Accounting for similarities in air transport route choice

Michael VAN EGGEMOND
Institute for Transport Planning and Systems (IVT)
ETH Zurich
Faculty of Civil Engineering & Geosciences, Transport & Planning
Delft University of Technology (TU Delft)
HIL F 31.1.
CH-8093 Zurich
Phone: +41-44-633 67 37
Fax: +41-44-633 10 57
Email: michaelva@student.ethz.ch

Nadine SCHUESSLER
(Corresponding Author)
Institute for Transport Planning and Systems (IVT)
ETH Zurich,
HIL F 51.1
CH-8093 Zurich
Phone: +41-44-633 30 85
Fax: +41-44-633 10 57
Email: schuessler@ivt.baug.ethz.ch

Kay W. AXHAUSEN
Institute for Transport Planning and Systems (IVT)
ETH Zurich,
HIL F 32.2
CH-8093 Zurich
Phone: +41-44-633 39 43
Fax: +41-44-633 10 57
Email: axhausen@ivt.baug.ethz.ch

Words: 5297
Figures: 1 (=250 words)
Tables: 4 (=1000 words)
Total: 6547
ABSTRACT
Overcoming the “independence of irrelevant alternatives” (IIA) property of the basic multinomial logit (MNL) model is a major research issue in the field of discrete choice modelling. In recent years, several approaches have been developed to achieve this goal. This paper focuses on similarities in the context of air connection choice. Due to the extensive choice sets in this study a similarity factor approach is applied, which does - even for this huge data set - not prolong estimation time.

The scrutinized similarity factor has been specifically developed for public transport connections. It is capable to account for the complex and multi-dimensional nature of similarities of public transport alternatives. The optimal parameter setting for the similarity factor with respect to the dataset at hand is found. The estimation results emphasize that the similarity between alternatives exerts a significant influence on air connection choice. Similarity is perceived positive. It appears that decision-makers have a strong preference for certain attributes of the alternatives such as a specific departure time, travel time or fare. Thus, they prefer alternatives that are similar with respect to these attributes.

ACKNOWLEDGMENTS
The authors would like to thank Henrik Imhof and Karl Isler (Swiss International Airlines) for providing the data set containing the ticket booking through CRS system and the webbots to observe the fares. Their contribution was central to the success of this study. In addition, we would like to thank Michel Bierlaire for providing Biogeme for the logit estimations.
INTRODUCTION

Compared to ground-based public and private transportation, the number of studies using discrete choice models to represent passenger behavior in the aviation sector is fairly limited. Potential application areas for discrete choice models are airport choice and air connection choice (itinerary) modeling, or combinations of both. This challenging research area includes complex choices across multiple dimensions. Air connection choice modeling can aid airlines with their medium and long-term planning as it provides carriers with an understanding of the relative importance of different service factors. This understanding can also aid airlines and online travel portals with the listings of different air connections on their websites.

Another important challenge in the field of transport modeling is how to overcome the IIA (independence of irrelevant alternative) property of the classic Multinomial Logit model (MNL). Recent research has focused on three different general approaches: changing the variance-covariance structure, nesting alternatives, and introducing similarity factors in the deterministic part of the utility function. The main issue is to find a solution that is first flexible and able to represent complex correlations, allows second a more thorough understanding of people’s transport behavior and is third easy to compute and applicable to large choice sets.

This paper focuses on a similarity factor for the particular problem of air connection choice. Since similarities between alternatives in this context are complex and have several dimensions (spatial, temporal and monetary), similarity factors used so far in private transport route choice are not sufficient. Instead, the Autonomy of a Connection (AC) concept by (1) will be scrutinized.

Several datasets have been necessary for this research: a dataset that contains tickets bookings through computer reservation systems (CRS) for November 2006, a dataset with fares observed in the period September 2006 – November 2006 for flights departing in November 2006 on 70 origin-destination pairs and the official airline guide database (2).
They have been combined to form a comprehensive dataset for the analysis of air connection choice as well as an appropriate similarity measure.

