KM	1.000	0.000	0.000	1.000
TotVersato a Seneca	0.000	1.000	0.000	1.000

Then, variables are aggregated and selected (within sets of overlapping/equivalent ones, see above) combining logical considerations, sign and significance of the chi-squared association and results of a more naive analysis, based on the difference between the conditional distribution of Y_f, to each single level of each single variable, and the marginal distribution of Y_f, displayed below (if the difference is zero, then the two variables are absolutely independent in distribution).

Table 3.26: Differences from the marginal masses

Table 26: differences between conditional and marginal distributions

MARGINAL	1	0.852	0.008	0.020	0.014	0.077	0.027
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
BANK	421	-0.842	0.037	0.208	0.065	0.393	0.109
CORPORATE	26079	0.031	-0.001	-0.003	0.004	-0.032	0.001
PA	9811	0.059	-0.003	0.003	-0.010	-0.034	-0.017
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
OPERAIO	1	-0.852	-0.008	-0.020	-0.014	0.923	-0.027
PROF	32	-0.852	0.023	0.042	-0.014	0.704	0.036
HOST	13	-0.775	0.146	0.056	-0.014	0.615	-0.027
UFFICIALIinferiori	19	-0.799	-0.008	0.137	-0.014	0.607	-0.027
MAGISTRATO	3	-0.852	-0.008	0.313	-0.014	0.590	-0.027
IMPIEGATO	700	-0.806	0.044	0.101	0.028	0.517	0.100
GENERALE	19	-0.746	0.098	0.085	-0.014	0.502	0.026
GRADUATI	30	-0.652	0.026	0.146	0.020	0.490	-0.027
PROFESSIONISTA	9	-0.852	-0.008	0.091	-0.014	0.479	0.306
DIRETTOREadPres	50	-0.812	0.052	0.120	0.126	0.443	0.033
UFFICIALIsuperiori	10	-0.552	-0.008	0.180	-0.014	0.423	-0.027
DIRIGENTE	1073	-0.810	0.030	0.110	0.074	0.422	0.160
QUADRO	556	-0.821	0.048	0.118	0.087	0.420	0.124
SOTTOufficiali	34	-0.705	-0.008	0.244	0.016	0.364	0.003
TITOLARE	14	-0.852	-0.008	0.122	0.058	0.352	0.330
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Nazionale	21273	-0.007	0.001	-0.002	0.003	0.002	0.004
Internazionale	14013	0.010	-0.001	0.002	-0.004	-0.001	-0.006
Intercontinentale	3603	0.003	-0.002	0.007	-0.003	-0.006	0.000
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
7	3	-0.185	-0.008	-0.020	-0.014	0.256	-0.027
5	114	-0.062	-0.008	0.023	-0.014	0.046	0.008
1	10917	-0.061	0.003	0.002	0.007	0.039	0.009
6	114	-0.071	0.001	0.050	0.013	-0.007	0.017
2	23165	0.025	-0.001	-0.002	-0.004	-0.014	-0.005
3	729	0.030	-0.004	0.007	-0.001	-0.019	-0.013
4	3846	0.020	-0.003	0.000	0.003	-0.024	0.004
8	1	-0.852	-0.008	-0.020	0.986	-0.077	-0.027
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Mar	7790	-0.020	0.001	0.002	0.004	0.006	0.007
Mer	7915	0.002	0.000	-0.002	0.002	0.000	-0.002
Lun	7101	-0.002	-0.001	0.000	0.000	-0.001	0.003
Ven	7879	0.008	-0.001	0.001	-0.002	-0.001	-0.005
Giov	7488	0.009	0.001	-0.001	-0.003	-0.002	-0.003
Dom	216	0.037	-0.008	-0.002	-0.014	-0.003	-0.013
Sab	500	0.032	-0.002	0.014	-0.008	-0.027	-0.013
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Ven	4426	-0.005	0.001	0.000	-0.004	0.014	-0.008
Giov	5597	-0.008	-0.001	-0.001	0.002	0.005	0.002
Mer	7486	-0.021	0.003	0.001	0.005	0.001	0.012
Lun	8590	0.017	-0.001	-0.003	-0.004	-0.001	-0.008
Mar	7965	-0.006	0.001	0.002	0.002	-0.003	0.003

Sab	1505	0.021	-0.004	-0.001	-0.004	-0.009	-0.002
Dom	3320	0.030	-0.003	0.001	-0.005	-0.018	-0.005
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Dom	1551	-0.004	-0.002	-0.002	-0.005	0.017	-0.004
Ven	5908	0.017	-0.001	-0.002	0.000	-0.011	-0.002
Lun	2876	0.040	0.000	-0.007	-0.005	-0.018	-0.009
Mar	3723	0.030	-0.002	-0.003	-0.002	-0.019	-0.004
Sab	1417	0.028	-0.003	0.004	-0.011	-0.021	0.004
Giov	5511	0.027	-0.001	-0.002	-0.002	-0.021	-0.001
Mer	4984	0.031	0.000	0.000	-0.002	-0.027	-0.002
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Maschi	28072	-0.029	0.001	0.004	0.002	0.018	0.004
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Cliente_Target_Bu	5603	0.012	0.000	-0.005	0.004	-0.011	0.000
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
RT1_OW0	25834	0.026	-0.001	-0.002	-0.003	-0.017	-0.003
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
2	2173	0.016	-0.002	0.000	-0.007	0.000	-0.006
1	1058	0.033	-0.006	-0.002	-0.006	-0.010	-0.009
4	1512	0.032	-0.003	-0.003	-0.001	-0.017	-0.008
3	3195	0.040	-0.001	-0.004	-0.007	-0.018	-0.009
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Eco	9718	-0.002	0.001	0.001	-0.001	0.001	0.000
Bus	19450	-0.001	-0.001	0.000	0.001	0.001	0.001
Fir	9720	0.004	0.001	0.000	-0.001	-0.003	-0.001
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
LowCost	1044	-0.006	0.003	0.000	-0.004	0.020	-0.015
MERIDIANA_FLY	905	0.021	-0.003	-0.006	-0.009	0.019	-0.020
BRITISH_AIRWAYS	601	0.000	-0.004	0.005	-0.009	0.008	0.000
Other	6559	0.005	-0.002	0.004	-0.005	0.003	-0.005
ALITALIA_CAI_SPA	21838	-0.012	0.001	0.000	0.004	0.001	0.007
AIR_FRANCE	939	-0.037	0.003	0.011	0.013	0.000	0.008
LUFTHANSA	2615	0.002	0.000	0.001	0.002	-0.008	0.002
BRUSSELS_AIRLINI	1260	0.025	0.000	0.000	-0.009	-0.009	-0.012
RYANAIR	1486	0.069	-0.006	-0.012	-0.014	-0.010	-0.025
EASYJET	1642	0.073	-0.005	-0.008	-0.013	-0.020	-0.025
	n	FLOWN	NONflown	PartREFUN	TotREFUNI	PartCh	TotCh
Cuneo	32	-0.508	-0.008	0.073	0.205	0.204	0.036
Miami	17	-0.323	-0.008	0.215	-0.014	0.158	-0.027
Stuttgart	13	-0.083	-0.008	-0.020	-0.014	0.154	-0.027
Shanghai	18	-0.130	-0.008	0.035	-0.014	0.145	-0.027
Munich	45	-0.118	-0.008	0.002	0.008	0.123	-0.004
Sao_Paulo	28	-0.102	-0.008	0.015	-0.014	0.102	0.009
Bremen	17	-0.146	-0.008	-0.020	-0.014	0.100	0.032
Moscow	40	-0.077	-0.008	0.005	0.011	0.098	-0.027
Malta	23	-0.200	-0.008	0.154	-0.014	0.097	-0.027
Helsinki	29	-0.093	-0.008	0.014	0.021	0.096	-0.027
Dublin	30	-0.052	-0.008	0.013	-0.014	0.090	-0.027
Heraklion	121	-0.108	0.001	-0.004	0.019	0.080	0.015
WINNIPEG	13	-0.006	-0.008	-0.020	-0.014	0.077	-0.027

Ancona	42	-0.161	0.087	-0.020	-0.014	0.066	0.045
Alicante	15	0.015	-0.008	-0.020	-0.014	0.056	-0.027
Dubai	23	-0.026	-0.008	-0.020	-0.014	0.054	0.017
Tirana	70	-0.037	-0.008	0.008	0.001	0.052	-0.012
Valencia	16	0.023	-0.008	-0.020	-0.014	0.048	-0.027
Istanbul	41	-0.022	-0.008	0.004	-0.014	0.045	-0.002
Olbia	108	0.009	-0.008	-0.020	0.005	0.044	-0.027
Ljubljana	25	0.028	-0.008	-0.020	-0.014	0.043	-0.027
Madrid	90	-0.018	0.037	-0.009	-0.014	0.034	-0.027
New_York	45	-0.118	-0.008	0.024	0.031	0.034	0.040
Barcelona	109	-0.035	0.001	-0.002	-0.005	0.033	0.001
Naples	1134	-0.116	0.011	-0.003	0.035	0.028	0.046
Krakow	29	-0.024	-0.008	0.014	-0.014	0.027	-0.027
Frankfurt	305	-0.032	0.009	-0.001	-0.001	0.025	-0.004
Bologna	417	0.000	-0.008	0.004	-0.009	0.024	-0.012
Sevilla	10	-0.052	-0.008	-0.020	-0.014	0.023	0.073
Ankara	21	0.053	-0.008	-0.020	-0.014	0.018	-0.027
Lisbon	43	0.055	-0.008	-0.020	-0.014	0.016	-0.027
Zagreb	43	0.032	-0.008	0.003	-0.014	0.016	-0.027
London	280	0.023	-0.008	-0.006	-0.007	0.016	-0.020
Pescara	120	0.040	-0.008	-0.004	-0.014	0.015	-0.027
Thessaloniki	11	-0.034	-0.008	0.070	-0.014	0.014	-0.027
Milan	8956	-0.035	0.002	0.004	0.007	0.011	0.012
Paris	467	-0.023	0.012	0.016	-0.005	0.011	-0.012
Dusseldorf	23	-0.113	-0.008	0.067	-0.014	0.010	0.017
Chicago	24	0.023	-0.008	0.021	-0.014	0.006	-0.027
San_Francisco	12	-0.018	-0.008	-0.020	0.070	0.006	-0.027
Athens	291	-0.024	0.013	0.014	-0.010	0.006	0.001
Turin	1666	0.020	-0.002	-0.004	-0.007	0.005	-0.010
Cagliari	973	0.030	-0.003	-0.007	-0.008	0.001	-0.016
Amsterdam	90	0.037	-0.008	-0.009	-0.003	0.001	-0.016
Zurich	39	0.046	-0.008	0.005	-0.014	0.000	-0.027
Basel	13	0.071	-0.008	-0.020	-0.014	0.000	-0.027
Marseille	13	0.071	-0.008	-0.020	-0.014	0.000	-0.027
Toronto	13	0.071	-0.008	-0.020	-0.014	0.000	-0.027
Copenhagen	645	0.041	-0.003	-0.010	-0.008	-0.001	-0.019
Verona	205	0.012	-0.003	-0.006	-0.004	-0.004	0.007
Rome	12419	0.005	-0.001	-0.001	0.000	-0.005	0.002
Nice	56	0.041	-0.008	-0.020	0.004	-0.005	-0.009
Brussels	432	0.023	-0.003	0.007	-0.009	-0.007	-0.020
Trieste	375	-0.044	0.014	0.001	0.016	-0.008	0.019
Palermo	1311	0.032	-0.003	-0.003	-0.006	-0.008	-0.011
Bari	958	0.044	-0.001	-0.006	-0.009	-0.011	-0.014
Catania	1078	0.052	-0.005	-0.007	-0.010	-0.012	-0.019
Vienna	94	0.031	-0.008	0.011	-0.014	-0.013	-0.005
Brindisi	425	0.052	-0.008	-0.009	-0.009	-0.013	-0.010
Beijing	16	-0.102	0.055	0.042	-0.014	-0.014	0.036
Podgorica	16	0.023	-0.008	-0.020	-0.014	-0.014	0.036
Pisa	197	0.062	-0.003	0.000	-0.014	-0.016	-0.027
Alghero	250	0.052	0.000	0.000	-0.014	-0.017	-0.023

Casablanca	17	0.031	-0.008	-0.020	-0.014	-0.018	0.032
Florence	428	0.008	0.006	-0.004	0.003	-0.018	0.008
Trapani	52	0.091	-0.008	-0.020	-0.014	-0.019	-0.027
Washington	18	0.037	-0.008	-0.020	0.042	-0.021	-0.027
Tunis	37	0.094	-0.008	-0.020	-0.014	-0.023	-0.027
Mexico_City	19	0.096	-0.008	-0.020	-0.014	-0.024	-0.027
Bucharest	39	0.046	-0.008	0.005	-0.014	-0.026	-0.001
Reggio_Calabria	216	0.056	-0.003	-0.002	-0.014	-0.026	-0.008
Venice	675	0.059	-0.006	-0.013	-0.002	-0.027	-0.010
Lyon	20	0.048	-0.008	0.030	-0.014	-0.027	-0.027
Luxembourg	42	0.077	-0.008	0.003	-0.014	-0.029	-0.027
Riga	21	0.053	-0.008	0.027	-0.014	-0.029	-0.027
Tallinn	21	0.101	-0.008	-0.020	-0.014	-0.029	-0.027
Warsaw	109	0.029	0.001	-0.002	-0.005	-0.031	0.001
Prague	66	0.088	-0.008	-0.020	-0.014	-0.031	-0.011
Hamburg	22	0.103	-0.008	-0.020	-0.014	-0.031	-0.027
Genoa	835	0.046	0.002	-0.006	-0.007	-0.035	0.001
Pantelleria	25	0.028	-0.008	0.020	-0.014	-0.037	-0.027
Bolzano	77	0.044	-0.008	0.019	-0.014	-0.038	-0.014
Sofia	26	0.110	-0.008	-0.020	-0.014	-0.038	-0.027
Tokyo	26	0.071	-0.008	-0.020	0.025	-0.038	-0.027
Buenos_Aires	54	-0.130	0.029	0.165	-0.014	-0.040	-0.008
Berlin	136	0.053	-0.008	0.016	-0.014	-0.040	-0.005
Belgrade	31	0.084	0.025	-0.020	-0.014	-0.045	-0.027
Geneva	31	-0.013	0.057	-0.020	-0.014	-0.045	0.006
Lamezia_Terme	340	0.086	0.001	-0.012	-0.011	-0.047	-0.015
Manchester	49	0.067	-0.008	-0.020	0.007	-0.056	0.014
Lampedusa	52	0.129	-0.008	-0.020	-0.014	-0.058	-0.027
Budapest	42	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Stockholm	31	0.116	0.025	-0.020	-0.014	-0.077	-0.027
Skopje	20	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Tel_Aviv_Yafo	20	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Algiers	19	-0.010	-0.008	-0.020	0.039	-0.077	0.079
Sarajevo	19	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Cairo	17	-0.028	-0.008	0.098	-0.014	-0.077	-0.027
Pristina	16	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Rio_De_Janeiro	15	-0.252	0.059	0.180	0.120	-0.077	-0.027
Las_Vegas	14	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
BAKU	13	-0.083	0.069	0.056	-0.014	-0.077	0.050
Edinburgh	13	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Amman	12	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Birmingham	12	0.065	-0.008	-0.020	-0.014	-0.077	-0.027
Newark	12	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Boston	11	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Vilnius	11	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Hong_Kong	10	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Tbilisi	10	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Muscat	12	-0.518	0.076	0.063	-0.014	0.256	0.057
Monastir	11	-0.125	-0.008	-0.020	-0.014	0.196	-0.027
Oslo	22	-0.215	-0.008	0.070	-0.014	0.196	-0.027

TEHERAN	11	-0.215	-0.008	-0.020	0.077	0.196	-0.027
Chisinau	20	-0.302	-0.008	0.180	-0.014	0.173	-0.027
Sao_Paulo	90	-0.274	0.026	0.046	0.031	0.156	0.018
Strasbourg	28	-0.209	-0.008	0.123	-0.014	0.137	-0.027
Bangalore	19	-0.063	-0.008	-0.020	-0.014	0.134	-0.027
Tirana	270	-0.196	0.000	0.057	0.012	0.127	-0.008
Miami	35	-0.166	-0.008	0.008	-0.014	0.123	0.059
Entebbe	10	-0.152	-0.008	-0.020	-0.014	0.123	0.073
Riyadh	16	-0.164	-0.008	0.042	-0.014	0.111	-0.027
Glasgow	11	-0.034	-0.008	-0.020	-0.014	0.105	-0.027
Moscow	140	-0.059	-0.008	-0.013	-0.007	0.087	0.002
Delhi	25	-0.132	0.032	0.060	-0.014	0.083	-0.027
EDMONTON	13	-0.006	-0.008	-0.020	-0.014	0.077	-0.027
Nantes	13	-0.006	-0.008	-0.020	-0.014	0.077	-0.027
Thessaloniki	13	-0.083	-0.008	-0.020	-0.014	0.077	0.050
Belgrade	35	-0.137	-0.008	-0.020	0.015	0.066	0.088
Singapore	35	-0.195	-0.008	-0.020	0.015	0.066	0.116
Tokyo	53	-0.173	0.011	0.055	0.024	0.055	0.030
Warsaw	114	-0.036	0.001	0.006	-0.005	0.055	-0.018
Sydney	23	-0.113	0.036	0.067	-0.014	0.054	-0.027
Beijing	33	-0.064	-0.008	0.010	-0.014	0.044	0.034
Berlin	263	-0.011	0.000	0.002	-0.010	0.041	-0.019
Bucharest	79	-0.016	0.005	-0.020	0.012	0.037	-0.014
Birmingham	44	-0.102	-0.008	0.025	-0.014	0.037	0.064
Guangzhou	62	-0.045	-0.008	0.028	0.002	0.036	-0.011
Basel	80	-0.014	-0.008	0.005	0.011	0.035	-0.027
Athens	381	-0.014	-0.005	-0.007	-0.003	0.033	-0.003
Stuttgart	38	0.043	-0.008	-0.020	-0.014	0.028	-0.027
Doha	19	0.043	-0.008	-0.020	-0.014	0.028	-0.027
Bologna	202	0.020	-0.003	-0.011	-0.014	0.027	-0.017
Lima	29	-0.093	-0.008	0.118	-0.014	0.026	-0.027
Los_Angeles	30	0.015	-0.008	-0.020	-0.014	0.023	0.007
Shanghai	61	-0.147	0.009	0.029	0.035	0.022	0.039
Dubai	31	-0.045	-0.008	0.012	-0.014	0.020	0.038
Ottawa	21	0.005	-0.008	0.027	-0.014	0.018	-0.027
Palermo	1201	0.008	-0.004	-0.004	-0.006	0.017	-0.013
Bangkok	11	-0.124	-0.008	-0.020	0.077	0.014	0.064
COMISO_RAGUSA	11	0.057	-0.008	-0.020	-0.014	0.014	-0.027
Detroit	11	0.057	-0.008	-0.020	-0.014	0.014	-0.027
Johannesburg	11	0.057	-0.008	-0.020	-0.014	0.014	-0.027
Mexico_City	11	0.057	-0.008	-0.020	-0.014	0.014	-0.027
Milan	5874	-0.043	0.004	-0.001	0.011	0.014	0.017
Luxembourg	156	0.001	-0.001	-0.001	-0.014	0.013	0.005
Heraklion	67	-0.046	0.007	-0.006	0.001	0.013	0.033
San_Francisco	56	-0.066	0.010	0.015	-0.014	0.012	0.045
Guayaquil	34	-0.146	0.022	0.156	-0.014	0.011	-0.027
Verona	229	-0.026	0.001	-0.012	-0.001	0.010	0.026
Katowice	23	-0.069	0.036	-0.020	-0.014	0.010	0.060
Tallinn	23	-0.026	0.036	-0.020	-0.014	0.010	0.017
TUNIS	116	0.019	-0.008	0.014	-0.014	0.009	-0.018