The remainder of the paper is structured as follows: After a short overview over recent developments in air travel route choice modeling, the problem of accounting for similarities specifically in connection choice will be highlighted. Subsequently, the dataset used in this study will be described, before the utility function of the air connection choice model and the modeling approach for the similarity factor including the parameter settings are presented. Afterwards, the findings regarding the connection choice model in general and the treatment of similarities in particular will be discussed in the results section. Special emphasis will be laid on the optimal formulation of the similarity factor. The paper concludes with a summary of the main findings.

**AIR TRAVEL ROUTE CHOICE MODELING AND THE ACCOUNT FOR SIMILARITIES**

From a passenger point of view, several choice stages can be recognized with regard to air travel. Amongst these choices are the choice to make a trip, destination choice, the choice when and how to book, mode choice, airport choice and itinerary choice. These choices are not necessarily sequential. On the contrary: it is highly probable that multiple feedback loops exist between them and that some are even carried out simultaneously.

To account for this, Hess and Polak consider the simultaneous choice of airport access mode, airport and airline in both the San Fransisco Bay area (3) and the Greater London area (4) using stated preference data.

Several studies have been carried out with regard to air connection modeling, either using stated preference data (5) (6) or revealed preference data (7, 8). The latter studies provide an excellent overview of the research carried out in aviation modeling.
The modeling framework used in these studies as well as in the work presented here is
discrete choice modeling. In the most basic of these models, the Multinomial Logit (MNL)
model, the utility $U_{in}$ of an alternative $i$ for a decision-maker $n$ is defined by

$$U_{in} = V_{in} + \varepsilon_{in} = f(\beta, x_{in}) + \varepsilon_{in}$$

with a deterministic part $V_{in}$ that consists of a function $f(\beta, x_{in})$ of the vector $\beta$ of taste
parameters and the vector $x_{in}$ of attributes of the alternative, the decision-maker and the
choice situation, whereas the error terms $\varepsilon_{in}$ are independently type I extreme values
(Gumbel) distributed. The probability of an alternative to be chosen is then:

$$P_{in} = \frac{e^{\mu_{in}}}{\sum_{j \in C_n} e^{\mu_{jn}}}$$

A major shortcoming of the MNL model is the “independence of irrelevant alternatives”
(IIA) property. Since the error terms are independently Gumbel distributed, no correlations
between the alternatives of a choice set are taken into account, though they often exist in
reality.

In general, three different approaches have been established to overcome the IIA property:

- changing the variance-covariance structure,
- subdividing alternatives into nests, and
- introducing similarity factors in the deterministic part of the utility function.

The first group contains on the one hand the (multinomial) Probit model, applied for
example by (9) to a private transport route choice problem. Multivariate normal distributed
error terms replace the i.i.d. Gumbel distributed ones of the MNL. On the other hand there is
the family of Mixed Multinomial Logit (MMNL) models presented first by (10). They
preserve the i.i.d. Gumbel distributed error terms and add (multivariate) random variables to
account for correlations. Different distributional assumptions for the non-Gumbel variable
have been successfully tested for different choice situations by for instance (11), (12), (13), (14) or (15). The aforementioned study by (3) applied it to an aviation context.

All of these models are able to account for any kind of correlation structure. However, they require a lot of effort in terms of specification, as elaborated in (16), and computation and are thus still hardly applicable to choice situations with large numbers of alternatives.

The second group contains the nested approaches which have been generalized by (17) to the Generalized Nested Logit (GNL) model and (18) to the Network GEV model. Alternatives are subdivided into groups, the so-called nests. Correlations may remain within the nests, but between the nests they are eliminated. In the classic Nested Logit (NL) (19) model the nests are completely disjoint whereas in the Cross Nested Logit (CNL) model (20) each alternative can belong to more than one nest. In the context of aviation choices, (8) applied a nested logit model structure to air-travel itinerary modeling in the United States, whereas (4) used a CNL structure to model airport choices in the Greater London area.