Vienna	383	0.007	0.003	-0.007	-0.009	0.009	-0.003
Naples	1214	-0.094	0.002	-0.002	0.038	0.009	0.046
Madrid	282	-0.033	0.003	0.004	-0.007	0.008	0.023
Helsinki	71	0.036	-0.008	0.008	-0.014	0.008	-0.027
Istanbul	226	0.002	-0.003	-0.007	-0.009	0.007	0.009
Abidjan	24	0.065	-0.008	-0.020	-0.014	0.006	-0.027
Vancouver	12	-0.102	-0.008	-0.020	-0.014	0.006	0.140
Rome	7977	-0.008	0.001	0.003	-0.001	0.006	-0.002
Barcelona	262	0.018	0.000	-0.009	-0.002	0.003	-0.015
Nice	25	0.068	-0.008	-0.020	-0.014	0.003	-0.027
Sofia	25	0.068	-0.008	-0.020	-0.014	0.003	-0.027
Trieste	364	-0.058	0.006	-0.004	0.025	0.003	0.028
Budapest	64	0.054	-0.008	-0.020	-0.014	0.001	-0.011
Ancona	39	0.071	-0.008	-0.020	-0.014	0.000	-0.027
Kuala_Lumpur	26	0.071	-0.008	-0.020	-0.014	0.000	-0.027
Cagliari	743	0.027	0.000	-0.008	-0.007	0.000	-0.009
Brussels	1754	0.023	-0.001	-0.002	-0.009	-0.001	-0.012
Genoa	724	0.018	-0.001	-0.002	0.000	-0.002	-0.014
Florence	261	0.014	0.004	0.003	-0.006	-0.004	-0.011
Pisa	153	0.037	-0.001	-0.020	-0.007	-0.005	-0.007
London	967	0.024	-0.005	-0.003	-0.004	-0.007	-0.005
Venice	526	0.038	-0.004	-0.007	-0.010	-0.007	-0.010
Turin	731	0.013	-0.002	0.000	0.001	-0.007	-0.008
Paris	1022	-0.003	0.000	0.005	0.001	-0.007	0.008
Santiago	29	-0.128	-0.008	0.014	0.021	-0.008	0.077
Frankfurt	1627	0.015	-0.003	0.004	-0.005	-0.009	-0.003
Porto	15	0.082	-0.008	-0.020	-0.014	-0.010	-0.027
Munich	350	-0.017	0.012	-0.003	0.015	-0.011	0.005
Washington	62	-0.029	-0.008	0.028	0.019	-0.012	0.006
Pantelleria	32	0.086	-0.008	-0.020	-0.014	-0.014	-0.027
Prague	120	0.082	0.001	-0.020	-0.014	-0.019	-0.027
Chicago	35	0.034	-0.008	-0.020	0.015	-0.020	0.002
Marseille	35	0.091	-0.008	-0.020	-0.014	-0.020	-0.027
Riga	35	0.034	-0.008	0.008	0.015	-0.020	-0.027
Alghero	286	0.054	-0.004	-0.010	-0.014	-0.021	-0.006
San_Diego	18	0.037	-0.008	-0.020	-0.014	-0.021	0.029
Seoul	18	0.093	-0.008	-0.020	-0.014	-0.021	-0.027
Valencia	37	0.013	-0.008	-0.020	0.013	-0.023	0.027
Zurich	169	0.012	-0.008	0.009	-0.002	-0.024	0.015
Alicante	19	0.043	0.045	-0.020	-0.014	-0.024	-0.027
Lamezia_Terme	384	0.065	-0.003	-0.013	-0.009	-0.025	-0.014
New_York	213	0.040	-0.003	-0.011	-0.014	-0.025	0.016
Mumbai	39	-0.031	0.018	-0.020	0.012	-0.026	0.050
Malta	60	0.065	0.009	-0.004	-0.014	-0.027	-0.027
Catania	906	0.059	-0.003	-0.006	-0.005	-0.027	-0.017
Krakow	101	-0.010	0.002	-0.011	0.006	-0.027	0.043
Manchester	101	0.010	0.002	0.019	0.006	-0.027	-0.007
Geneva	81	-0.012	0.017	0.004	0.011	-0.027	0.010
Bari	860	0.059	-0.001	-0.003	-0.011	-0.028	-0.014
Lampedusa	105	0.072	-0.008	-0.001	-0.014	-0.029	-0.017

Hamburg	64	-0.039	0.008	-0.005	0.033	-0.030	0.036
Casablanca	65	0.071	0.008	-0.020	-0.014	-0.031	-0.027
Melbourne	22	0.057	-0.008	-0.020	-0.014	-0.031	0.019
Brindisi	447	0.061	-0.005	-0.005	-0.012	-0.032	-0.007
Malaga	23	0.105	-0.008	-0.020	-0.014	-0.033	-0.027
Reggio_Calabria	210	0.091	-0.008	-0.011	-0.009	-0.034	-0.027
Ankara	24	0.065	-0.008	0.021	-0.014	-0.035	-0.027
Lyon	24	0.065	-0.008	-0.020	-0.014	-0.035	0.015
Amsterdam	456	0.041	0.005	-0.007	0.008	-0.037	-0.009
Edinburgh	27	0.074	-0.008	0.017	-0.014	-0.040	-0.027
Toronto	27	0.074	-0.008	0.017	-0.014	-0.040	-0.027
Zagreb	85	0.042	-0.008	0.003	-0.002	-0.042	0.009
Dublin	86	0.102	0.004	-0.020	-0.014	-0.042	-0.027
Buenos_Aires	29	0.045	-0.008	-0.020	0.021	-0.042	0.008
Lisbon	117	0.097	0.001	-0.020	-0.014	-0.043	-0.018
Cologne	31	0.084	-0.008	-0.020	-0.014	-0.045	0.006
Hong_Kong	96	-0.008	-0.008	0.000	0.049	-0.046	0.015
Copenhagen	575	0.091	-0.006	-0.003	-0.010	-0.047	-0.025
Bolzano	102	0.050	0.012	0.029	-0.014	-0.047	-0.027
Philadelphia	43	0.009	-0.008	0.026	-0.014	-0.054	0.043
Dusseldorf	93	0.062	-0.008	-0.020	-0.003	-0.055	0.016
Boston	51	0.089	-0.008	-0.001	-0.014	-0.057	-0.007
Olbia	104	0.081	-0.008	-0.001	-0.014	-0.058	-0.007
Stockholm	61	0.083	0.009	-0.004	-0.014	-0.060	-0.010
Tel_Aviv_Yafo	749	0.139	-0.008	-0.020	-0.012	-0.069	-0.027
Trapani	50	0.128	-0.008	0.000	-0.014	-0.077	-0.027
Pescara	49	0.087	0.013	0.020	-0.014	-0.077	-0.027
Toulouse	44	0.103	-0.008	-0.020	-0.014	-0.077	0.019
Dallas	28	0.077	-0.008	0.015	-0.014	-0.077	0.009
Ljubljana	27	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Sevilla	23	0.105	-0.008	-0.020	0.030	-0.077	-0.027
Dresden	21	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Newark	21	0.101	-0.008	-0.020	-0.014	-0.077	0.021
Podgorica	19	0.043	-0.008	0.032	-0.014	-0.077	-0.027
Amman	18	0.037	-0.008	0.035	-0.014	-0.077	-0.027
Basra	18	-0.130	-0.008	0.091	0.042	-0.077	0.084
Bilbao	18	0.093	-0.008	0.035	-0.014	-0.077	-0.027
Skopje	18	0.093	-0.008	0.035	-0.014	-0.077	-0.027
Vilnius	17	0.089	-0.008	-0.020	-0.014	-0.077	0.032
St_Petersburg	14	0.077	-0.008	-0.020	-0.014	-0.077	-0.027
Tbilisi	14	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Beirut	13	0.071	0.069	-0.020	-0.014	-0.077	-0.027
Rotterdam	13	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Hanover	12	-0.018	-0.008	0.063	0.070	-0.077	-0.027
Jakarta	12	0.065	-0.008	-0.020	0.070	-0.077	-0.027
Palma_Mallorca	12	-0.019	-0.008	0.147	-0.014	-0.077	-0.027
Rio_De_Janeiro	12	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Dakar	11	0.057	-0.008	-0.020	-0.014	-0.077	0.064
Larnaca	11	0.148	-0.008	-0.020	-0.014	-0.077	-0.027
Maputo	11	-0.397	-0.008	0.525	-0.014	-0.077	-0.027

Atlanta	10	0.048	-0.008	-0.020	-0.014	-0.077	0.073
Gothenburg	10	0.148	-0.008	-0.020	-0.014	-0.077	-0.027

The selected and aggregated candidate explanatory variables are the following:

Table 3.27: Specified independent variables

Table 27: final candidate Explanatory Variables

Variable	Levels
Age	-34 35 - 44 45 - 54 55 -
Professional Status	in Army Not in Army
Sex	Male Female
Type of Itinerary	Round Trip One Way
Target_Buy Contract	Yes No
Class	Economy, Business, PremiumEconomy First
Type of Client	Corporate Public Administration Bank
Type of Flight	National International, Intercontinental
Day of Issuance	Weekend Working days
Day of Departure	Weekend Working days
Day of Return	Weekend Working days
Number of Routes	1 2 3 4 and more
Route	FCO_GOA_FCO FCO_LIN_FCO FCO_TLV_FCO LIN_FCO_LIN OtherRoute
Airline	AIR_FRANCE ALITALIA_CAI_SPA EASYJET Other
Distance KM	
Advance Booking Days	

Distanza_KM
AnticipoPrenotazioneG
AIRLINE

Chapter 4

Modelling the risk of non-fly

4.1 Simplifying the business problem

The research problem that the TMC explicitly required to address is to calculate the probability that a ticket is not flown and, in particular, that it is, either partially or totally, changed or refunded. But this problem is not as easy as it appears at a first sight. In fact, once bought an air ticket, the business traveler can:

1. depart,
2. waive the flight completely, without refund,
3. waive the flight completely, with total refund,
4. fly only a part of the routes included in the ticket and ask refund for the remaining,
5. entirely change the ticket (for getting a new date and/or new routes),
6. fly only a part of the routes included in the ticket and change the remaining (dates and/or routes).

But, if the traveler changes the ticket, partially or totally, then he can again:

1. depart,
2. waive the flight completely, without refund,
3. waive the flight completely, with total refund,
4. fly only a part of the routes included in the ticket and ask refund for the remaining,
5. entirely change the ticket (for getting a new date and/or new routes),
6. fly only a part of the routes included in the ticket and change the remaining (dates and/or routes).

And if the traveler changes the ticket, totally or partially, a second time, then he has again the 6 options listed above. Considering that a business traveler can potentially change ticket an in(de)finite number of times, the theoretical dimension of this problem is virtually infinite. Therefore a first difficulty consists in simplify the business problem in a way that is logically consistent, economically meaningful, computationally feasible and suitable to reflect the impact of each ticket outcome on the TMC's business.

It is clear that tickets' outcomes result from business travelers' choices, listed above. Namely, a ticket can display one of the following outcomes:

1. flown,
2. not flown,
3. total refund,
4. partial refund,
5. total change,
6. partial change.

And if the ticket is partially changed:

- 1 partial change and flown,
- 1 partial change and not flown,
- 1 partial change and total refund,
- 1 partial change and partial refund,
- 1 partial change and 1 total change,
- 2 partial changes.

And similarly, if the ticket is totally changed. Again, each time a ticket is changed, all the above 6 outcomes are possible. It is to be noted that outcomes cannot be aggregated based neither on their final state, nor on their initial state, nor on any intermediate state. In fact, there is a great difference, from both a statistical and an economic perspective, between the apparently same outcome at subsequent 'levels' (so to say: it is possible to consider the 6 possible outputs of the first issued ticket as belonging to the first level, the 6 possible outputs of the second issued ticket, due to a change, as belonging to the second level, and so forth).

For example, the first-level outcome 'total refund' differs from a second-level 'total refund' because, from the statistical point of view, the 'trajectory' of the first one is 'issuance - total refund', thus the ticket underwent a single transition between the initial state and the absorbing state. While the trajectory of the second one is, for example, 'issuance - partial change - total refund', so the ticket passed from the initial state to an intermediate state, then to the absorbing state, realizing 2 distinct transitions. Indeed, a second-level 'total refund' also differs from the other possible second-level 'total refund', which would have trajectory: 'issuance - total change - total refund', as the intermediate state differs, though the number of transitions, the final and the absorbing states are the same.

The difference between the mentioned outcomes, in economic terms, is more complicated, because:

- Seneca stipulates personalized contracts with its clients, so there are nearly as many different economic conditions, producing different economic outputs for each ticket outcome, as client companies.

- Different airlines offer different conditions, with reference to the onerousness of changes and possibility of refund, through different fares and tickets.

The economic difference between 'equal' tickets outcomes at different levels depends on both the conditions applied to clients by Seneca and the conditions applied to Seneca by airlines, which vary with the degree of flexibility of the ticket.

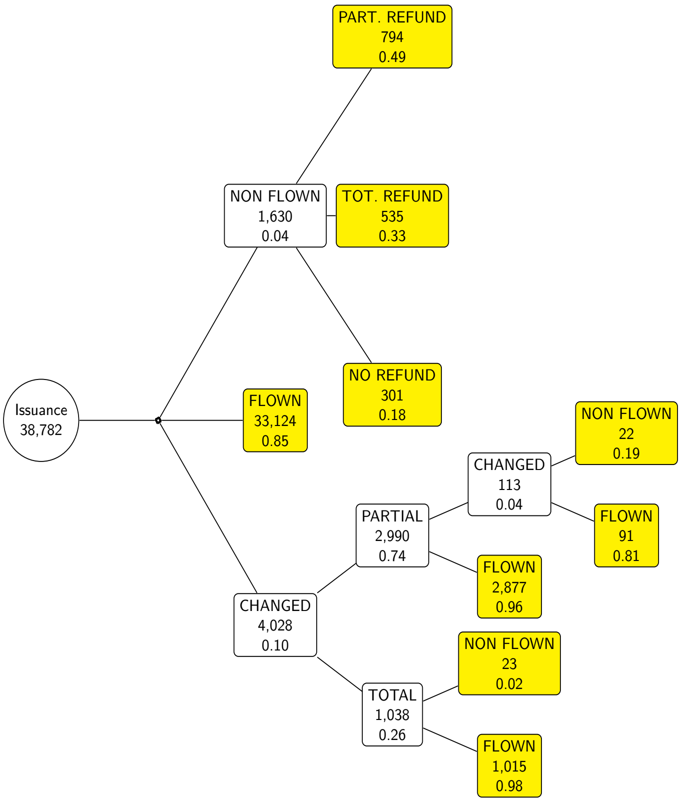
For example, imagine that a client company stipulated a Target Buy contract with the TMC, providing that it always pays 100 euros for flying in economy class from Rome to Brussels, plus a brokerage commission equal to 15 euros for each fight and a penalty of 10 euros for each change. Then, hypothesize that the client requests 2 tickets for the specified route, paying a total amount of 230 euros. Let Seneca purchase, on behalf of that client, 2 tickets priced 60 euros each, the degree of flexibility of which corresponds to the possibility of change with a penalty of 20 euros and the possibility of refund up to 50% of the price. If the outcome of one ticket is first-level 'total refund', then the economic output, for the TMC, is equal to: $100 + 15 - 60 - 100 + 30 = -15$ euros, while, if the outcome of the other ticket is second-level 'total refund', then the economic output, for the TMC, equals: $100 + 15 - 60 + 10 - 20 - 100 + 30 = -25$ euros.

A further example. Imagine that the Target Buy contract, for same client company, provides also that it always pays 1,000 euros for flying from Brussels to New York, with a penalty of 250 euros for each change. The client requires 2 tickets to reach Brussels from Rome, then New York from Brussels. Seneca buys, on behalf of that client, 2 non-refundable tickets, covering both routes, priced 500 euros, with a penalty of 70 euro in case of change. Then, one ticket is first totally changed, then totally refunded, producing an economic result, for the TMC, equal to: $100 + 1,000 + 15 - 500 + 250 + 10 - 70 - 1,100 = -295$ euros. With reference to the other ticket, the client asks to anticipate the date of the flight from Rome to Brussels, leaving the flight from Brussels to New York unchanged (partial change). But, just a few days before the date of departure, the business traveler has a hitch impeding him to do the journey, so he asks the total refund. Thus, the economic output of the transaction is: $100 + 1,000 + 15 - 500 + 10 - 70 - 1,100 = -545$ euros.

As shown in the above examples, the economic result of each transaction, for Seneca, depends on a lot of factors, most of them out of its control, but the more numerous the changes, the more probable that the TMC attains a bigger loss. And, of course, if the client company asks the refund, it is more probable that the economic result is negative, compared to the situation in which the refund is not required. It is also evident that a more expensive and flexible ticket is convenient in case of changes and/or refund, whence the importance of estimating the probability of each outcome of the ticket.

As there is no literature helping in the specification of the dependent variable(s), different options have been explored, based on the sample distribution and on the information collected through the qualitative investigation of the context. Finally, the following tree specification, reflecting both the logical and the dynamic structure of the business travelers' decision-making process, has been selected:

Figure X: Tree of air tickets' outcomes



The leaves of the tree, colored in yellow, represent the absorbing states of the tickets. The observed tickets for which the outcome is non-missing are 38,782, the missing data are indeed right-censored, as the departure date was posterior to that of the data collection and no change nor refund was required up to that time. Censored data are excluded from estimation sample and subsequently included in the forecasting one, where the full trajectory of those tickets will be recorded. The formalization shown in the tree allows to simplify the problem, assuming that it is composed by 6 variables:

- $Y1$ is a trinomial variable, which can assume the modalities: Flown, Non Flown, Changed;

- $Y2 \mid Y1 = NONflown$ is a trinomial variable, which can assume the modalities: No Refund, Total Refund, Partial Refund;

- $Y3 \mid Y1 = Changed$ is a binary variable, which can assume the modalities: Totally changed, Partially changed;

- $Y4 \mid Y3 = Totallychanged$ is a binary variable, which can assume the modalities: (Totally changed and) Flown, (Totally changed and) Non Flown;

- $Y5 \mid Y3 = Partiallychanged$ is a binary variable, which can assume the modalities: (Partially changed and) Flown, (Partially changed and, again, Partially) Changed;

- $Y6 \mid Y5 = Changed$ is a binary variable, which can assume the modalities: (twice Partially changed and) Flown, (twice Partially changed and) Non Flown.

Such a formalization implies a simplification, allowing to avoid overparametrization and to specify identified models, if compared with a simultaneous one, where only a variable is considered, but it would be a nominal variable, composed by 9 levels (one for each leave), each of which affected by all the relevant explanatory variables.

Conversely, this way a structure of conditional independence relations can hold, although must be empirically verified. Namely, if X is the vector of all candidate explanatory variables, 'subsettable' in J vectors, x_j, and if $Y1$ depends on x_1, then it is possible that $Y2 \perp x_1 \mid Y1 = NonFlown$ and $Y3 \perp x_1 \mid Y1 = Changed$. Similarly, if x_2 is the set of explanatory variables for $Y3$, it is possible that $Y4 \perp x_2 \mid Y3 = Total$ and $Y5 \perp x_2 \mid Y3 = Partial$. Finally, if x_3 is the set of variables correlated with $Y5$, it is possible that $Y6 \perp x_3 \mid Y5 = Changed$.

Thus, if no residual effect, of 'grandparents' variables is empirically detected, the loss of efficiency, due to the progressive diminution of the sample size at descending the branches of the tree, is compensated by the minor number of parameters to be estimated. But the tree formalization of the business problem requires other independence conditions to hold, namely that:

- the business traveler's choice of not asking the refund, asking a partial refund (while flying the rest of routes) or a total one, once he decided not to fly (the first bought ticket) is independent by the alternatives of flying the ticket or changing it;

- when choosing whether to completely or partially change the (first bought) ticket, the business traveler does not consider anymore the possibility of flying it or to waive the flight;

- once totally changed the ticket, the traveler considers only two possible choices: to fly it, or not to fly it (changes are included in the latter, but no further modelled, due to scarcity of observations), independently from the possibilities represented by other nodes of the tree;

- once partially changed the ticket, the traveler chooses between changing it again or flying it, independently from any other node of the tree;
- after changing the ticket twice, the traveler decides to leave or not to leave (further changes are included in the latter, but no further modelled, due to scarcity of observations), not influenced by any other alternative.

Thus, such conditions can be called 'irrelevance of higher-level alternatives', then the discussion on their validity may be included in the well known issue of Independence from Irrelevant Alternatives.

4.2 Models specification

Since the time elapsed between the moment of booking and the departure date is said, by the TMC's experts, to be crucial for predicting the tickets' outcomes, a first attempt was made to model it as a stochastic process, representing the ticket trajectory from the first issuance to the final leaf. But too many errors in recording dates made practically impossible to estimate duration models, namely the Proportional Hazard and the Multi-state Markov models (see: Florens et al, 2008).

Then, many different classifiers were tried:

- Classification Trees;
- Random Forests;

- Hierarchica Models;

- Bayesian models.

But no available independent variable resulted useful to subset observations according to a classification tree. Random forests semed to work pretty well on the training sample, but their performances on the test sample were even worse thant those discussed in the following chapter. No hierarchical model is feasible, because the dataset is too unbalanced, the identifiers of groups are missing in too many cases and it is not sure that an unidentified individual does not belong to an identified group. Bayesian models were considered because of the hope to get some prior information from the experts of the TMC, able to integrate the poor data provided by the corporate dataset. But no further information became available, so that the bayesian approach would have led to the same results of the frequentist one.

However, the qualitative study of the context and the decision tree specification, describing the phenomenon under investigation as the decision-making process of business travelers, make natural to employ discrete choice models. In particular, the MLM resulted the most viable methodological choice, in the present case, especially considering that the various types of MLM can be derived from the general Random Utility Model (Train, 2009).

Thus, the unknown utility of the j-th 'flying choice' for the n-th business traveler, $U_{n,j}$, is a function $V_{n,j}$ of the characteristics of the j-th alternative, $S_{n,j}$, and of the traveler-specific variables, Z_n, plus a random component, $\epsilon_{n,j}$, representing the unobservable elements of $U_{n,j}$. Unluckily, in the present case no information on $V_{n,j}$ is available, thus:

$$U_{n,j} = V_{n,j} + \epsilon_{n,j} = V_j(Z_n) + \epsilon_{n,j} \tag{4.1}$$

As it is logical to assume that the traveler chooses the option yielding the greater utility, then, if the baseline alternative is the j-th, the probability that the n-th business traveler makes the i-th decision is:

$$Pn,i = P[U_{n,i} > U_{n,j}] = P[V_{n,j}+\epsilon_{n,j} < V_{n,i}+\epsilon_{n,i}] = P[\epsilon_{n,j}-\epsilon_{n,i} < V_{n,i}-V_{n,j}] = \tag{4.2}$$

$$= F_{\epsilon_{n,i}-\epsilon_{n,j}}(V_{n,i} - V_{n,j}) = \int 1_{\{\epsilon_{n,i}-\epsilon_{n,j}<V_{n,i}-V_{n,j}\}} f(\epsilon_n)\delta\epsilon_n \tag{4.3}$$

Then, if $\epsilon_n \sim_{IID} Gumbel(\mu, \gamma)$ and, as in the present case, only the characteristics of the decision-maker are available as explanatory variables, the Pure MLN is obtained:

$$U_{n,j} = Z_n\beta_j + \epsilon_{n,j} \tag{4.4}$$

$$f(\epsilon_{n,j}) = exp\{-\epsilon_{n,j}\}exp\{-exp\{-\epsilon_{n,j}\}\} \tag{4.5}$$

$$F(\epsilon_{n,j}) = exp\{-exp\{-\epsilon_{n,j}\}\} \tag{4.6}$$

Where each (level of each) explanatory variable has an alternative-specific parameter, β_j, and the utility is 'normalized', so that it has a scale and the model is basically identified, once set $\beta_k = 0$ if the k-th category is the reference one, thanks to the constant variance of the error term: $Var[\epsilon_{n,j}] = \frac{\pi^2}{6}$. As the difference of two Gumbel-distributed variables has Logistic distribution, if $\epsilon^*_{n,j} = \epsilon_{n,j} - \epsilon_{n,i}$, then $\epsilon^*_{n,j} \sim Logistic(s, m)$ and:

$$F(\epsilon^*_{n,i}) = \frac{exp\{\epsilon^*_n\}}{1 + exp\{\epsilon^*_n\}} \tag{4.7}$$

$$Pn,i = P[\epsilon_{n,j} < Vn,i + \epsilon_{n,i} - Vn,j] = \tag{4.8}$$

$$= \int \prod_{j \neq i} exp\{-exp\{-(Z_n\beta_i + \epsilon_{n,i} - Z_n\beta_j)\}\} exp\{-\epsilon_{n,i}\} exp\{-exp\{-\epsilon_{n,i}\}\} \delta\epsilon_{n,i} = \tag{4.9}$$

$$= \frac{exp\{Z_n\beta_i\}}{1 + \sum_j Z_n\beta_j} \tag{4.10}$$

Generally, the Pure MLM is interpreted in terms of odds: $odd_{i,j} = exp\{Z_n(\beta_i - \beta_j)\}$ so that if δ_{z_m} is an increment in the m-th explanatory variable, then its effect on the probability that the decision-maker chooses the i-th alternative, over the reference choice, k, equals to:

$$\frac{odd_{i,k}(Z_n, z_{n,m} + \delta_{z_m})}{odd_{i,k}(Z_n, z_{n,m})} = exp\{\beta_{m,i}\delta_{z_m}\} \tag{4.11}$$

and the marginal effect of $z_{n,m}$ equals $\frac{\delta P_{n,i}}{\delta z_{n,m}}$. The Maximum Likelihood Estimate (MLE) of the model is simply, as the log likelihood function is globally concave. Its functional form is the following:

$$L = \prod_{n=1}^{N} (\prod_j P_{n,j}^{1_{\{n,j\}}}) \tag{4.12}$$

$$ln(L) = \sum_{n=1}^{N} (\sum_j 1_{\{n,j\}} ln \left(\frac{exp\{Z_n\beta_i\}}{\sum exp\{Z_n\beta_j\}} \right) \tag{4.13}$$

where $1_{\{n,j\}}$ is an indicator function, equal to 1 if the n-th business traveler chooses alternative j, 0 otherwise. To facilitate the economic interpretation of estimated coefficients, usually they are standardized: $\hat{\beta_{j,m,SD}} = [exp\{\beta_{j,m}\} - 1] * 100$, then $\hat{\beta_{j,m,SD}}$ equals the percentage variation in the odd ratio, produced by a unitary increment of the m-th independent variable.

Whether the random components $\epsilon_{n,j}$ are not identically distributed and, in particular, they are heteroskedastic, $Var[\epsilon_{n,j}] = \sigma_j^2 \neq Var[\epsilon_{n,i}] = \sigma_i^2, \forall i \neq j$, then the relation between σ_j^2 and the decision-maker's characteristics makes the estimates of the Pure MLN coefficients inefficient, biased and inconsistent, and an Heteroskedastic MLM, or Parametrized Heteroskedastic MLM, must be specified. Its form is the following:

$$U_{j,n} = Z_n\beta_j + \sigma_j\epsilon_{n,j} \tag{4.14}$$

Such a specification does not need the IIA to hold and allows to model also different variances for different decision-makers. For simplicity, considering an only individual, $\epsilon_j \sim Gumbel(0, \sigma_j)$ and

$$f(\epsilon_j) = \frac{1}{\sigma_j}exp\{-\frac{\epsilon_j}{\sigma_j}\}exp\{-exp\{-\frac{\epsilon_j}{\sigma_j}\}\} \tag{4.15}$$

$$F_{\epsilon_j}(z) = \int^{\epsilon_j = z} f(\epsilon_j)\delta\epsilon_j = exp\{-exp\{-\frac{z}{\sigma_j}\}\} \tag{4.16}$$

Then, the probability that the business traveler chooses alternative i is:

$$P_i = P[\epsilon_j < z\beta_i + \epsilon_i - z\beta_j] = \int \prod \Lambda \frac{z\beta_i + \epsilon_i - z\beta_j}{\sigma_j} \frac{1}{\sigma_i} \lambda \frac{\epsilon_i}{\sigma_i} \delta\epsilon_i \tag{4.17}$$

where Λ and λ are respectively the Gumbel density and the Gumbel cumulative function. σ_i can be interpreted as a measure of the degree of uncertainty associated with the expected value of the utility, $E[U_i] = z\beta_i$, due to the unobservable part of the decision-maker's utility, ϵ_i, thus it can be seen as the relative weight of the utility's random component, with respect to its systematic ones, $z\beta_i$, for estimating P_i. It also includes the effect of changes in the available set of alternatives, on the probability of choosing each alternative, in which the violation of IIA consists. Thus, the higher the estimated value of σ_i, the lower the impact of $(z\beta_j) - (z\beta_i)$ on P_i, the greater the effect of $\epsilon_i - \epsilon_j$, the smaller the elasticity effect on P_i. Also this model is estimated through MLE. Comparing the estimates of Pure MLM and Heteroskedastic MLM, including the same variables, is also useful as residuals diagnostics.

Given the tree formalization, presented above, it may be the case that the IIA holds between but not within some set of alternatives, namely those coded as levels of the same Yn variable. For example, it can be possible that, once decided not to fly, the business traveler's choice of not asking the refund is influenced by the possibilities of asking a total refund or asking a partial refund, but not by those of changing the ticket partially or totally. In such a case, each Yn variable can be seen as the s-th cluster of levels of a single variable, including all the nodes of the tree, so that alternatives within each cluster are correlated, but they are independent from alternatives in other clusters. Then, a Nested MLM should be specified:

$$U_{n,j} = Z_n\beta_j + \epsilon_{n,j,s} \tag{4.18}$$

where $Var[\epsilon_{n,j,s}] = \sigma_s^2 = Var[\epsilon_{n,i,s}], \forall j, i \in Nest_s; Var[\epsilon_{n,j,s}] = \sigma_s^2 \neq Var[\epsilon_{n,k,l}] = \sigma_l^2, \forall j \in Nest_s, k \in Nest_l$, i.e. the error variance is constant within each nest, but can differ between nests. Then, the actual probability of choosing alternative j is the joint probability of selecting an alternative within $Nest_s$ and of selecting alternative j between all the alternatives in $Nest_s$:

$$P[j, Nest_s] = P[Nest_s]P[J \mid Nest_s] \tag{4.19}$$

Therefore, the Nested MLM is composed by a two levels specification, plus a connecting term:

$$P[Nest_s] = \frac{exp\{Z_n\beta_j + \lambda_s IV_{n,s}\}}{\sum\limits_{\forall Nest_h} exp\{Z_n\beta_i + \lambda_h IV_{n,h}\}} \tag{4.20}$$

$$P[J \mid Nest_s] = \frac{exp\{\frac{Z_n\beta_j}{\lambda_s}\}}{\sum\limits_{i\in Nest_s} exp\{\frac{Z_n\beta_i}{\lambda_s}\}} \tag{4.21}$$

$$IV_{n,s} = ln\left(\sum_{j\in Nest_s} exp\{\frac{Z_n\beta_j}{\lambda_s}\}\right) \tag{4.22}$$

Where $IV_{n,s}$ is the so-called Inclusive Value, representing the expected utility, for the n-th business traveler, of chosing the $Nest_s$, $E[U_{n,s}] = E[max_{j\in Nest_s}(Z_n\beta_j + \epsilon_{n,j})]$ and $\lambda \in [0,1]$. Also this model can be estimated through MLE. All the 3 descripted models are considered in the search for a specification for the risk of non-fly.

4.3 Notes on the Independence from Irrelevant Alternatives

For a MLM to be correctly specified, it is necessary that all the following assumptions hold:

- with reference to their outcome, the n tickets can be considered n independent trials, happening in equal conditions;
- each outcome is not (necessarily) independent from the others;
- the probability of each outcome is constant in each trial;
- the listed tickets' outcomes are both exhaustive and mutually exclusive;
- the Independence from/of the Irrelevant Alternatives (IIA)

While the second hypothesis is especially convenient for the present case, in general, it is likely that both the independence and the 'constancy' assumptions hold conditionally to the explanatory variables, if the latter are actually able to synthesize the business and environmental factors determining the travelers' behavior. Also the fourth assumption should hold in the present case, with the limits mentioned above (concerning the virtual infinity of possible outputs). Conversely, the IIA deserves a dedicated discussion, because, if the IIA is violated, then the Maximum Likelihood Estimates (MLE) of the model's parameters are biased and vary with the considered set of alternatives - of tickets' outcomes, in the present case - (McFadden et al, 1978). There are 3 main formulations of IIA, which are not equivalent when individual preference ordering are aggregated through the rank-order method, to obtain a social choice function - a consistent sample distribution, in the present case - (Ray, 1973).

The easiest formulation of IIA is that by Radner and Marschak (1954), which, applied to the present work, results to be the following. Let T be the (infinite) set of all the possible tickets outcomes, y_t, $t = 1, ... + \infty$, and S the finite subset of T, which it is considered and modelled in this study. $y_j \in S \subset T$, $j = 1, ..., 9$. Let C be a choice function, defined on the set of the alternative outcomes, then IIA states that if $y_j \in Img(C(T)) \Rightarrow y_j \in Img(C(S))$. In other words, for IIA to hold it is necessary that the business travelers exhibit a particular behavior (generating y_j), independently on the other elements of the set of alternative behaviors they could assume, whether the set was more or less numerous. For example, IIA requires that, if a traveler needs to partially change the reserved ticket, e.g. because his travel aims at meeting 2 clients in 2 different places, 1 of which is no more available for the reserved date, he will change it partially (so that he will fly and meet the available client, but not the unavailable one), independently on whether or not the option of, e.g., partial refund is available (as getting the refund instead of the partial change would not allow him to reach the other client in a new date).

The most known formulation of IIA is surely that of Arrow (1963), which, in the present situation, can be recalled this way. Let $*$ and $\#$

denote 2 different business travelers, with individual preferences ordering: $\{y^*_{j1}, ..., y^*_{j9}\}$ and $\{y^{\#}_{j1}, ..., y^{\#}_{j9}\}$. Let $C^*(S)$ and $C^{\#}(S)$ be the two travelers' choice/utility functions defined on S, the set of alternatives including y_j. Let A and B be 2 alternatives available to the two business travelers: $A, B \in S$, then, according to IIA, for each possible ticket outcome (or equivalently travelr's possible choice) J, $Ay^*_jB \wedge Ay^{\#}_jB \Leftrightarrow C^*(S) = C^{\#}(S)$. In other words, according to both travelers alternative A is the preferred one and alternative B is the least desired one, independently on all the other possible alternatives, if and only if their utility functions, defined on the set of available choices, are the same. So, for IIA to hold it is necessary to admit that, for example, 2 business travelers prefer to totally change the ticket for a flight on Friday and their least desired option is that of flying the ticket on Friday, independently on the fact that they could or could not change the ticket only partially or ask the refund or not fly at all, because, e.g., their utility function is the same, as they want to spend the weekend with their family, while not giving up the travel.

Finally, the probabilistic formulation of IIA is that of Luce (1958), which, in the present case, can be viewed as follows. Let $P_s(A)$ indicate the probability of choosing alternative A among the elements of S, and $P(A/B)$ denotes the probability of choosing alternative A among A and B. For each couple of alternatives $A, B \in S \subset T < +\infty$, if $P(A/B) \neq 0 \Rightarrow \frac{P(A/B)}{P(B/A)} = \frac{P_s(A)}{P_s(B)}$. And, if IIA holds, then $\frac{P_s(A)}{P_s(B)} = \frac{P(A)}{P(B)}$. In other words, for IIA to be valid, the probability of each decision of business travelers, choosing among two different alternatives, must be independent on all the other available alternatives. So, for example, IIA requires that if the odd of fly is 0.8, then it stays equal to 0.8 whether the only alternative for the business traveler is that of not fly, whether he can also choose to change the ticket partially, whether he can also change it totally and so forth.

The formalization of Luce (1958) leads directly to understand the importance of IIA for MLM: in the MLM any category can indifferently be considered as the base in the logit transformation, as it is always possible to convert the ratio from one formulation to another, provided that the odd between two alternatives is invariant to the introduction or deletion of other alternatives in the considered set. I.e: if $k, l, m \in S$ and x_k, x_l, x_m are the characteristics of the corresponding choice, while s is the vector of the business traveler's characteristics, then IIA is valid $\Leftrightarrow \frac{P(k|S)}{P(m|S)} = \frac{exp\{C(x_k,s)\}}{exp\{C(x_m,s)\}} \Rightarrow P(k/m) \perp l|S$.

Although, considering the previous examples and the fact that business travelers' choices are not arbitrary, but led by working motivations mainly independent on their own will, the circumstance that the phenomenon under investigation should theoretically comply with IIA requirements does not guarantee that IIA is effectively valid in the MLM which can be employed. In fact, as highlighted by McFadden et al. (1981), violations of IIA do not concern the structure of the choices themselves, but the particular specification of the used model. Thus, for any set of choices, for any phenomenon, IIA can hold for one specification of the explanatory variables, but not for another one. Indeed, violations of IIA can be seen as a specific kind of omitted variable bias, as this property holds only if the omitted variables are independent and random.

Whence the difficulty of knowing a priori whether the MLM can be correctly used, as the omitted variables are absent because they are not observed. Of course, it is possible to hypothesize those omitted variables, their relations and their influences on the variables of interest, according to the available knowledge of the context of the study. But there is no way to ascertain a priori that such a guess is indeed correct and that no other omitted variable affects the choice. While, for an ex-post verification of IIA, various tests have been developed.

There are two types of statistical tests for IIA. The first one is very intuitive, it is the easiest to implement and consists in estimating MLM on different subsets (and also on the full set) of the possible alternatives, among which the choice is made. Then the parameters of the estimated models are compared and, if they are not significantly different, IIA is verified to hold. The most popular tests of this kind are: Hausmann-McFadden's test of misspecification (1984), and McFadden-Train-Tye's test (1981), Horowitz's test (1981) and Small-Hisiao's test (1985), which are all variants of the Likelihood Ratio test. Unfortunately, those tests, relying on asymptotic theory, have very disappointing finite sample properties. In particular, the sizes of these tests are affected by serious distorsions (Cheng and Long, 2007) and their power is substantially too low to make them reliable (Fry and Harris, 1996).