However, though particularly the CNL is able to represent nearly all kinds of correlations, a realistic nesting structure is highly complex and therefore cumbersome to estimate. Only the NL model is applicable to large choice sets as they are present in air travel connection choice.

The models of the third group are less complex and completely preserve the MNL model structure. Thus, they are easier to estimate and applicable to large choice sets. The deterministic part of the utility function is extended by a similarity factor that adapts the utility of an alternative with respect to its similarity with other alternatives in the choice set.

The basis for these approaches is the implicit availability/perception (IAP) model presented by (21). They state that a decision-maker is not able to consider all alternatives of the universal choice set because of the individual’s imperfect knowledge of the alternatives and limited information processing abilities. The underlying assumption attests that the similarity
of an alternative with other alternatives decreases its utility because it decreases its probability to be perceived an alternative. However, recent studies such as (22) or (15) suggest that this assumption does not hold for all choice contexts. In route choice situations for instance, the utility of a route increases with its similarity to other routes because the decision-maker gains the possibility to switch routes while traveling.

The most prominent similarity factors have been developed for route choice situations, where similarities between alternatives are most obvious and of spatial nature. The first one was the C-Logit proposed by (21) following their definition of the IAP model. A so-called commonality-factor decreases the utility with respect to the percentage of route length that the route shares with other routes.

Second, (23) proposed the Path Size Logit (PS) model. The path size measures decreases the utility of an alternative according to the length of its links and the length of the routes that share a link with it. A modification of the PS factor, the General Path Size (GPS) factor was presented by (24). However, (22) as well as (15) criticize this factor as it is behaviorally difficult to interpret and produces counterintuitive results.

Since most of the similarity factors presented above have been developed for private transport route choice, they focus on a spatial dimension of similarity. However, in public transport connection choice the spatial dimension is less decisive and mainly restricted to shared transfer points. For air connection choice even this influence is arguable. Instead, temporal aspects are highly relevant, especially for inter-urban public transport as shown by (22) and (25). While (22) found the use of the same public transport leg to be the significant description for overlaps in multimodal route choice, (25) demonstrated that correlations between departure times are much stronger than those between the same public transport modes. Another important aspect for a public transport similarity measure is the fare.

Thus, (1) designed a similarity measure specifically for public transport connection choice, the Independence of a Connection (IND) factor. The independence of a connection is defined
as the reciprocal of the sum of similarities with other alternatives in the choice set. The similarity is measured considering the time gap between corresponding departure and arrival times and the differences in speed and price. An application of this factor can be found in (26). The actual formulation of the IND factor and its application is discussed below.

DATASET DESCRIPTION

Three revealed preference datasets are available for the purpose of this research:

- tickets bookings through computer reservation systems (CRS) in November 2006
- fares observed in the period September – November 2006 for departures in November 2006 on 70 origin-destination pairs

The first dataset contains detailed records of passenger bookings in November 2006 through Computer Reservation Systems (CRS). CRS systems included in the dataset are Amadeus, Abacus, Galileo, Worldspan and Apollo. A rough comparison with Eurostat figures has led us to believe that the CRS data cover between the 40% and 90% of the passenger bookings on any one route. Variables included in the CRS dataset are: booking date, trip origin, trip destination, leg origin, leg destination, departure date, return date, departure and arrival times, carrier abbreviation, and flight number per leg.

The second dataset was obtained by webbots querying Expedia (http://www.expedia.de) on a nearly daily basis in the period September – November 2006 for flights departing in November 2006 on 70 origin-destination pairs in Europe. Three durations of stay were queried: a trip returning the same day, a trip returning on the next day and a trip returning in two weeks time. Variables obtained from Expedia include query date, trip origin, trip destination, departure date, return date, departure and arrival times, carrier name(s), flight number(s) and most notably fare.
Third, detailed information of carrier schedules was obtained from the Official Airline Guide (OAG, 2006). This dataset contains variables such as operating day, operating airline, code share airlines, departure and arrival time and type of aircraft operated.