The second type of statistical tests for IIA is model-based. It consists in estimating more general models, not requiring the IIA property, then testing the constraints on parameters to zero, which reconduct the models to the bond of IIA. Such general models, which also constitute an alternative to MLM whenever IIA does not hold, are: DOGIT (Gaudry and Dagenais, 1979), Multinomial Probit, Nested Logit and Mixed Logit (Train, 2003). The constraints on parameters can be tested through the usual Wald test, Likelyhood Ratio test and Lagrange Multiplier (LM) score test (in the formulation of Tse, 1987). The finite sample properties of 2 tests of this kind, namely the LM and its extension by King and Wu (1993): the Locally Most Mean Powerful (LMMP) test, were investigated by Fry and Harris (1996) throug a Monte Carlo study. They found that the empirical size of LM is far below the nominal one for any sample size, but also the LMMP test is affected by undersize distorsion, leading to over-accptance of the null hypothesis of IIA. Nonetheless, their power is more satisfying, and their size distorsion less serious than that of the tests based on the partitioning of the choice set. However, the main problem with this kind of tests is that of identification, as the more general is the model, the more numerous are the parameters to be estimated, the number of which, moreover, affects the power of the test.

Though the problem of verifying IIA seems irresolvable, McFadden et al (1978) made an encouraging consideration: whenever IIA does not hold, the usual residuals diagnostics and goodness-of-fit tests highlight that there is a problem of misspecification. Therefore, it can be assumed that if the residuals are well behaved, the model fits well and the qualitative knowledge of the context leads to hypothesize that IIA can hold, then it is very likely that it actually holds.

4.4 Variables selection

Whenever, as in the present work, there is no literature guiding the model specification, any specification search procedure must be necessarily data-driven, implying some difficulties. First, from a theoretical perspective, the contingency, fortuity chatacter of the relations between variables, portrayed by the obtained model. As any sample is somehow randomly selected, the correlations found in sample may not apply to other samples of the reference population, nor to the latter. Whether it is the case, the selected specification, missing some important predictors or including some unrelevant ones, describes only the sample, thus the knowledge retrieved from the data is not extensible to the population, and, more important to this work, the model is not able to reliably predict out of sample realizations of the phenomenon of interest. The cause of the sample non-representativeness is essentially a selection bias.

For the present work there was not a real sampling procedure: with reference to the cross-sectional dimension, the whole population of air tickets intermediated by Seneca is observed, but with respect to the time dimension, the dataset covers only a small temporal interval. Therefore, a model selected based on the available dataset could display a poor forecasting performance due to 2 sources of selection bias. First, the eventual difference in business travelers' behavior in different periods of the year (as just some moth of observation are available) and in different years (e.g. it may be the case that in a certain year an exibition, a war, an epidemic caused a modification of their propensity to change or waive the flights). Second, as there are some missing data, the eventual difference in proportions of clients and tickets with certain modalities of the explanatory variables, between the observed and the unobserved realizations of the 'non-fly phenomenon'. While the bias induced by the latter can be tested, as it is done in the chapter dedicated to the description of the data, and eventually corrected through weights, there is no way to quantify or avoid the bias due to the small time window.

A second problem with data-driven specification search is that, substantially, the only feasible practices are forward and backward stepwise procedures, eventually combined. Indeed, in the present case, a backward stepwise is unviable, due to the limited degrees of freedom available. But the use of stepwise procedures is greatly discouraged, because the selected explanatory variables can be non-significant, although with small P-values, and the excluded ones can be important predictors, though their P-values appear over the (arbitrarily) set acceptance threshold (Harrell, 2001).

Freedman (1983) showed the paradoxical results obtainable through these methods: he generated independent normal variables, applied stepwise and found significant correlations between them. Many other researches confirmed that stepwise techniques lead to the selection of models with a small number (if compared to that of all the actually significant regressors) of explanatory variables, with highly significant coefficients and inflated values of R-squared and R-squared-like statistics, even though covariates are not even minimally correlated with the dependent variable, when the number of candidate regressors is very close to the size of the estimation sample (Lukacs et al., 2010; Flom and Cassell, 2007; Foster and Stine, 2006).

This so called 'Freedman Paradox' is due to the downward biasedness of standard errors of regression coefficients, producing too narrow confidence intervals, estimates biased in absolute value and too small p-values for the selected model, which will then show a poor forecasting performance (Shatland et al., 2008). Moreover, the selection is usually unstable, sensitive to small perturbations in the data: adding or deleting a small number of observations, the model resulting to be the 'best' one can change (Steyerberg et al, 2000).

In particular, when the aim is obtaining a reliable predictor, the main problem is the "agonizing process of choosing the 'right' critical p-value in stepwise regression" (Shatland et al., 2008, p. 2). Many authors suggest to keep all the candidate regressors, or to set a high acceptance threshold (e.g. p-values of 0.5, see: Harrell, 2001; Steyerberg et al, 2000; Steyerberg et al, 2001), but if the problem is the scarcity of degrees of freedom, this option is likely to be unviable.

Although many researchers conclude that stepwise should be completely avoided, in practice it is necessary, in absence of an economic theory, as in the present case. Thus, a forward stepwise is employed for the model selection, but cautiously taking into account all the mentioned issues. Moreover, here stepwise is only the first step of the predictive specification search, it is not done automatically and some useful shrewdness are added to the standard procedure. In particular, if K is the number of all available candidate explanatory variables and k is the number of those entering a single model, in the case of the MLM, the iterative procedure is composed by $J \leq K-1$ iterations, each consiting in the followinf steps.

1. For $j = 0, 1, ..., J$, the j-th iteration starts estimating $3 * (K-j)$ models with $k = j+1$: a Pure MLM, an Heteroskedastic MLN and a Nested MLM, for each of the $K-j$ candidate regressors (since iteration 1, this is done augmenting $b_{(j-1)}$ specification - see the following - with the remaining candidate regressors).

2. Within each triplet, the 3 different specifications, including the same regressor(s), are compare in terms of log-likelihood value, McFadden R-squared and likelihood ratio test, to find an eventual correctly specified model. In fact, such a comparison has also a diagnostic function: if Pure MLN is the 'best' one, residuals are homoskedastic and IIA holds; if Nested MLM is to be preferred, IIA holds only within 'clusters' of alternatives; if Heteroskedastic MLM is the most sound, residuals are heteroskedastic and IIA does not hold.

3. All the correctly specified models, with all the coefficients' p-values minor or equal to 0.3 are included in the bouquet of 'best' specifications (the 0.3 threshold is fixed on empirical basis, to obtain a manageable number of 'best' models, while including at least an explanatory variable for the last branches of the tree, as it is necessary to make predictions).

4. Among these models, the best specification b_j is selected, again based on log-likelihood value, McFadden R-squared and likelihood ratio test.

5. Steps 1-4 are iterated until no candidate regressor remains, or it does no entry b_j with a significant coefficient.

Thus, the output of this specification search procedure is not a single model, but rather a bouquet of models, found along a single trajectory b in the space of all the possible specifications. In practice, the 'pure' stepwise was employed only for reducing the cardinality of the models space, equal to 6^K, to the manageable number of $3K$. The selected models should correctly describe the relations between variables, a first check of it consists in verifying that the coefficients for the same independent variable display the same sign and not too different values in different, nested estimated models. Afterwards obtained models must be validated, in order to both exclude the bias, which could be present due to the use of (though a more sofisticated) stepwise, and evaluate their forecasting performance.

To the first aim, cross-validation and bootstrap techniques are not reliable enough, because subsetting a database which is already small, compared to the number of the candidate (mainly nominal) explanatory variables (and of their levels), is likely to increase the biasedness of estimates, while implying a loss of efficiency; while re-sampling from a dataset eventually affected by selection bias can propagate the bias throughout the replicas. Luckily, for the present work completely new data, collected from the same source, are going to be available, allowing external validation, which permits a trustworthy evaluation of the prediction error (Shtatland et al, 2008).

But, given the explained issues with data-driven specification search procedures, two other alternatives, obtained from the previously selected models through averaging, are added to the set of options to be validated, checking whether corresponding forecasts are more accurate. Coefficients and forecasts averaging are easy shrinkage techniques, which in some works were shown to substantially improve the performance of predictors (see: Clemen, 1989 for a wide review). Thus, besides the forecasts produced by the set of selected models, the two alternative forecasts sets, obtained averaging prediction probabilities (as it makes no sense to average predicted choices) are considered. Shtatland et al. (2008) suggest the following easy and useful formula, for computing the averaged prediction probability $\bar{P}$:

$$\bar{P} = \sum_{m=1}^{M} w_m \frac{1}{1 + exp\{-\beta_{0,m} - \beta_{1,m} Z_1 - ... - \beta_{k,m} Z_m\}} \tag{4.23}$$

Finally, the relations between the dependent Y_f and independent X_k variables, portrayed in the selected models, are verified in a nonparametric framework, in order to check the distribution-free association between variables. To this aim, chi-squared tests of independence, between the dependent variables and each candidate explanatory variable, are performed.

This method is chosen because the resulting P-values are measures of association comparable across different variables, of various nature (nominal, ordinal, continuous, which are temporarily discretized in classes) and support. For each dependent variable, a nonparametric selection of explanatory ones, parallel to the forward stepwise procedure, is realized as follows. Let K be the number of candidate explanatory variables and S the number of those which will be selected in the end.

1. The independence of Y_f and X_k, $\forall k = 1, ..., K$ is tested.

2. The variable for which the P-value is the lowest, X_s^*, is selected as significantly related with Y_f.

3. The independence of Y_f and X_k, $\forall k = 1, ..., K - s$, conditional to X_s^*, is tested.

Steps 2 and 3 are iterated $S \leq K$ times, until no further variable significantly related to the dependent one, conditional to all the previously selected variables, is found.

Although such a procedure is implemented in order to seek confirmation of the results of the parametric variables selection, it too is not free from problems. In fact, it suffers from the gradual loss of efficiency, due to the progressive shrinking of the sample, necessary to test the relations of conditional independence. Therefore, it can happen that variables 'truly' highly associated are not selected, because in some shrunk samples not all of their levels are observed, at worst all the tickets display the same value. Therefore, the chi-square based procedure tends to underestimate the number of significant explanatory variables.

4.5 Estimations resuts

As already mentioned, the business traveler is identified in 25,749 cases in the estimation sample and just in 7,506 cases in the forecasting sample, therefore the variables selection procedure described above is repeated on this subsample of the estimation sample, in order to avoid favouring the inclusion of the variables with less missing data (model estimated on this subsample are marked with "a").

Moreover, variables sex and age of the traveler are recorded just in 2,432 cases in the estimation sample and in 1,868 cases in the forecasting sample, thus the variables selection is repeated once more on this subsample, but only for descriptive purposes (model estimated on this subset of data are marked with "b").

With reference to the residuals diagnostic, the Pure MLM is the best fitting specification, between the 3 discrete choice models considered, so the assumptions of homoskedasticity and non-correlated alternatives hold. Signs and values of different estimates for coefficients referring to the same variable are consistent and stable throughout models. But the goodness of fit measures are very disappointing, highlighting the very low correlation with the dependent variable and the definitely insufficient explanatory power of available independent variables. The full outputs for the bundle of selected models is available upon request.

Regrettably, also the comparison between the chi-square based variables selection and the results of the stepwise based procedure, shown below, are not encouraging.

Table 4.1: Comparison of parametric and non-parametric variables selections results

	Change/Flown/Not Flown		Non/Total/Partial Refund	
	CHI-SQ.	**MLM**	**CHI-SQ.**	**MLM**
Advance Booking Days	X		X	
Airline	X	X		
Class				X
Departure in Weekend	X			
Issuance in Weekend		X		X
Return in Weekend		x		X
Distance KM	X		X	x
Number of Routes	X	X	X	
Route	X		X	
Target_Buy Contract	X	X	X	
Type of Client	X		X	X
Type of Flight			X	X
Type of Itinerary		x		
	Total/Partial Change		Partial Change and Flown/Change	
	CHI-SQ.	**MLM**	**CHI-SQ.**	**MLM**
Advance Booking Days	X	X		X
Airline	X		X	
Class				X
Departure in Weekend				X
Issuance in Weekend				
Return in Weekend		X	X	
Distance KM		X	X	
Number of Routes	X		X	
Route				
Target_Buy Contract		X		X
Type of Client	X			X
Type of Flight		X		x
Type of Itinerary		X		
	Partial Changes and Flown/N		Total Change and Flown/Not	
	CHI-SQ.	**MLM**	**CHI-SQ.**	**MLM**
Advance Booking Days				X
Airline				
Class			X	X
Departure in Weekend		x		
Issuance in Weekend				
Return in Weekend		x		
Distance KM		x	X	
Number of Routes				
Route			X	
Target_Buy Contract				X
Type of Client				
Type of Flight				
Type of Itinerary	X			

Red crosses indicate that the corresponding variable is selected in MLM but not in Chi-square selection. The reason of concern is not only the inconsistency of most of selected/dropped variables, but also the fact that the non-parametric method is expected to select less variables than the whole set of actually significant regressors, while here it includes more variables than the stepwise procedure for MLM. This evidence confirms that the available information is insufficient to effectively model the phenomenon under investigation, thus results must be read very cautiously.

Estimation results indicate that the smaller the number of routes, the higher the propensity, of the business traveler, to fly the first purchased ticket, while if the flight covers 5 routes or more, it is more likely that the first ticket is not flown. An eventual 'speculative' behavior of Target-Buy clients is not supported by estimates, in fact the probability that a client with this type of contract flies the first ticket is slightly higher than the one that he changes or waivers the flight. A higher positive effect on the propensity to fly is found for the variable indicating that the first ticket was issued during the weekend. This evidence suggests that tickets reserved in the weekend, when business traveler should not work, are for 'emergencies' and, as such, they are normally flown.

When the flight is provided by Alitalia and Easy Jet the probability that it is flown are higher that those for (other) low cost airlines, but smaller that those for (other) non-low cost air companies. If the date of return is during the weekend, then it is likely that the business traveler changes the flight, possibly to be able to spend the weekend together with his family. On the contrary, flights departing during the weekend, which can represent 'mandatory' journeys, and round trip (RT) itineraries are more likely to be flown, maybe because a RT is often chosed to save money, being less expensive than 2 one-way tickets, when the date of the way back is basically sure.

Clients with Target-Buy contracts are nontheless clearly more likely to ask a total refund for non-flown tickets, thus maybe the procedure for asking directly Seneca a refund are much easier and shorter than those required by the airlines, and Target-Buy clients profit from them, while not abusing of the possibility to waive the flight often. Passengers flying in first class are less propense to ask refunds, maybe because money is not an issue for them. The contrary appears for business travelers working in banks, while refunds are rarely asked by those working in the public administration and other corporations. Refunds are few allso for national flights, when compared with international and intercontinental ones. Consistently with the high propensity to change (the date of, then just partially change) flight when the date of return is during the weekend, this variable also positively affect the possibility that a partial, rather than a total, refund is asked.

Given that the first ticket is changed, then Target-Buy clients tend to change it totally. Total changes are also more likely than partial ones if the covered distance is great or if the itinerary is RT. Conversely, partial changes are more probable if the flight is national, if the booking is made with abundant advance and if the date of return is provided in the weekend.

Consistently with these findings, whether business travelers with first class tickets totally change the flight, that rarely happens, then they tend to fly the changed ticket, that is also true for clients with Target-Buy contract. The less the days between the total change of the ticket and the departure date, for the new flight, the higher the propensity of the traveler to fly the changed ticket.

Once partially changed the ticket, the higher the number of days of advance booking, the lower the probability that the changed ticket is flown. While passengers in first class are more likely to fly the changed ticket, as well as those departing during the weekend. The opposite relation is found for Target-Buy clients, which are more likely to ask multiple changes, once partially changed the first flight, for business travelers working in the public administration and in non-bank corporations, and for national flights.

Among business travelers who partially changed the ticket twice (not flying the ticket after a first partial change), those who should depart or return during the weekend are less likely to fly. Moreover such a trend is slightly more marked if the flights cover long distances.

The estimations outputs for selected models is reported in the tables below.

Table 4.2: Estimation output for Y1

Variable	Estim.Coeff.	*Signif.*
FLOWN:(intercept)	1.356	$***$
NONflown:(intercept)	-0.685	$***$
FLOWN:Nroutes2	0.726	$***$
NONflown:Nroutes2	0.135	$**$
FLOWN:Nroutes3	0.927	$***$
NONflown:Nroutes3	0.517	$**$
FLOWN:Nroutes4	0.682	$***$
NONflown:Nroutes4	0.366	$***$
FLOWN:TargetBuy	0.220	$***$
NONflown:TargetBuy	0.094	
FLOWN:IssWknd	0.490	$***$
NONflown:IssWknd	0.279	
FLOWN:ALITALIA	0.116	
NONflown:AIRALITALIA	-0.301	$*$
FLOWN:EASYJET	1.072	$***$
NONflown:EASYJET	-0.641	$**$
FLOWN:Other	0.345	$**$
NONflown:AIROther	-0.453	$**$
LogLikelihood:	-19172	
McFadden Rsq.	0.017	
Likelihood ratio test : chisq :	679	

Sign. Codes: $*$: $Pvalue <= 0.1$; $**$: $Pvalue <= 0.05$: $***$: $Pvalue <= 0.01$

Table 4.3: Estimation output for Y1 a

Variable	Estim.Coeff.	*Signif.*
FLOWN:(intercept)	1.096	$***$
NONflown:(intercept)	-2.398	$***$
FLOWN:OccCh	0.170	$***$
NONflown:OccCh	0.177	$***$
FLOWN: OccTrav	0.550	$***$
NONflown:OccTrav	0.630	$**$
FLOWN:FreqWai	1.050	$***$
NONflown:FreqWai	0.956	$**$
FLOWN:ALITALIA	-0.133	$***$
NONflown:ALITALIA	-0.164	$**$
FLOWN:EASYJET	0.596	$***$
NONflown:EASYJET	-0.424	
FLOWN:AIR.FRANCE	-0.377	$***$
NONflown:AIR.FRANCE	0.317	$**$
FLOWN:AdvBook	-0.005	$***$
NONflown:AdvBook	-0.002	
FLOWN:RT	0.587	$**$
NONflown:RT	0.783	
FLOWN:OccTravC	-0.004	
NONflown:OccTravC	-0.010	
LogLikelihood:	$-10,979$	
McFadden Rsq.	0.037	
Likelihood ratio test : chisq :	843	

Sign. Codes: $*$: $Pvalue <= 0.1$; $**$: $Pvalue <= 0.05$: $***$: $Pvalue <= 0.01$

Table 4.4: Estimation output for Y1 b

Variable	Estim.Coeff.	$Signif.$
FLOWN:(intercept)	2.382	$***$
NONflown:(intercept)	-0.906	$***$
FLOWN:Age	0.109	$***$
NONflown:Age	0.168	$***$
FLOWN:AdvBook	-0.021	$***$
NONflown:AdvBook	-0.010	$**$
FLOWN:National	-0.360	$***$
NONflown:National	-0.567	$**$
FLOWN:TargetBuy	-0.363	$***$
NONflown:TargetBuy	0.367	$**$
FLOWN:Distance	$-4.9e^{-5}$	$***$
NONflown:Distance	$-2.6e^{-5}$	
LogLikelihood:	$-2,180$	
McFadden Rsq.	0.019	
Likelihood ratio test : chisq :	86	

Sign. Codes: $*$: $Pvalue <= 0.1$; $**$: $Pvalue <= 0.05$: $***$: $Pvalue <= 0.01$