In order to use these datasets, two steps had to be taken. First, the datasets were matched to obtain complete air connection information. Second, choice sets were extracted.

Air connection fares were added to the passenger bookings by matching the CRS dataset and Expedia dataset on query date, departure date, duration of stay and outbound and inbound flight number combination. OAG information was added by matching flight numbers and carrier abbreviations, taking into account code shares. In the end, nearly 19,000 choices with fare information are observed. Unfortunately, no characteristics of the decision-makers could be included. Therefore, no trip purposes are known and no fare parameter per user category and parameters for deviation of preferred departure time could be estimated, as done by Garrow, Jones and Parker (2006).

Choice sets were generated by assuming that a traveler considered all alternatives available on Expedia on the day of booking for the booked origin-destination pair and flights departing on the booked departure day.

In Table 1 an overview of the used variables, a brief description and some descriptive statistics are given.
Table 1  Variables of the connection choice model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier</td>
<td>Dummy for operating carrier</td>
<td></td>
</tr>
<tr>
<td>Code share</td>
<td>Dummy for code share on connection</td>
<td>3.3 % 11.4 %</td>
</tr>
<tr>
<td>Total travel time</td>
<td>Average travel time between departure and arrival, including waiting time at transfers</td>
<td>90 158</td>
</tr>
<tr>
<td>Number of transfers</td>
<td>Number of transfers in the air connection.</td>
<td>0.4% 99.6%</td>
</tr>
<tr>
<td>Fare</td>
<td>Fare as listed on Expedia (without taxes)</td>
<td>€ 377.94 € 512.95</td>
</tr>
<tr>
<td>Regional aircraft</td>
<td>Dummy if connection includes a regional jet</td>
<td>42.2% 48.6%</td>
</tr>
<tr>
<td>Propeller aircraft</td>
<td>Dummy if connection incl. propeller aircraft</td>
<td>7.2% 11.5%</td>
</tr>
<tr>
<td>Afternoon flight out - same day return</td>
<td>Dummy - departure time of outbound connection between 9 a.m. and 4 p.m.</td>
<td>6.6% 4.7%</td>
</tr>
<tr>
<td>Evening flight out - same day return</td>
<td>Dummy - departure time of outbound connection after 4 p.m.</td>
<td>1.0% 1.0%</td>
</tr>
<tr>
<td>Afternoon flight out - overnight stay</td>
<td>Dummy - departure time of outbound connection between 9 a.m. and 4 p.m.</td>
<td>33.3% 41.8%</td>
</tr>
<tr>
<td>Evening flight out - overnight stay</td>
<td>Dummy - departure time of outbound connection after 4 p.m.</td>
<td>27.4% 27.9%</td>
</tr>
<tr>
<td>Afternoon flight out - fortnight stay</td>
<td>Dummy - departure time of outbound connection between 9 a.m. and 4 p.m.</td>
<td>41.2% 49.5%</td>
</tr>
<tr>
<td>Evening flight out - fortnight stay</td>
<td>Dummy - departure time of outbound connection after 4 p.m.</td>
<td>34.2% 29.4%</td>
</tr>
<tr>
<td>Afternoon flight in - same day return</td>
<td>Dummy - departure time of inbound connection between 4 p.m. and 6 p.m.</td>
<td>32.7% 41.6%</td>
</tr>
<tr>
<td>Evening flight in - same day return</td>
<td>Dummy - departure time of inbound connection after 6 p.m.</td>
<td>64.7% 57.3%</td>
</tr>
<tr>
<td>Afternoon flight in - overnight stay</td>
<td>Dummy - departure time of inbound connection between 4 p.m. and 6 p.m</td>
<td>32.2% 43.5%</td>
</tr>
<tr>
<td>Evening flight in - overnight stay</td>
<td>Dummy - departure time of inbound connection after 6 p.m</td>
<td>64.0% 51.0%</td>
</tr>
<tr>
<td>Afternoon flight in - fortnight stay</td>
<td>Dummy - departure time of inbound connection between 4 p.m. and 6 p.m</td>
<td>13.9% 15.3%</td>
</tr>
<tr>
<td>Evening flight in - fortnight stay</td>
<td>Dummy - departure time of inbound connection after 6 p.m</td>
<td>4.8% 4.9%</td>
</tr>
</tbody>
</table>
FORMULATION OF THE UTILITY FUNCTION