Table 4.5: Estimation output for $Y2 \mid Y1$

Variable	Estim.Coeff.	*Signif.*
TOTchange:(intercept)	-1.758	$**$
TOTchange:TargetBuy	0.648	$***$
TOTchange:National	-0.404	$***$
TOTchange:AdvBook	-0.008	$**$
TOTchange:Distance	$2.1e^{-5}$	$**$
TOTchange:RT	1.0142	
TOTchange:RitWknd	-0.207	
LogLikelihood:	-1280	
McFadden Rsq.	0.019	
Likelihood ratio test : chisq :	49	

Sign. Codes: $* : Pvalue <= 0.1; ** : Pvalue <= 0.05 : *** : Pvalue <= 0.01$

Table 4.6: Estimation output for $Y3 \mid Y2$

Variable	Estim.Coeff.	*Signif.*
PChNONflown:(intercept)	-3.782	$***$
PChNONflown:AdvBook	-0.022	$*$
PChNONflown:First	-0.841	$**$
PChNONflown:DepWknd	-1.146	
PChNONflown:TargetBuy	0.425	
PChNONflown:CORPORATE	1.063	
PChNONflown:PA	0.839	
LogLikelihood:	-234	
McFadden Rsq.	0.043	
Likelihood ratio test : chisq :	21	

Sign. Codes: $* : Pvalue <= 0.1; ** : Pvalue <= 0.05 : *** : Pvalue <= 0.01$

Table 4.7: Estimation output for $Y3a \mid Y2a$

Variable	Estim.Coeff.	*Signif.*
PChNONflown:(intercept)	-3.862	$***$
PChNONflown:AverTrav	0.596	$**$
PChNONflown:AdvBook	-0.024	$**$
PChNONflown:First	0.831	$**$
PChNONflown:OccTravC	0.017	$*$
PChNONflown:DepWknd	-1.161	
PChNONflown:NflightsC	$5.7e^{-5}$	
LogLikelihood:	-233	
McFadden Rsq.	0.047	
Likelihood ratio test : chisq:	23	

Sign. Codes: $* : Pvalue <= 0.1; ** : Pvalue <= 0.05 : *** : Pvalue <= 0.01$

Table 4.8: Estimation output for $Y3b \mid Y2b$

Variable	Estim.Coeff.	*Signif.*
PChNONflown:(intercept)	-5.454	$***$
PChNONflown:Age	0.785	$**$
LogLikelihood:	-52	
McFadden Rsq.	0.051	
Likelihood ratio test : chisq :	6	

Sign. Codes: $* : Pvalue <= 0.1; ** : Pvalue <= 0.05 : *** : Pvalue <= 0.01$

Table 4.9: Estimation output for $Y5 \mid Y3$

Variable	Estim.Coeff.	$Signif.$
PCh2NONflown:(intercept)	-2.0565	$***$
PCh2NONflown:Distance	$1e^{-4}$	$*$
LogLikelihood:	-23	
McFadden Rsq.	0.065	
Likelihood ratio test : chisq :	3.3	

Sign. Codes: $*$: $Pvalue <= 0.1; **: Pvalue <= 0.05 : *** : Pvalue <= 0.01$

Table 4.10: Estimation output for $Y4 \mid Y2$

Variable	Estim.Coeff.	$Signif.$
TChNONflown:(intercept)	-2.455	$***$
TChNONflown:AdvBook	-0.076	$*$
TChNONflown:TargetBuy	-1.806	$*$
TChNONflown:First	-1.594	
LogLikelihood:	-71	
McFadden Rsq.	0.088	
Likelihood ratio test : chisq :	14	

Sign. Codes: $*$: $Pvalue <= 0.1; **: Pvalue <= 0.05 : *** : Pvalue <= 0.01$

Table 4.11: Estimation output for $Y4b \mid Y2b$

Variable	Estim.Coeff.	*Signif.*
TChNONflown:(intercept)	-6.475	
TChNONflown:Male	-5.790	$* * *$
TChNONflown:Age	1.768	
TChNONflown:National	2.219	
LogLikelihood:	-7	
McFadden Rsq.	0.494	
Likelihood ratio test : chisq :	14	

Sign. Codes: $* : Pvalue <= 0.1; ** : Pvalue <= 0.05 : * * * : Pvalue <= 0.01$

Table 4.12: Estimation output for $Y6 \mid Y1$

Variable	Estim.Coeff.	*Signif.*
PARTrefund:(intercept)	2.121	$***$
TOTrefund:(intercept)	1.327	$**$
PARTrefund:TargetBuy	−0.371	
TOTrefund:TargetBuy	0.821	$**$
PARTrefund:First	−0.305	
TOTrefund:First	−0.456	$**$
PARTrefund:CORPORATE	−1.045	$**$
TOTrefund:CORPORATE	−0.491	
PARTrefund:PA	−0.828	$**$
TOTrefund:PA	−1.312	$**$
PARTrefund:National	−0.210	
TOTrefund:National	−0.523	$**$
TOTrefund:RitWknd	−0.435	
LogLikelihood:	−908	
McFadden Rsq.	0.05	
Likelihood ratio test : chisq :	96	

Sign. Codes: $*$: $Pvalue <= 0.1$; $**$: $Pvalue <= 0.05$: $***$: $Pvalue <= 0.01$

Estimated coefficients show that the probability of changing the first issued ticket is especially low for occasional business travelers. Frequent flyers normally fly the first air ticket, but when they do not, it is more probable that they renounce to the flight, rather than changing the ticket. Tickets for flights operated by Alitalia are more likely changed, those issued by Easyjet are more probably flown, while flights operated by Air France are more often subject to renounce.

As it may be expected, the longer the advance booking time, the higher the probability of change. Moreover, it is more likely that a business traveler renounces to a round trip (than a one way journey) but it is less probable that he changes the ticket. Travelers working for companies occasionally requiring them to fly tend to change ticket more often. While the probability of changing ticket decreases as the flyer's age increases, the probability of renounce becomes higher for older workers. It is more likely that a national flight is changed, but the effect of the distance covered by the journey is very small. Single-route tickets are changed more often, multi-route first issued ones are more likely flown.

If the ticket is booked during the weekend, then it is highly probable that it will be flown, it will be hardly changed, maybe because journeys organized outside of work are motivated by emergencies. Once totally changed the first purchased ticket, it is much more likely that it will be changed again, or that the worker will renounce to the flight, if the business traveler is a female, maybe because women are often in charge of caregiving, for both children and the oldest members of the family. The particular contract offered by Seneca, Target-buy, appears to make travelers incline to renounce, rather than to change flight.

As already warned, these results are not very reliable. Typically, Discrete Choice Models work better with information about the decision-maker and alternative-specific variables, but in this case they are not available (in sufficient quantity and quality). As a consequence, the TMC is recommended to collect, in the future, more data about the characteristics of the business travelers, as the problems in the present empirical analysis and the qualitative study of the business issue suggest that they should be an important predictor for the behavior of business travelers.

However, a general finding important from the phenomenic perspective, is that there are some characteristics, of the flight, the ticket, the company and the traveler, that can help predicting the behavior of the business flyer. If TMCs or corporate travel departments extensively collect and take them into account, they should be able to choose the optimal fare for each traveler, for each flight.

Chapter 5

Predicting air tickts' outcomes

5.1 Prediction problems

The models, the estimation outputs of which are reported above, describe the estimation sample 'to the best', given the available information. To find the best predictor, all of the models ($m = 1, \ldots, M$) obtained at each step of the first two Stepwise procedures are subjected to external validation on the forecasting sample. According to the ML principle, the outcome, predicted by the m-th model, for the i-th flight, is that ($j*$) for which the estimated conditional probability is the highest of the row vector $\hat{P}_{i,m}[y_i]$:

$$\hat{y}_{i,m} = j* \Leftrightarrow \hat{P}_{i,m}[y_i = j*] = MAX \left\{ \hat{P}_{i,m}[y_i = j]; \forall j \in R_y \right\} \tag{5.1}$$

But the discriminatory power of the available independent variables is exceptionally low, as expected looking at the values of McFadden R-square. Thus, the estimated conditional probabilities are either very close to the sample marginal frequencies or largely biased. In fact, right from the first node ($Y1$):

- the Exact Classification Rate (ECR) of models ("a") including variables describing the past behavior of travelers is less than 33%, so less than that of a completely random classification;

- for the other models ECR = 91% but they always classify any ticket as flown.

Therefore, these models are useless in prediction and do not allow to classify outcomes following the first one.

The causes of such a low discriminatory power, are essentially three:

- In order to predict the travelers' decisions correctly, through Discrete Choice Models, information about the decision-makers and alternative-specific variables would be needed, while sufficient information, in terms of both quantity and quality, is available only about flights and tickets' characteristics (nonetheless, the MLM is the only viable specification, as there is no enough information to estimate more complex models).

- A few client companies purchase most of the flights (80/20 rule of sales), that are nearly identical (they always depart from the airport closest to the company, are operated by the same airline because travelers want to accumulate miles on their frequent flyer programs, are always of the same class and so forth), so that they show the same values of explanatory variables and, among numerous tickes, just very few are not flown.

- The phenomenon could be in influenced by some factors not observed or appreciable in the estimation sample, but present in the forecasting sample.

As there is no possibility to collect further information and the composition of Seneca's client portfolio is given, an attempt is made to act on the last cause of prediction problems. So what element could influence the phenomenon, but is not observed in the estimation sample?

- As mentioned in the literature review, some extant studies about business travel suggest that it may be a sort of seasonal pattern, related to vacation periods.
- The qualitative study of the phenomenon highlights that changes and waivers, in business travel, are often determined by personal and relational reasons, especially by the desire to spend the festivities with the family.

If the estimation sample covered a whole year, this would prompt to insert a dummy indicator of vacation periods D and estimate its effect δ. But, since the observations used for estimation span only seven months, it is just possible to guess a 'vacation effect'.

5.2 Guess-based predictions

Guesses have been used in Statistics since its birth, in the form of subjective prediction (eg. in Delphy method) or, slightly more recently, as prior information, in the Bayesian approach. In the present case, none of the two ortodox solutions is viable. In fact, the literature about business travel and the qualitative study of the context provide a too vague suggestion, not directly expressible in the out-of-sample classification, nor elicitable in a prior distribution of the coefficients of the models.

Thus, an alternative and very simple method is adopted: the guess on the 'vacation effect', δ, is added directly in the estimated models for $Y1$. In practice, a dummy variable D is constructed as the interaction: $D1 * D2$, where each element of $D1$, $d1_i = 1$ if the i-th flight is booked in December or August, zero otherwise; each element of $D2$, $d2_i = 1$ if the departure date of the i-th fligh is in December or August, zero otherwise.

Then, guessing that during vacations the probability that a ticket is not flown should be higher than the probability that it is flown, in consideration of the fact that renounce (Nf) is the most difficult outcome to be forecasted (and the most rarely observed), δ is chosen such that:

$$\delta_{NO} = \hat{\alpha}_{NO} + \hat{\beta}_{NO} - q_{0,NO} \tag{5.2}$$

$$\delta_{Fl} = -(\hat{\alpha}_{Fl} + \hat{\beta}_{Fl}) + q_{0,Fl} \tag{5.3}$$

with q_0 real number, small enough such that:

$$\exists i : if \hat{P}_i[y1_i = Nf] = MAX\left\{\hat{P}[y1 = Nf]\right\} \wedge D_i = 1 \tag{5.4}$$

$$\Rightarrow \hat{P}_i[y1_i = Nf] = MAX\left\{\hat{P}_i[y1]\right\} \tag{5.5}$$

$$\Rightarrow \hat{y}_i = Nf \tag{5.6}$$

but big enough so that the model does not collapse on the intercept and vacation effect only:

$$logit(Y1) = \alpha + \delta D + e \tag{5.7}$$

By adding δD, all of the models' predicting capability improves, from a corporate perspective: although the ECR decreases, now it is possible to correctly classify some tickets as changed and non-flown, the two most economically relevant outcomes (the correct prediction of which generates the highest saving). So the best predictor is selected, as the one yielding:

- Max ECR;

- Max Sensitivity to Ch;
- Max Sensitivity to Nf.

Model 1 (+ δD), the same model that resulted the best in description, is selected. The improvement in forecasting performance is highlighted by the measures reported above (PAS and chi-square test are explained in the following), compared to the benchmark no-change model (predicting always flown - Fl -).

Table 5.1: Guess-based prediction: results for Model 1 + δD

Guess-augment. Model 1	Perf.	**Benchmark**	Perf.
ECR	79 %	ECR	91 %
PAS	2.35	PAS	1.88
ChiSq	43,226	$\chi^2_{0.005}$	16.75
CHANGE		CHANGE	
TP	179	TP	0
TN	22,799	TN	0
Sens.	7 %	Sens.	0
Spec.	85 %	Spec.	100 %
NON FLOWN		NON FLOWN	
TP	3	TP	0
TN	22,975	TN	0
Sens.	5 %	Sens.	0
Spec.	79 %	Spec.	100 %

TP = True Positive; TN = True Negative; Sens. = Sensitivity; Spec. = Specificity.

The comparison shows that, adding the guess on the vacation effect to Model 1, its overall forecasting performance does improve, in economic terms.

But what would be the value of δ if it was possible to estimate it on data actually affected by this vacation effect? In order to approximate it:

- 500 replicas of the independent variables are drawn from the whole database (estimation sample + forecasting sample, for getting a whole year), through a non−parametric, non−stratified bootstrap with replacement (X_{boot});

- For each replica, the dependent variable $\tilde{Y1}$ is generated from the guess-augmented model (1):

$$logi\tilde{t}(Y1) = \hat{\alpha}_1 + \hat{\beta}_{NT1} NumTratte_{boot} + \hat{\beta}_{EW1} EmissWknd_{boot} + \hat{\beta}_{air1} Airline_{boot} + \hat{\beta}_{TB1} TargetBuy_{boot} + \delta D_{boot} + e_1 \tag{5.8}$$

  where $e_1 \sim Gumbel(0, 1)$.

- D is added to specification 1 and all the coefficients (including δ) are estimated ML on each bootstrap sample.

- All the coefficients are significant at the chosen 0.3 P−value threshold, except for $\hat{\delta}$, which is significant only in the 10% of cases.

- Each coefficient is averaged over the 500 samples (α and δ included), $\hat{\beta}_{h,boot} = \frac{\sum\limits_{r=1}^{500} \beta_{h,r}}{500}$.

- 'boot' coefficients are used to predict the real out-of-sample data, plus 500 further bootstrap replicas of the whole database.

The table below shows the results of the guess-based prediction through bootrap coefficients, compared to the benchmark prediction, on the real forecasting sample. The same measures of forecasting accuracy are, on average, 1% higher on the 500 bootstrapped forecasting samples.

Table 5.2: Guess-based prediction: results for bootstrapped coefficients

Boot. M.	Perf.	**Benchmark**	Perf.
ECR	79 %	ECR	91 %
PAS	2.35	PAS	1.88
ChiSq	43,226	$\chi^2_{0.005}$	16.75
CHANGE		CHANGE	
TP	112	TP	0
TN	22,802	TN	0
Sens.	4 %	Sens.	0
Spec.	85 %	Spec.	100 %
NON FLOWN		NON FLOWN	
TP	6	TP	0
TN	22,908	TN	0
Sens.	9 %	Sens.	0
Spec.	78 %	Spec.	100 %

TP = True Positive; TN = True Negative; Sens. = Sensitivity; Spec. = Specificity.

It is possible to see a slight increase in the forecasting performance of the model, thanks to the bootstrap. But the most interesting result is that, comparing predictions made through the selected model, as re-estimated with the vacation effect, with those obtained using the coefficients estimated without the vacation-effect, on the 500 bootstrap forecasting samples (including observations for the whole year), the forecasting performance improves also on the first half of the samples (corresponding to the estimation sample, but in its bootstrap replica). In fact, on average, 2 tickets are classified as renounces (Nf), and 4 as changes (Ch), while none was classified differently than flown (Fl) through the model estimated without δD.

But is this improvement enough to help the TMC choosing the optimal fare, in order to minimize the cost of flights?

5.3 Economic evaluation of the forecasting performances

Since the prediction of air tickets' outcomes is aimed at deciding which fare to buy ($F*$), in order to minimize the cost of flights, the problem addressed in the present work is economic in nature. Therefore, the wanted predictor is not the one yielding the highest values of statistical measures of forecasting capability, but that producing the biggest saving. As a consequence, the forecasting performance of estimated models should be assessed non just through statistical measures, but also and more relevantly through the economic result, obtainable thanks to the predictors.

In general, the economic performance of a predictor should be evaluated based on

$$C_i - C_{T*,i}(E[Y_i]) \tag{5.9}$$

But:

- for choosing $F*_i$ it would be necessary to know the price of every possible T_i;

- to simulate the price of the 3 X N T_i, a reference price, with constant average difference from the price of all $T, \forall i$ (for flights with different characteristics), is required;

- as reference price, those of the full fare and the IATA fare (reference fare, as calculated by the International Air Transport Association) are available, but none of them has that property;

- the type of fare is recorded only in 108 cases over 29,252, in the forecasting sample;

- moreover, in the forecasting sample, in 19,974 cases over 29,252 C_i is not computable, as the price of each ticket and the cost of each change, is missing.

As a consequence, a simplification and an approximation are required, for developing a cost function useful to the aim of assessing predictors' performances and also business-specific.

First, the price of each ticket Pr_i is approximated with the average cost per Km of the corresponding fare, times the kilometers covered by the i-th flight ($\hat{Pr_i}$).
Afterward, the following hypotheses are made:

1. The possible fares are only two: fixed and flexible. This assumption is definitely meaningful, as, if the business traveler flies the first issued

ticket, than the optimal fare is the cheapest one and there is no cheaper fare than the fixed one. If the ticket is changed, the optimal fare is the one charging the lowest penalty and the flexible fare allows as many changes as the flyer wants, all for free. But also if the traveler renounces to the flight the flexible fare is oprimal, as it is the only full refundable one.

2. In the absence of a reliable predictor, the benchmark purchasing strategy consists in buying always the fixed fare. It is based on the consideration that 85% of the first issued tickets are flown, in the estimation sample, but it is also the strategy commonly prescribed by corporate travel departments, in the corporate travel policy of big corporations.

3. If the traveler changes the ticket, the change is made close to the departure date, when discounted fares tend to be no more available. Moreover, the flyer is assumed to guess that maybe he will have to change more than once, or, finally, could even renounce to the flight, so that he will find more convenient to get a flexible fare.

4. As a consequence, once purchased a fixed fare, in case of change, a flexible ticket is chosen.

5. Prices increase as the date of departure approaches. This is often true in reality, even if the price depends on the plane filling rate. However, this means that a flexible fare, purchased at the time of the first issuance, has a lower price than a flexible fare bought closer to the departure dare.

Accordingly, the cost functions, for the benchmark strategy and for predictors (M) respectively, are:

Table 5.3: Cost function

Obs.	Benchmark	$E_M[y_i] = Ch$	$E_M[y_i] = Nf$	$E_M[y_i] = Fl$
Ch	$Pr_{fix} + Pr_{flex,t+k}$	$Pr_{flex,t}$	$Pr_{flex,t}$	$Pr_{fix} + Pr_{flex,t+k}$
Nf	Pr_{fix}	0	0	Pr_{fix}
Fl	Pr_{fix}	$Pr_{flex,t}$	$Pr_{flex,t}$	Pr_{fix}

where t is the date of issuance and $t + k$ is the date of change.

Then, the loss function for predictors is given by the difference between their cost function and that of the benchmark:

Table 5.4: Loss function

Obs.	$E_M[y_i] = Ch$	$E_M[y_i] = Nf$	$E_M[y_i] = Fl$
Ch	$Pr_{fix} + Pr_{flex,t+k} - Pr_{flex,t}$	$Pr_{fix} + Pr_{flex,t+k} - Pr_{flex,t}$	0
Nf	Pr_{fix}	Pr_{fix}	0
Fl	$-[Pr_{flex,t} - Pr_{fix}]$	$-[Pr_{flex,t} - Pr_{fix}]$	0

Thus, if the $m-$th model predicts that the first issued ticket will be flown, a fixed fare is bought and in case it is changed, the cost of the flight is given by the sum of the price of the fixed fare and the price of the flexible fare, bought closer to the departure date; in both cases that the traveler renounces to the flight and that he flies the first ticket, the cost equals the price of the fixed fare. In this case the cost function for the benchmark is identical to that of the predictor, so that the loss function in zero each times the model predicts Fl.

Whether the model predicts that the first issued ticket will be changed, whether it forecasts that the traveler will renounce to the flight, a fixed fare is purchased. Thus, the cost of the flight is always equal to the price of the flexible fare bought at the time of the first issuance t, except in case the traveler renounces to the flight. In that case, the price of the ticket is completely refundend, so that the non-flown flight costs nothing. Therefore, if the predictor forecasts change or renounce and the traveler changes the ticket, the gain (saving) equals the fixed price plus the (more expensive) flexible price (for the change), minus the lower price of the flexible fare bought time before. In case the traveler renounces to the flight, then the gain is equal to the fixed price.

Only whether the first issued ticket is flown and the model forecasts one of the other two outcomes, it yields a loss, equal to the difference in price between a fixed and a flexible fare (in red in the above table).

The value of such loss function for the guess$-$based predictor is $-1,372,852$, thus the benchmark strategy greatly outperforms it. Clearly, the misclassification rate for flown tickets is very high, because, in order to correctly classify at least a few tickets as changed and renounced, the vacation effect is so heavy, that leads to classify as non-flown also a lot of actually flown tickets. This evidence points out that, even if there seems to be a vacation effect, increasing the probability that a ticket is changed or non flown, the probability that it is flown is still higher, also for flights booked in August or

December and departing in the same months.

Therefore, another alternative solution, to try to increase the prediction capability of models estimated with the poor available information, must be searched.

5.4 A new classification algorithm

For trying to improve the economic performance of predictors, it seems appropriate to replace the ML principle with the economic one, because the aim of the present work is minimizing the cost of flights.

According to the economic principle, the classification rule should be:

$$if \hat{P}_i[y_i = y_j] > h* \Rightarrow E_{h*}[y_i] = y_j \tag{5.10}$$

where

$$h* : \sum_i \{E_{y,h*}[C_{F*_i,i}] - C_i\} = MIN \left\{ \sum_i \hat{E}_{yh}[C_{F*_i,i}] - C_i; h \in [0,1] \right\} \tag{5.11}$$

But, as mentioned above, to claculate the cost of a flight corresponding to its fare $C_{F,i}$, the price of each ticket Pr_i and its fare typology F_i should be known. The cost function described in the previous section can be employed to approximate the expected value of the flight, for the fare that the model suggests to be optimal $\hat{E}_{yh}[C_{F*,i}]$. But the lack of data about the actual cost of the flight cannot be meaningfully replaced with any approximation, in the present context, because here the aim is to compare the amount spent buying a fare chosen with no predictor with the figure that would have been spent if the proposed models were used. Therefore a further simplification is necessary.

Analyzing to the loss function illustrated above, it is found that:

- The only cases in which a predictor yields a loss, if compared to the benchmark strategy, is when it forecasts that the ticket's outcome will be change or renounce, but the traveler actually flies the first issued ticket.
- Each time a predictor forecasts that the flyer will change or waiver the ticket and one of these two behaviors happens, compared to the benchmark strategy, it produces a gain that is higher than the possible loss:

$$Pr_{fix} + Pr_{flex,t+k} - Pr_{flex,t} > Pr_{fix} > Pr_{flex,t} - Pr_{fix} \tag{5.12}$$

- Nonetheless, it is very difficult to beat the benchmark strategy, because a lot of tickets are classified as not flown, if the hand is forced enought to correctly predict a few changes and renounces.

As a consequence, the aim of a new classification rule should be to obtain the highest possible sensitivity to changes and renounces, while keeping the number of incorrectly classified actually flown tickets as low as possible. Therefore, the threshold $h*$ is chosen maximizing the sensitivity to Ch and Nf (outcomes generating relative gains), once controlled for the misclassification rate of Fl:

$$if \hat{P}_i[y_i = y_j] > h* \Rightarrow E_{h*}[y_i] = y_j \tag{5.13}$$

where h*:

$$\frac{\sum_{i=1}^{N} 1_{E_{h*}[y_i]=Nf=y_i}}{\sum_{i=1}^{N} 1_{y_i=Nf}} = MAX\left\{\frac{\sum_{i=1}^{N} 1_{E_h[y_i]=Nf=y_i}}{\sum_{i=1}^{N} 1_{y_i=Nf}}; h \in [0,1]\right\} \tag{5.14}$$

$\wedge$

$$\frac{\sum_{i=1}^{N} 1_{E_{h*}[y_i]=Ch=y_i}}{\sum_{i=1}^{N} 1_{y_i=Ch}} = MAX\left\{\frac{\sum_{i=1}^{N} 1_{E_h[y_i]=Ch=y_i}}{\sum_{i=1}^{N} 1_{y_i=Ch}}; h \in [0,1]\right\} \tag{5.15}$$

$\wedge$

$$\frac{\sum_{i=1}^{N} 1_{E_{h*}[y_i]\neq y_i=Fl}}{\sum_{i=1}^{N} 1_{y_i=Fl}} = MIN\left\{\frac{\sum_{i=1}^{N} 1_{E_h[y_i]\neq y_i=Fl}}{\sum_{i=1}^{N} 1_{y_i=Fl}}; h \in [0,1]\right\} \tag{5.16}$$

Many attempts are made to identify such a threshold (on the estimation sample) on the probabilities estimated by all the models, considered for external validation. Unluckly, no value for $h*$ can satisfy the above desiderata, because available predictors are not 'accurate' enough for assigning the highest probability to the actually observed outcome in a sufficient number of cases. This problem is due, of course, to the extremely low discriminatory power of explanatory variables.

Whence the need to exploit the whole matrix of predicted probabilities, to retrieve more relevant information about the outcome of each single ticket, than it is expressed in the row vector estimated for that unit. The basic idea is to consider not only the ordinal relations between probability masses for different outcomes and the same ticket $P_i[y_i] = (P_i[y_i = Fl]; P_i[y_i = Ch]; P_i[y_i = Nf])$, but also those between probability masses for the same outcome and different tickets $P[y_j]$. In fact, it seems likely that, if the models are not completely misleading, the highest values of the vector $P[y_j]$ should be predicted for tickets actually classified as y_j. For example, if the probability that a certain ticket is changed is higher than the probability that any other one is changed, then that ticket should be actually changed. This information, lost using the MAP rule, can be combined with that present in the individual row vector, for each unit, through a function.

Therefore, predicted probabilities are transformed through a function able to, so to say, filter out the marginal mass, amplifying the faint signal, the discriminatory power of regressors. We call this function a; if $j = 1, 2, \ldots, J$ indicate the classes sorted in decreasing order of observation frequency, its general form is as follows:

$$a(\hat{P}[Y])_i = \frac{\hat{P}_i[y_i = y_1]}{Med(\hat{P}[Y = y_1])} - \sum_{j=2}^{J} \frac{\hat{P}_i[y_i = j]}{Med(\hat{P}[Y = j])} \tag{5.17}$$

Dividing each estimated probability value by the median of the vector of probabilities predicted for the corresponding class, emphasizes the effect of the independent variables in the rarest cases, those for which the probability is especially high in the longitudinal dimension of the estimated matrix. The median is chosen because most of (all, in many models) the explanatory variables are categorical, in particular nominal, too, so that the estimated masses are substantially discrete. In general, in case regressors are continuous variables, it is likely that replacing the median with the mean, in the above function, will yield better results.

Subtracting the summation of the masses predicted for the less frequently observed outcomes (observed in the estimation sample), divided by the respective medians, from the same transformation of the probability estimated for the most frequently observed class, allows to identify those individuals (tickets) for which it is nearly sure that they will be actually classifyed as y_1. The functional form of a is such that it is expected to work especially well in cases, like the present one, where y_1 is observed definitely much more often than the other outcomes (85% of the times). However, it is exactly in this kind of situations that such a function is more needed in business environment where the database is affected by the 80/20 rule of sale or similar specificities of the corporate activities.

Then, 2 thresholds $h*_1$ and $h*_{unc}$ are identified on the vector $a = [a_1, \ldots, a_i, \ldots, a_N]$, on the estimation sample, instead than $j - 1$ thresholds on the raw vectors of estimated probabilities. In general:

$$If \quad a_i(\hat{P}[Y]) \geq h*_1 \Rightarrow E_{h*}[y_i] = y_1 \tag{5.18}$$

where:

$$h*_1 : \frac{\sum_i 1_{E_{h*_1}[y_i]=y_i}}{\sum_i 1_{a_i(\hat{P}[Y]) \geq h*_1}} = MAX \left\{ \frac{\sum_i 1_{E_h[y_i]=y_i}}{\sum_i 1_{a_i(\hat{P}[Y]) \geq h_1}} \right\} \tag{5.19}$$

$$if \quad h*_{unc} < a_i(\hat{P}[Y]) < h*_1 \Rightarrow E_{h*}[y_i] =? \Rightarrow E_{h*}[y_i] = y_1 \tag{5.20}$$

Then:

$$if \quad a_i(\hat{P}[Y]) \leq h{*}_{unc} \wedge \frac{\hat{P}_i[y_i = y_2]}{Med(\hat{P}[y = y_2])} = MAX \left\{ \frac{\hat{P}[y = j]}{Med(\hat{P}[Y = j])}; j = 2, \ldots, J \right\} \Rightarrow E_{h*}[y_i] = y_2 \tag{5.21}$$

$$if \quad a_i(\hat{P}[Y]) \leq h{*}_{unc} \wedge \frac{\hat{P}_i[y_i = y_3]}{Med(\hat{P}[y = y_3])} = MAX \left\{ \frac{\hat{P}[y = j]}{Med(\hat{P}[Y = j])}; j = 2, \ldots, J \right\} \Rightarrow E_{h*}[y_i] = y_3 \tag{5.22}$$

... otherwise $E_{h*}[y_i] = y_J$

The first threshold $h{*}_1$ leaves the individuals that will be 'nearly surely' observed as belonging to the first class (in decreasing order of observation frequency) above its value. Its value is chosen maximizing the sensitivity to this class, considering only individuals whose value of a_i is above $h{*}_1$. It also constitutes the upper limit of an 'uncertainty interval', defined by $h{*}_{unc}$ as its lower bound.

Units for which a_i belongs to this uncertainty interval, are characterized by a lack of information sufficient to correctly classify them, even in a small percentage of cases. The discriminatory power of independent variables is essentially null in these cases, so that the uncertainty about their outcome is as wide as if no regressor were available. Therefore, the estimation sample marginal masses are considered, so that the predicion, for these individuals, coincides with the most frequently observed outcome. This choice is especially appropriate in the present case, because whenever a ticket is forecasted as flown, the purchase of a fixed fare is suggested. And the fixed fare is the one always chosen through the benchmark strategy, so that the relative loss, for the model, is null, as well as the relative gain (saving).

Units for which the value of a_i is lower than $h*_{unc}$ display $\frac{\hat{P}_i[y_i=y_{j\neq 1}]}{Med(\hat{P}[y=y_{j\neq 1}])}$ large enough to be classified with an acceptable degree of reliability. Thus, the values of the transformations for the remaining $J-1$ classes are compared and the highest one identifies the predicted outcome.

In the present case, the algorithm is specifically as follows:

$$a_i(\hat{P}[Y]) = \frac{\hat{P}_i[y_i = Fl]}{Med(\hat{P}[y = Fl])} - \left\{ \frac{\hat{P}_i[y_i = Ch]}{Med(\hat{P}[y = Ch])} + \frac{\hat{P}_i[y_i = Nf]}{Med(\hat{P}[Y = Nf])} \right\} \tag{5.23}$$

Then, the threshold $h*_{fl}$, for the most frequently observed class, is determined on the estimation sample:

$$if \quad a_i(\hat{P}[Y]) \geq h*_{fl} \Rightarrow E_{h*}[y_i] = Fl \tag{5.24}$$

where

$$h*_{Fl} : \frac{\sum_i 1_{E_{h*}[y_i]=y_i=Fl}}{\sum_i 1_{a_i(\hat{P}[Y])\geq h*_{fl}}} = MAX \left\{ \frac{\sum_i 1_{E_h[y_i]=y_i=Fl}}{\sum_i 1_{a_i(\hat{P}[Y])\geq h_{fl}}} \right\} \tag{5.25}$$

below which, the uncertainty interval is defined, with lower bond $h*_{unc}$:

$$if \quad h*_{unc} < a_i(\hat{P}[Y]) < h*_{fl} \Rightarrow E_{h*}[y_i] = ? \tag{5.26}$$

because the available information does not allow a conditional classification of the observations. But, in these cases, the marginal probability of Fl (85%) suggests to buy a fixed F, predicting $y_i = Fl$.

Finally, as Nf is the least frequently observed outcome, if

$$a_i(\hat{P}[Y]) \leq h *_{unc} \wedge \frac{\hat{P}_i[y_i = Ch]}{Med(\hat{P}[y = Ch])} > \frac{\hat{P}_i[y_i = Nf]}{Med(\hat{P}[y = Nf])} \tag{5.27}$$

$\Rightarrow E_{h*}[y_i] = Ch$

otherwise $E_{h*}[y_i] = Nf$

This classification algorithm is applied to all the possible predictors estimating at least 14 different conditional probability values for each outcome (most of the regressors are coded as binary variables, so that many models, containing just a few explanatory variables yield only 2 or 4 different mass values for each outcome. It would be meaningless to try to apply this algorithm in these cases).

Results obtained for the first node of the decision tree ($Y1$), applying the relative loss function described in the previous section, are displayed in the following table:

Table 5.5: Results for Y1

Guess-based	Model1	Model2	Model1A
-1,513,775	-148,388	-894,904	109,011

Thanks to this classification algorithm, inspired to the economic principle, the forecasting capability of all the models improves. In particular, when applied to model 1A:

$$logit(Y1_a) = \alpha_{a,1} + \beta a, c10, 1OccCh + \beta a, v11, 1OccTrav + \beta a, RF, 1FreqWai + \beta a, FWC, 1FreqWaiC + \beta_{a,air,1} Airline + \beta_{a,Ab,1} AdvBK + \beta_{a,RT,1} RT + e_{a,1} \tag{5.28}$$

Even if in statistical terms the forecasting performance is still very unsatisfying :

ECR = 37%
Sensitivity to:
Fl = 25% Ch = 55% Nf = 55%
Specificity to:
Fl = 55% Ch = 25% Nf = 37%

it is globally better than before the application of this algoritm. However, what matters the most in the present work is that an estimated global gain (saving) of 109,011 euros is obtained, compared to the benchmark purchasing strategy. It equals about 5% of the global cost incurred by the company in the considered 6 months.

Therefore, model 1A, the one including variables describing the past behavior of the traveler (that appears to be crucial for predicting his future behavior), is selected as the best predictor for $Y1$. Moreover, now it is possible to descend throughout the decision tree, conditioning to the class forecasted for the first node, predicting subsequent outcomes.

For binary variables, the algorithm becomes simpler, as $P[y = y_2] = 1 - P[y = y_1]$, thus :

$$a(\hat{P}[Y]) = \frac{P[y = y_1]}{Med\{P[Y = y_1]\}} \tag{5.29}$$

And it is sufficient to identify only one threshold $h*$ (one for each binary node):

$$if \;\; a_i(\hat{P}[Y]) \geq h* \Rightarrow E_{h*}[y_i] = y_1 \tag{5.30}$$

otherwise $E_{h*}[y_i] = y_2$

For the sake of completeness, it is worth underlining that forecasts obtained through model averaging are even more biased than those yielded by single predictors. In fact, averaging models with identical ECR, all predicting always the most frequently observed class leads, obviously, to the same ECR and the same predictions. While averaging models with ECR less than that of a completely random classification produces an intermediate result, which is worse than that of the 'best' predictor.

Moreover, it makes little sense to apply the proposed algorithm to probability matrices obtained by model averaging, because, by construction, the averaging procedure filters out the extreme values of probability, extreme with reference to the whole set of masses estimated by all the averaged models, exactly the information that makes the difference between a 'good' and a poor model, once amplified by the algorithm, and allows to get some improvements of the forecasting capability.

5.5 Selecting 'primus inter pares'

In order to classify the subsequent nodes of the business travelers' decision tree, it is necessary to find out a model for each node, that reliably predicts the corresponding variable. Using the traditional MAP rule, with the estimated models, produces the same problems highlighted above, for $Y1$.

Thus, the proposed algorithm is applied again, on all of the models yielding more than 14 different values of predicted probabilities, for each outcome of the node. But a problem for choosing the best predictors emerges, because for nodes 2 (discriminating between total and partial change, given that the first issued ticket is changed) and 4 (classifying tickets in totally changed and flown or not flown, given that the firsst issued ticket is totally changed) all the predictors, to which the algorithm can be applied, show the same forecasting performance, as assessed through the traditional statistical measures. Though they estimate different conditional probabilities.

Of course, it would not be appropriate to pick up a model randomly, because, if one day the TMC will be able to collect a sufficient number of data about the tickets' prices and fares, it would be able to calculate the expected cost of flights and implement the optimal solution to the problem of purchasing the most convenient fare. As shown in equation (1), the reliability of the expected value of such a cost depends on the accuracy of predicted probabilities for the business traveler's behavior.

Therefore it matters not only which output is predicted by the models, but also the reliability of estimated probabilities. On this side of the problem, clearly there is nothing that the developed classification algorithm can do. Moreover, it is no more possible to employ the loss function described above, as there is no other optimal fare for outcomes following the first one. In fact a fixed fare is not changeable nor refundable and the flexible fare can be freely changed, both totally and partially, as many times as the traveler wants.

Therefore, an alternative measure of forecasting performance is developed, considering the J vectors N X 1 $\hat{P}_m[Y_j]$ of estimated probabilities, instead of the predicted classes, as it is the case for extant methods.

- $\forall j, i$ identify $\bar{m}_{i,j+} : \hat{P}_{i,\bar{m}_{j+}}[y_i = y_{i,obs}] = MAX\left\{\hat{P}_{i,m}[y_i = y_{i,obs}], m = 1, \ldots, M\right\}$ the model assigning to the actually observed class the highest mass than any other model.
- $\forall m$ construct J indicators $I_{m,j+}$, with elements $i_{i,m,j+} = 1$ it the $m-$th model $= \bar{m}_{i,j+}$, 0 otherwise.
- $\forall m, j$, compute $S_{m,j+} = \frac{\sum\limits_{j} i_{i,m,j+}}{N}$ as the nuber of missing values can vary for different explanatory variables of different models. Thus, $S_{m,j+}$ is the proportion of cases for which the m-th model is the most accurate one, because it estimates the highest probability for the actually observed outcome, not among the masses for the same individual and different outcomes, but for the same outcomes and each individual, compared to all the other models.

Since both sensitivity and specificity to each outcome is considered throughout the present work, for retrieving deeper insights, especially about the prediction problems, their causes and how to outflank them, the same is done for the negative observations.

- $\forall j, i$ identify $\bar{m}_{i,j-}$: $\hat{P}_{i,\bar{m}_{i,j-}}[y_i \neq y_{i,obs}] = MIN\left\{\hat{P}_{i,m}[y_i \neq y_{i,obs}], m = 1,\ldots,M\right\}$, the model assigning to any class other than the actually observed one the lowest mass than any other model.
- $\forall m$ construct J indicators $I_{m,j-}$, with N elements $i_{i,m,j-} = 1$ if the $m-$th model $= \bar{m}_{j-}$, 0 otherwise.
- $\forall m, j$, compute $S_{m,j-} = \frac{\sum_j i_{i,m,j-}}{N}$. Thus, $S_{m,j-}$ is the proportion of cases for which the m-th model is the most accurate one, in the meaning that it estimates the lowest probability for any outcome other than the actually observed one, not among the masses for the same individual and different outcomes, but for the same outcomes and each individual, compared to all the other models.

According to this descriptive test, the best predictor is:

$$m* : \sum_{m*} S_{m*,j+} + S_{m*,j-} = MAX\left\{\sum_m S_{m,j+} + S_{m,j-}; \forall m\right\} \tag{5.31}$$

the one maximizing the sum of the two proportions, calculated as illustrated above. To simplify, $\sum_m S_{m,j+} + S_{m,j-}$ is named Predictive Accuracy Score (PAS, which, for completeness was also reported in the tables of results for guess-based predictions).

Starting from $S_{m,j}$ it is also possible to inferentially verify the significance of the difference in forecasting accuracy pp_m of the models, through a Chi-square test of identity in proportions:

$$H_0 : pp_{m*} = pp_{m2} \Leftrightarrow S_{m*,j} = S_{m2,j};$$

$$H_1 : pp_{m*} \neq pp_{m2} \Leftrightarrow S_{m*,j} \neq S_{m2,j}$$

$$\chi^2 stat_m = \sum_j \frac{[S_{exp,m,j} - S_{obs,m,j}]^2}{S_{exp,m,j}} \sim \chi^2((J-1)*(n_m-1)) \tag{5.32}$$

If $\chi^2 stat_m > \chi^2 crit_\alpha \Rightarrow$, H_0 is to be refused, otherwise accepted.

Therefore, models for nodes 2 and 4 are compared through PAS and the Chi-square test. The results of the comparison are reported below. Although all the $S_{m,j}$ for all the models are simultaneously considered in the inferential test, to save space only the 2 best PAS are displayed.

Results - first 2 predictors for Y2 | Y1

$$ModelI : logit(Y2 \mid Y1) = \alpha_I + \beta_{I,adv} AdvBk + e_I \tag{5.33}$$

$$ModelII : logit(Y2 \mid Y1) = \alpha_{II} + \beta_{II,adv} AdvBk + \beta_{II,TB} TargetB + e_{II} \tag{5.34}$$

$$PAS_I = 0.94$$

$$PAS_{II} = 0.7$$

$\chi^2 stat = 945$; df= 3 (4 predictors compared) ; $\chi^2 crit_{0.005} = 12.838$

According to the PAS and Chi-squared values reported above, predictor I is significantly more accurate than predictor II and all the other ones. Moreover, it appears that variable Advance Booking includes all the sufficient information for the classification, while variable Target Buy adds only noise. Nonetheless, as only 2 partial changes are correctly classified, it is not possible to predict outcomes following the first partial change.

Results for Y4 | Y2

$$ModelI : logit(Y4 \mid Y2) = \alpha_I + \beta_{I,adv} AdvBk + e_I \tag{5.35}$$

$$ModelII : logit(Y4 \mid Y2) = \alpha_{II} + \beta_{II,adv} AdvBk + \beta_{II,TB} TargetB + e_{II} \tag{5.36}$$

$$PAS_I = 0.65$$

$$PAS_{II} = 1.35$$

$\chi^2 stat = 15.63$; df= 2 (3 predictor compared) ; $\chi^2 crit_{0.005} = 10.597$

Thus, based on the values reported above, predictor II is significantly more accurate and, this time, the variable Target Buy adds useful information for the classification.

Thanks to this alternative measure of predictive accuracy, just one node is left to be forecasted.

Predicting the last node (the leaves of which are partial refund, total refund and no refund, given that the traveler renounces to the flight) is much more difficult. In fact, the relative frequencies of the three outcomes are extremely different in the estimation sample and in the forecasting sample. It seems like if the two samples were drawn from two completely different distributions.

In particular, in the estimation sample the great majority of tickets are partially refunded, while out-of-sample no ticket is partially refunded. This weird fact prevents the proposed classification algorithm from yielding good results, by construction. Nonetheless, once again the predictive performance of all the models improves thanks to the application of the algorithm.

The selected predictor is:

$$logit(Y6 \mid Y1) = \alpha_6 + \beta_{TB6}TargetBuy + \beta_{f6}First + \beta_{COR6}Corporate + \beta_{PA6}PA + \beta_{Nt6}Nat + \beta_{wkr6}WkndRit + e_6 \tag{5.37}$$

Results are reported in the next section.

5.6 Results

In general, forecasts can be either static or dynamic (see: Guizzardi, 2002). When the information set, conditioning to which the prediction is made, includes only observed data, the forecast is static. While the prediction is dynamic if the conditioning information set is composed (also) by previously predicted values of the variables.

In case the dependent variable is modeled as a stochastic function of only the explanatory variables, the prediction can be dynamic whether the independent variables are time-varying. For example, if:

$$Y_{i,t} = f(X_{i,t}) + \epsilon i, t \tag{5.38}$$

$$\hat{Y}_{i,t+k} = f(X_{i,t+k}) \quad is\ a\ static\ forecast; \tag{5.39}$$

$$\hat{Y}_{i,t+k} = f(\hat{X}_{i,t+k}) \quad is\ a\ dynamic\ forecast. \tag{5.40}$$

Where one or more or all the independent variables can be time-varying. So, in the static framework it is necessary to wait for the new (referring to time $t + k > t$) realization of the time-varying explanatory variable(s) to be observed, while in the dynamic approach its value is forecasted through a function of other variables.

In case the dependent variable is modeled as a function of (also) its own past values, the static forecast is made conditioning to the observed values of previous realizations of the dependent variable, while the dynamic prediction is obtaned conditionng to the values previously predicted (whether the forecasting horizon is longer than 1 period). For example, in the simple autoregressive model of order 2:

$$Y_{i,t} = \alpha + \beta_1 Y_{i,t-1} + \beta_2 Y_{i,t-2} + \epsilon i, t \tag{5.41}$$

$$\hat{Y}_{i,t+3} = \hat{\alpha} + \hat{\beta}_1 Y_{i,t+2} + \hat{\beta}_2 Y_{i,t+1} \quad \textit{is a static forecast}; \tag{5.42}$$

$$\hat{Y}_{i,t+3} = \hat{\alpha} + \hat{\beta}_1 \hat{Y}_{i,t+2} + \hat{\beta}_2 \hat{Y}_{i,t+1} \quad \textit{is a dynamic forecast}. \tag{5.43}$$

Of course, the forecast can be made conditioning to both time-varying independent variables and past realizations of the dependent variable itself (or also, as in Vector autoregressive models, of other dependent variables). The dynamic prediction is affected by additional uncertainty, deriving from the forecast of the other values, to which it is conditioned. While the static forecast is burdened only by the uncertainty related to the estimate of the parameters for the dependent variable. Therefore normally static forecasts are more accurate, but dynamic predictions are more timely.

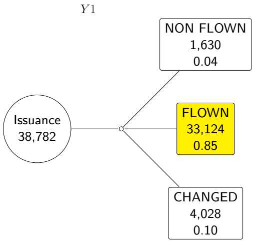
In the present case, no time-varying independent variable is included in the information set, conditionally to which the prediction of the ticket's outcome is made. Clearly, the panel structure of the data is not modeled, because it is too unbalanced and it is not sure that observations for which the cross-sectional unit identification is missing do not refer to the unit identified in other cases. Thus, no time structure is considered and no autoregressive nor dynaic model is employed.

Nonetheless, the phenomenon investigated in the present work is modeled as a decision tree, where each node Y_k is conditioned to its parent one Y_{k-1}. Thus, forecasts too are conditional to the parent node. Although for node 2 (partial or total change, given that the first issued ticket is changed) the parent node is not foregoing in a chronological meaning, but in a logical sense, in all the other nodes it is. Therefore it is meaningful to consider both static and dinamic predictions:

$$\hat{Y}_{i,k} = E_{h*}[Y_{i,k} \mid X_i, Y_{i,k-1}] \quad \textit{is a static forecast}; \tag{5.44}$$

$$\hat{Y}_{i,k} = E_{h*}[Y_{i,k} \mid X_i, \hat{Y}_{i,k-1}] \quad \textit{is a dynamic forecast}. \tag{5.45}$$

that, of course, for the first node ($Y1$) coincide. Results for static predictions follow:



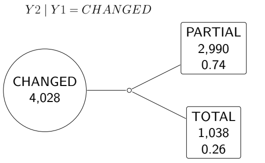
ECR $= 37\%$

Sensitivity to:

Fl $= 25\%$ Ch $= 55\%$ Nf $= 55\%$

Specificity to:

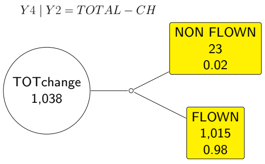
Fl $= 55\%$ Ch $= 25\%$ Nf $= 37\%$



ECR = 92%

Sensitivity to TOTch = 100%

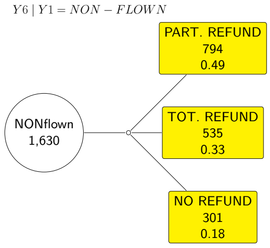
Specificity to TOTch = 2%



ECR = 91%

Sensitivity to TOTch&Fl = 100%

Specificity to TOTch&Fl =0

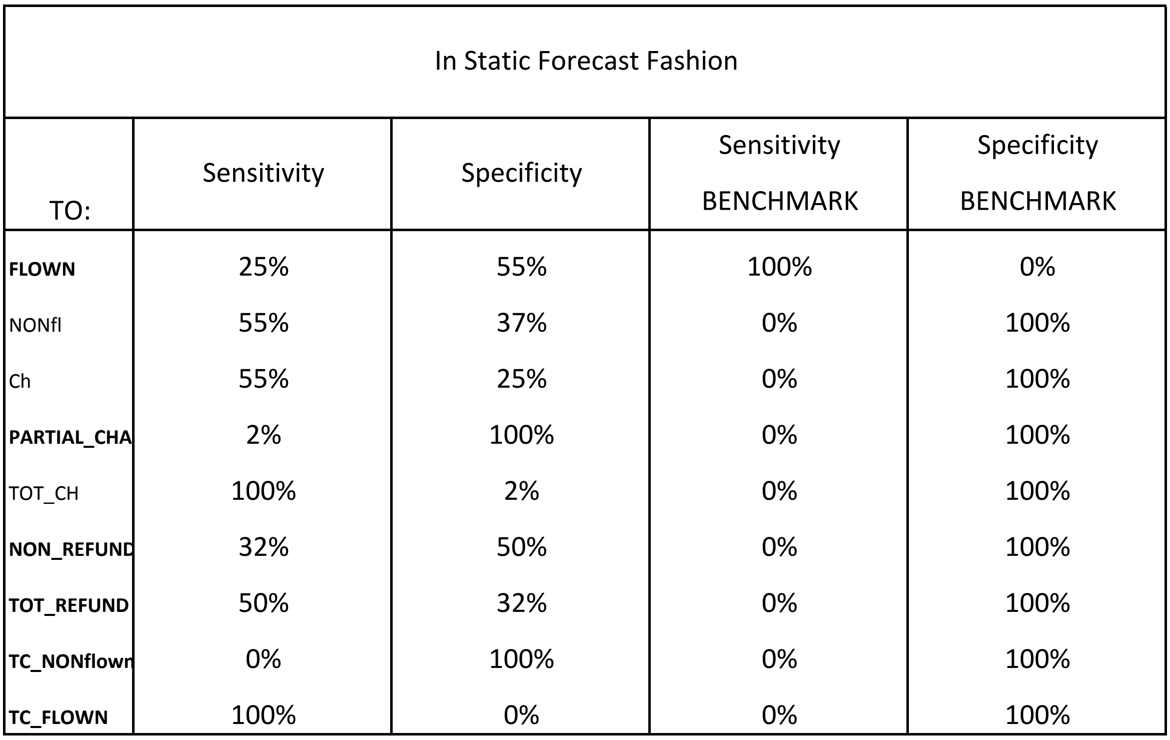


ECR = 35%
Sensitivity to:
TotRef = 50% NonRef = 32%
PartRef = not observed
Specificity to:
TotRef = 32% NonRef = 50%
PartRef = 35%

Because of the low discriminatory power of explanatory variables, even after applying the developed classification algorithm, the ECR is very low, although slightly higher than that of a completely random classifier (equal to 33% for variables having 3 levels and 50% for those with 2 classes), except in case all (or the great majority) of the out−of−sample observations are classified in only one class. In fact, the latter situation is encountered when the observed outcomes are actually nearly degenerate (all belonging to the same class).

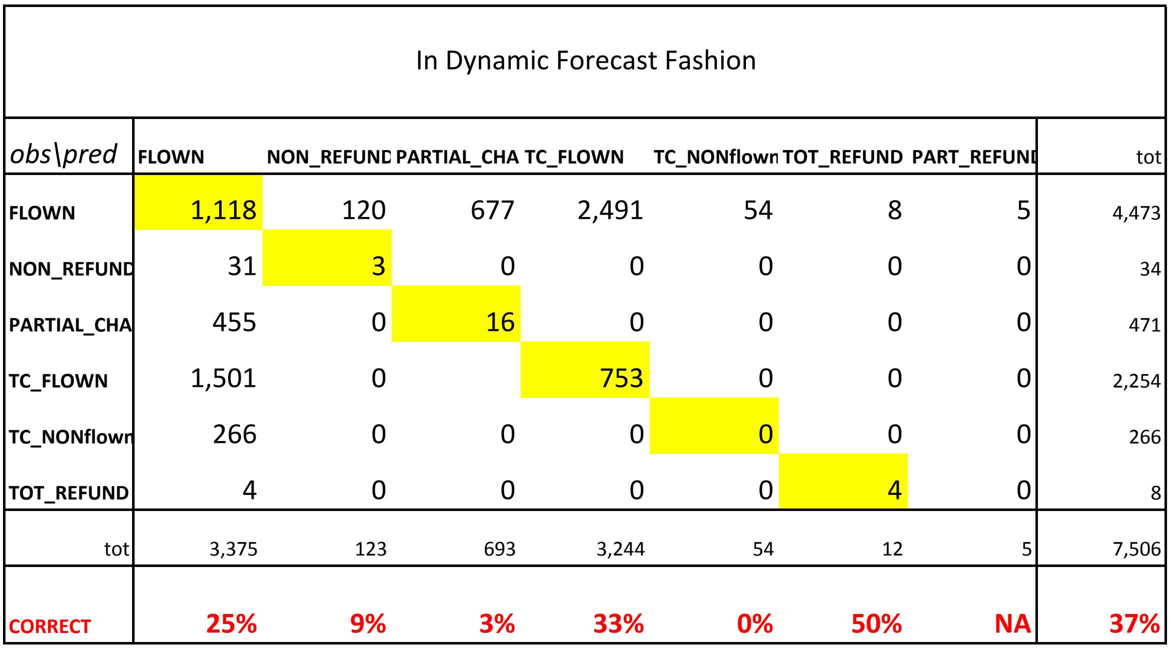
As the first table below highlights, even if the reported statistical measures of forecasting performance are not exciting, they are absolutely better than those obtainable withouth the 'last shore solutions' developed in the previous sections. The latter is named 'Benchmark'.

Table 5.6: Results: Dynamic forecasts



In Static Forecast Fashion

TO:	Sensitivity	Specificity	Sensitivity BENCHMARK	Specificity BENCHMARK
FLOWN	25%	55%	100%	0%
NONfl	55%	37%	0%	100%
Ch	55%	25%	0%	100%
PARTIAL_CHA	2%	100%	0%	100%
TOT_CH	100%	2%	0%	100%
NON_REFUND	32%	50%	0%	100%
TOT_REFUND	50%	32%	0%	100%
TC_NONflown	0%	100%	0%	100%
TC_FLOWN	100%	0%	0%	100%



In Dynamic Forecast Fashion

obs\pred	**FLOWN**	**NON_REFUND**	**PARTIAL_CHA**	**TC_FLOWN**	**TC_NONflown**	**TOT_REFUND**	**PART_REFUNI**	tot
FLOWN	1,118	120	677	2,491	54	8	5	4,473
NON_REFUND	31	3	0	0	0	0	0	34
PARTIAL_CHA	455	0	16	0	0	0	0	471
TC_FLOWN	1,501	0		753	0	0	0	2,254
TC_NONflown	266	0	0	0	0	0	0	266
TOT_REFUND	4	0	0	0	0	4	0	8
tot	3,375	123	693	3,244	54	12	5	7,506
CORRECT	**25%**	**9%**	**3%**	**33%**	**0%**	**50%**	**NA**	**37%**

The final leaves of the decision tree are written in bold font, in contrast to the intermediate nodes. The second table sumarizes the final results (in statistical terms) for the dynamic forecasts. The number of correctly classified outcomes, for each leaf, is colored in yellow. The red percentages are the values of sensitivity to each class. Only the best performing predictors are considered.

Predicted outcomes are listed by columns and observed ones by row. There is one extra column (Partial Refund) because this class is never observed in the forecasting sample, although it is forecasted quite often, as it was the most frequently observed one in the estimation sample.

The same considerations expressed for the performance of the static predictions hold also for the dynamic ones: these values should be read taking into account the poor information available, the complexity of the phenomenon to be modeled and the improvement in forecasting accuracy, relatively to the original one, rather than in absolute terms.

However, the result that really matters for the present work is the economic one, which is quite satisfying: choosign which fare to buy based on the predictions of Model 1A, after the proposed algorithm is applied, allows a global saving of 109,011 euros, that represents approximately the 5% of the total cost of flights.

Conclusions

This thesis was commissioned by a Travel Management Company (TMC), aiming at minimizing the overall cost of business flights it intermediates. The cost of a flight depends on various factors: the class on board, the route, the time of booking, the airline, that are chosen by the client company, but also by the eventual changes or renounce to the journey, made by the business traveler, and by the fare: the only lever the TMC can operate. Thus, the objective of this work was to provide a statistical tool able to help the TMC choosing the optimal fare for each ticket, the one which, coeteris paribus, allows to make flights as cheap as possible.

Airlines offer fares with different levels of flexibility. The higher the price of a fare, the lower the cost of eventual changes and waivers. Thus, the optimal fare depends on the business traveler's behavior, as he can change the ticket, either partially or totally, fly the first issued ticket or renounce to the journey, eventually asking for a partial or total refund. How the flyer will behave is highly uncertain, because in business environments travelers tend to undergo to events (changed meetings' dates, epidemics or wars at the destination, mechanical emergencies and so forth), as a qualitative study of the context highlighted. As a consequence, the natural statistical approach to this problem is deciding to buy the fare minimizing the expected value of the cost of each flight.

To compute the expected value of the cost of a flight it is necessary to know the price of the ticket, the amount of the penalty for changes and the fare; and to estimate the probability of each addend. Unfortunately, the mentioned data are missing in too many cases, in the available corporate dataset. Nonetheless, it is still possible to provide helpful indications, through the probability distribution of tickets' outcomes (consequent to the business traveler's behavior). So the specific problem, addressed in the present work, was to try to predict if a traveler will fly the first issued ticket or opt for another alternative, of those listed above.

Analyzing the correlation structure between available data and performing a qualitative study on a panel of frequent business flyers, it emerged that the tickets' outcomes derive from the decision-making process of the traveler, rather than from the characteristics of tickets and flights. This evidence complicated the task, as most of the available data refer exactly to such characteristics, instead than to those of the traveler, its company and alternative-specific variables, which would have been very useful. However, the phenomenon was modeled as a decision tree. It allowed to identify the models and to simplify the problem, through relations of independence from irrelevant alternatives, conditional to the parent node, at the cost of a progressive loss of efficiency, descending throughout the tree. For each node a variable is specified and modeled through a Pure Multinomial Logit Model.

Due to the scarce correlation between the available variables and the traveler's behavior, the goodness of fit of estimated models is not very satisfying, but the prediction performances are definitely worse. In fact, the Exact Classification Ratio (ECR) of the models is either less than that of a completely random guess, or very high, but for models predicting always the same (most often observed) outcome. This is the consequence of the poor discriminatory power of independent variables, due to the fact that the available information is poor, especially in quality.

In fact, to predict the travelers' behavior, alternative-specific and subject related variables would be needed, but were unavailable or insufficient in quantity, because the database was collected for purposes different from that of the investigation. Moreover, it was not possible to specify hierarchical models, because the sample is too unbalanced and many observations' identifiers are missing and it is not sure that some of them do not belong to the identified ones' groups. Furthermore, due to the 80/20 rule of sales, most of the tickets are purchased for a few client companies, all with the same characteristics, so same values of the explanatory variables, and, among many tickets, just very few are not flown after the first issuance. Besides, given that the dataset covers only one year, divided in estimation and forecasting sample, it was not possible to model time components.

In order to try to get useful predictors, even with the scarce relevant information available, some alternative solutions were proposed. A first hint was searched in the literature and in the qualitative study of the phenomenon. They suggested that business travel follows a sort of seasonal pattern, tied to vacation periods. Thus, an attempt was made to insert a guess, for a vacation effect, directly in estimated models, because the guess was not precise enough to exploit it within a Bayesian framework. Through bootstrap, complete data were generated from the model augmented with the guess, to approximate the vacation effect as would have been estimated on samples actually affected by it. Then the whole model was re-estimated on each bootstrapped dataset and the averaged coefficients used for predicting new bootstrapped data and the real forecasting sample.

Adding the guess about the vacation effect, all the models' predicting capability improved and it became possible to correctly classify some of the two most economically relevant outcomes. However statistical measures of forecasting performance do not matter as much as the economic result of decisions made as a function of predictions, in the present work. Whence the need to develop a business-specific loss function, for evaluating predictive performances, given that the too many missing data about penalties and prices did not allow to estimate the expected value of the cost of flights through the models. The loss function is based on some assumptions simplifying the very messy market of air ticket fares and compares the performance of proposed predictors to that of a benchmark strategy. The latter consists in buying always the fixed fare.

Unfortunately, the value of such a loss function for the guess-based predictor is very unsatisfying. Therefore, a new classification algorithm, amplifying faint signals, exploiting the whole matrix of estimated probabilities, for each prediction, was developed. It is especially useful when a class is much more frequently observed than the most economically relevant ones, but also very flexible, as it can be applied to any matrix of probabilities, estimated by any classifier. Thanks to the proposed algorithm, the predictive performance of all the models improved and, although in statistical terms the performances are still very low, in economic terms an estimated global gain of 109,011 euros is obtained, through the selected predictor. Finally, predicting the two crucial nodes, the algorithm allowed to classify outcomes following the first one.

However, a problem for selecting the best models for nodes two and four emerged. In fact, after the application of the new algorithm, all the candidate predictors displayed identical forecasting performance, as assessed through the traditional measures for nominal data. The latter are based on the predicted outcomes, but do not consider the estimated probabilities of each outcome, that were different, in this case. Conditional probabilities are important, because one day the cost data could be available and the expected value of the cost of flights computed. The reliability of such expectation will be higher the more accurate the estimated masses are. Thus, a further measure of forecasting accuracy, based on the probabilities, called Predictive Accuracy Score is proposed and employed.

The present work has important implications from the business travel management perspective. In particular, it provided an initial and partial (as based on limited information) answer to a current topical question:

in business air travel, does 'best buy' mean always choosing the cheapest ticket? The reply appears to be: most of the time it does, but it is not always the decision that minimizes the cost of flights. Based on the available evidence, in the 85% of cases purchasing the cheapest fare allows to make good saving. But 5% of the times it is more convenient to buy a refundable ticket, otherwise the cost of the ticket is entirely lost, and for the 10% of flights a changeable fare is preferable to the more expensive purchase of a new ticket.

So, how can a corporate travel department know when a flexible fare is worth the higher price? The present work showed that it is possible to identify some characteristics, of the traveler, of the company, of the flight and of the ticket, that influence the probability of ticket's outcomes and reduce the uncertainty about this issue. It is also worth of noting that information about the familiar and personal situation of the travelers would add important clues, along with the identification of the flyer's professional status, that is known and can be modeled by corporate travel departments. Therefore, to the aim of minimizing the cost of flights, personalized corporate travel policies can be more effective than an undifferentiated policy and this research provided some useful methodologies in case of informative problems, that are not rare in corporate datasets.

Concluding, the present thesis made three original contributions to the statistical methodology for predicting nominal data, in presence of poor information. First, a guess-based prediction technique, to exploit external hints (available instead of prior information). Then a new classification algorithm, emphasizing the faint discriminatory power of scarcely correlated explanatory variables. Finally, the Predictive Accuracy Score. As informative problems, of the kind encountered during this work, are not a rarity in business statistics, there is room for possible future developments.

For example, it would be valuable to develop a more specific inferential test for the significance of differences in the Predictive Accuracy

Scores of alternative models. More in general, it would be useful to test the performances of the proposed solutions, and especially of the new classification algorithm, in different fields, for instance in weather forecast, or to predict outstanding financial shocks. Finally, the basic ideas grounding the proposed methods can be exploited in other directions, as they tend to extract as much relevant information as possible from nominal data, on which not many operations and computations can be done, because of their qualitative and non-ordinal nature.

Bibliography

[1] Levere, J. (2000). Akturk , D., Gun, S., Kumuk , T. (2007). Multiple Correspondence Analysis Technique Used in Analyzing the Categorical Data in Social Sciences. *Journal of Applied Sciences.* , 7, 585-588.

[2] Alam, S. S. and Yasin, N. M. (2010). What factors influence online brand trust: evidence from online tickets buyers in Malaysia. *Journal of theoretical and applied electronic commerce research.* 5, 3.

[3] Alamdari, F. (2002). Regional development in airlines and travel agents relationship. *Journal of Air Transport Management.* 8, 339-348.

[4] American Express Travel. *Amex investigations*, retrievable from: https://travel.americanexpress.com/home.

[5] Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes, a large sample study. *Annals of Statistics.* 10, 1100-1120.

[6] Anderson, J. E. and Kraus, M. (1981). Quality of Service and the Demand for Air Travel. *The Review of Economics and Statistics*, 63(4), 533-540.

[7] Arrow, R. D. (1963). *Social Choice and Individual Values.* 2nd Ed. New Haven: Yale University Press.

[8] Baspunar, E. and Mendes, M. (2000). The usage of Correspondence Analysis technique at the contingency tables. *Journal of Agricultural Sciences*. 6, 98-106.

[9] Bartke, P., Gorin, T., Walczak, D., Friedemann, M. (2012). Managing Cancel & Rebook Behavior in Airline Revenue Management. An application of fare adjustment techniques. *Lufthansa Information Management Passage*. Lufthansa editions.

[10] Berger, V. W. and Exner, D. V. (1999). Detecting Selection Bias in Randomized Clinical Trials. *Controlled Clinical Trials*. August, 20, 4, 319-327.

[11] Bitner, M. J. and Booms, B. H. (1982). Trends in Travel and Tourism Marketing : The Changing Structure of Distribution Channels *Journal of Travel Research*. April, 20, 4, 39-44.

[12] Borenstein, S., and Rose, N.L. (1995). Bankruptcy and pricing behavior in us airline markets. *The American Economic Review*. 85, 2, 397-402.

[13] Botimer, T., and Belobaba, P. (1999). Airline pricing and fare product differentiation: A new theoretical framework. *Journal of the Operational Research Society*. 50, 11, 1085-1097.

[14] Brady, S.P., and Cunningham, W.A. (2001). Exploring predatory pricing in the airline industry. *Transportation journal*. 41, 1, 5-15.

[15] Brons, M., Pels, E., Nijkamp, P., and Rietveld, P. (2002). Price elasticities of demand for passenger air travel: a meta-analysis. *Developments in Air Transport Economics*, 8(3), 165-175.

[16] Chen, S. (2002). *Differential pricing on the web: The case of online air travel market. Business Economics and Public Policy*. Kelley School of Business, Indiana University.

[17] Cheng, S. and Long, J. S. (2007). Testing for IIA in the Multinomial Logit Model. *Sociological Methods & Research.* May, 35, 4, 583-600.

[18] Clay, K., Krishnan, R., and Smith, M. (2001). The great experiment: Pricing on the internet. *The handbook of electronic commerce in business and society.* 139-152. Watson Eds, New York, CRC Press.

[19] Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting.* 5, 559-583.

[20] Clemons, E.K., Hann, I.H., abd Hitt, L.M. (2002). Price dispersion and differentiation in online travel: An empirical investigation.*Management Science.* 48, 4, 534-549.

[21] Collins, D. and Tisdell, C. (2002). Gender and Differences in Travel Life Cycles. *Journal of Travel Research.* November, 41, 2, 133-143.

[22] Collins, D. and Tisdell, C. (2000). Travel Life Cycles Vary Significantly with the Purpose of Travel. Department of Economics Discussion Papers No. 276, University of Queensland, Brisbane.

[23] Cox D. R., Miller H. D. (1965). *The Theory of Stochastic Processes.*Chapman and Hall, London.

[24] Dilthey, W. (1883).*Introduction to the Human Sciences.* Princeton University Press.

[25] Duzgunes, O., Kesici, T., Gurbuz, F. (1983). *Statistical Methods I.* Ankara Univeristy, Faculty of Agriculture Publications, 861.

[26] Flom, P. L. and Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use *NESUG Statistics and Data Analysis.* 1-7.

[27] Florens, J-P., FougÃÍre, D., Mouchart, M. (2008). Duration Models and Point Processes. *The Econometrics of Panel Data: Handbook of Theory and Applications.* Eds. Sevestre P. and Lazlo M., 3rd edition, Springer, 547-601.

[28] Fossey, E., Harvey, C., McDermott, F., Davidson, L. (2002). Understanding and evaluating qualitative research. *Australian and New Zealand Journal of Psychiatry.* November, 36, 6, 717-732.

[29] Foster, D. P. and Stine, R. A. (2006). Honest confidence intervals for the error variance in stepwise regression. *Working papers.* The Wharton School of the University of Pennsylvania.

[30] Fourie, C., and Lubbe, B. (2006). Determinants of selection of full-service airlines and low-cost carriers——A note on business travellers in South Africa. *Journal of Air Transport Management*, 12(2), 98-102.

[31] Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician.* 37, 152-155.

[32] Fry, T. R. L. and Harris, M. N. (1996). A Monte Carlo study of tests for the independence of irrelevant alternatives property.*Journal of Transportation Research.* 30, 1, 19-30.

[33] Gaudry, M. J. I. and Dagenais, M. G. (1979). The DOGIT model. *Transportation Research.* 13B, 2, 105-111.

[34] Gordon,S. C. and Smith, A. (2004).Quantitative Leverage Through Qualitative Knowledge: Augmenting the Statistical Analysis of Complex Causes. *Oxford Journals: Political Analysis.* 12, 3, 223-255.

[35] Gorin, T., Walczak, D., Bartke, P., Friedemann, M. (2012). Incorporating cancel and rebook behavior in revenue management optimization. *Journal of Revenue and Pricing Management.* 2, 117-126.

[36] Graneheim, U. H. and Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*. February, 24, 2, 105-112.

[37] Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

[38] Guizzardi, A. (2002). *La previsione economica. Problemi e metodi statistici*. Guaraldi. Rimini.

[39] Harrell, F.E. (2001). *with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag, New York. Miller, A. J. (2002), Subset selection in regre

[40] Hausman, J. A. and McFadden, D. (1984). Specification tests for the multinomial Logit model. *Econometrica*. 52, 1219-1240.

[41] Horowitz, J. (1981) Identification and diagnosis of specification errors in the multinomial Logit model. *Transportation Research Record*. 58, 345-360.

[42] Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. I, 221-33.

[43] Kim, L., Qu, H., Kim, D. J. (2009). A study of perceived risk and risk reduction of purchasing air tickets online. *Journal of Travel and Tourism Marketing*. 26, 3, 203-224.

[44] King, M. L. and Wu, P. X. (1993).*Locally optimal one-sided tests for multiparameter hypotheses*. Monash: Mimeo.

[45] Kulendran, N. and Wilson, K. (2000). Modelling business travel *Tourism Economics*. 6, 1, 47-59.

[46] Kulendran, N. and Witt, S. F. (2003). Forecasting the Demand for International Business Tourism. *Journal of Travel Research.* February , 41, 3, 265-271.

[47] Kruskal, W. H. (1960).Some Remarks on Wild Observations. *Technometrics.* 2,1, 1-3 .

[48] International Air Transport Association. (2000). Airline Economic Results and Prospects, Part 1 *Summary Report IATA.*

[49] International Air Transport Association. (2014). Airline Economic Results and Prospects, Part 1 *Summary Report IATA.*

[50] ISTAT. (2013). *Rapporto Annuale.* Il mercato del lavoro tra minori opportunitÃă e maggiore partecipazione. Retrieved the 09/10/2014 from: http://www.istat.it/it/files/2013/05/cap3.pdf.

[51] L'osservatorio sul Business Travel. (2014). *Turismo d'affari.* Milano. Ediman.

[52] Law, R. and Chang, M. M. S. (2007). Online Pricing Practice of Air Tickets: The Case of Hong Kong. *Information and Communication Technologies in Tourism.* 513-522.

[53] Law, R. and Leung, R. (2000). A Study of Airlines' Online Reservation Services on the Internet. *Journal of Travel Research.* November, 39, 2, 202-211.

[54] Levere, J. (2000). Changing Roles. *Airline Business.* October, 48-76.

[55] Lin, P-C., Chen, C-C., Song, M-H. (2009). Price dispersion of online air tickets for short distance international routes *The Service Industries Journal.* 29, 11, 1597-1613.

[56] Luce, R. D. (1958). *Individual Choice Behavior.* New York: John Wiley.

[57] Lukacs, P. M., Burnham, K. P., Anderson, D. R. (2010). *Model selection bias and FreedmanâĂŹs paradox.* Annals of the Institute of Statistical Mathematics. 62, 117-125.

[58] Mason, K.J. and Gray, R. (1999). Stakeholders in a hybrid market: the example of air business passenger trave *European Journal of Marketing.* 33, 9/10, 844-858.

[59] Malterud, K. (2001). Qualitative research: standards, challenges, and guidelines. *The Lancet.* August, 358, 9280, 11, 483-488.

[60] McAfee, R. P. and Velde, V. (2006). Dynamic Pricing in the Airline Industry. *Handbook on Economics and Information Systems.* 1-43.

[61] McFadden, D., Tye, W. B., Train, K. (1978). An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model. *Transportation Research Record.* January, 637, 39-45.

[62] McFadden, D., Train, K.,Tye, W. B. (1981). An Application of Diagnostic Tests for the Independence From Irrelevant Alternatives Property of the Multinomial Logit Model. *Transportation Research Board Record.* 637, 39-46.

[63] McNichols, M. and O'Brien, P. (1997). Self-Selection and Analyst Coverage. *Journal of Accounting Research.* Studies on Experts and the Application of Expertise in Accounting, Auditing, and Tax. 35, 167-199.

[64] Moler, C. and van Loan, C. (2003). Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review.* 45, 1, 3-49.

[65] Nako, S. M. (1992). Frequent flyer programs and business travellers: An empirical investigation. *Logistics and Transportation Review*, 28(4), 395-402.

[66] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal.* 7, 308-313.

[67] Park, J. Y., and Jang, S. C. S. (2014). Sunk costs and travel cancellation: Focusing on temporal cost. *Tourism Management.* 40, 425−435.

[68] Proussaloglou, K., and Koppelman, F. S. (1999). The choice of air carrier, flight, and fare class. *Journal of Air Transport Management*, 5(4), 193-201.

[69] Radner, R. and Marschak, J. (1954). Note on Some Proposed Decision Criteria. In: R. M. Thrall, C. H. Coombs, R. L. Davies. *Decision Process.* New York, John Wiley.

[70] Ray, P. (1973). Independence of Irrelevant Alternatives. *Econometrica.* September, 41, 5, 987-991.

[71] Ritchie, JR.B. and Beliveau, D. (1974).Hallmark Events: An Evaluation of a Strategic Response to Seasonality in the Travel Market. *Journal of Travel Research.* October, 13, 2, 14-20.

[72] Sauerbrei, W. (1999). The use resampling methods to simplify regression models in medical statistics. *Applied Statistics.*48, 313-329.

[73] Shatland, E.S., Kleinman, K., Cain, E.M. (2008). A new strategy of model building in PRO LOGISTIC with automatic variable selection, validation, shrinkage and model averaging. *Proceeding of the twenty-ninth annual SAS users group international conference.* 1-10.

[74] Levere, J. (1985). Small, K. A. and Hsiao, C. (1985). Multinomial Logit specification tests. *International Economics Review.* 16, 471-486.

[75] Steyerberg, E. W., Eijkemans, M. J. C., Harrell J., F. E., Habbema, J. D. F (2000). Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine.* 19, 1059-1079.

[76] Steyerberg, E. W., Eijkemans, M. J. C., Harrell J., F. E., Habbema, J. D. F. (2001). Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making.* 21, 45 -56.

[77] Swarbrooke, J. and Horner, S. (2001).*Business Travel and Tourism.* Butterworths-Heineman, London.

[78] Train, K. (2003).*Discrete Choice Methods With Simulation.* New York: Cambridge University Press.

[79] Train, K. *Discrete Choice Methods with Simulation.* (2009). Cambridge University Press, second edition.

[80] Levere, J. (2000). Tse, Y. K. (1987) A diagnostic test for the multinomial Logit model. *Journal of Business and Economic Statistics.* 5, 283-286.

[81] Tunstall, R. (1989). Catering for the female business traveller. *Travel and Tourism Analyst.* 5, 26-40.

[82] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Fourth edition. Springer.

[83] *Turismo dâĂŹaffari.* (2014). Milano. Ediman.

[84] Vowles, T.M. (2000). The effect of low fare air carriers on airfares in the US. *Journal of Transport Geography.* 8, 2, 121-128.

[85] Wallenberg, F. (2000). *A study of airline pricing.* Berkeley, CA: University of California.

[86] Wilfried, B. and Tarnai, C. (1999). Content analysis in empirical social research. *International Journal of Educational Research.* January, 31, 8, 659-671.

[87] Wunsch,G., Mouchart, M., Russo, F. (2014). *Des Causes et des Effets.* L'Academie en poche.

[88] Yoon, M. G., Yoon, D. Y., Yang, T. W. (2006). Impact of e-business on air travel markets: Distribution of airline tickets in Korea. *Journal of Air Transport Management.* 12, 5, 253-260.