Based on the information available and the studies discussed above, eight attributes of the alternatives are considered: Journey and waiting time, fare, departure time period, duration of stay, carrier, code share, and type of aircraft. The alternatives are return trips between two airports without access or egress times. The return date was at the same day as the outbound flight, a day afterwards or a fortnight later.

Two variants have been tested to derive the best way to treat the relationship between journey time and transfers of the outbound flight: first, the combination of total journey time and number of transfers and second, the combination of in-vehicle time and waiting time at transfer points.

The fare variable represents the total price a passenger had to pay at his particular booking date without taxes and service charges.

Another interesting aspect was the influence of the departure time period (morning, afternoon, evening) for the outbound and the inbound flight and the duration of stay at the destination (same day return, overnight stay, fortnight stay). First, independent dummy variables for these attributes have been included. However, further analysis showed that variables representing combinations of these attributes had a higher explanatory power. Thus, dummy variables for meaningful combinations of departure time period and duration of stay have been included and each been evaluated against the base case morning flight.

Regarding the perception of carrier attributes, models with dummy variables for carrier attributes such as domestic and non-domestic or low cost carrier, flag carrier and regional carrier, have been tested against models with a dummy variable for each individual carrier. Additionally, in all models a dummy variable was included, that indicated, whether the flight was offered as a code share.

Finally, the type of aircraft considered. The dummy variables regional aircraft and propeller aircraft have been evaluated against the base case mainline jet.
To account for similarities the Independence of a Connection factor by Friedrich, Hofäs and Wekeck (2001) has been applied, which is calculated as for connection \( c \) of choice set \( C \):

\[
IND(c) = \frac{1}{\sum_{c \in C} f_c(c')} + \frac{1}{\sum_{c \in C} f_c(c')} \sum_{c \in C} \sum_{c' \neq c} f_c(c')
\]

Thus, the calculation of the choice probabilities is changed as following:

\[
P_{in} = \frac{e^{V_{in}} \cdot IND(i)}{\sum_{j \in C} e^{V_{in}} \cdot IND(j)} = \frac{e^{V_{in} + ln(IND(i))}}{\sum_{j \in C} e^{V_{in} + ln(IND(j))}}
\]

Thereby, \( f_c \) is an appropriate non-negative evaluation function, with \( f_c(c) = 1 \) and \( f_c(c') \leq 1 \), \( c' \in C \). \( f_c \) models the relative influence of other connections on \( c \) with respect to departure time, perceived journey time and fare and is defined as follows:

\[
f_c(c') = \left( 1 - \frac{x_c(c')}{s_x} \right) \left( 1 - \gamma \cdot \min \left( \frac{y_c(c')}{s_y}, \frac{z_c(c')}{s_z} \right) \right)
\]

With

\[
x_c(c') = \left[ \frac{DEP(c) - DEP(c')}{2} \right] + \left[ \frac{ARR(c) - ARR(c')}{2} \right]
\]

\[
y_c(c') = PJT(c') - PJT(c)
\]

\[
z_c(c') = Fare(c') - Fare(c)
\]

\[
PJT(c) = JT(c) + 2 \cdot WT(c) + 2 \cdot NT(c)
\]

And

\[
s_y = \begin{cases} s_y^+ & y(c') \leq 0 \\ s_y^- & y(c') > 0 \end{cases}
\]

\[
s_z = \begin{cases} s_z^+ & z(c') \leq 0 \\ s_z^- & z(c') > 0 \end{cases}
\]

\( s_x, s_y \) and \( s_z \) set the range of influence of \( x_c(c') \), \( y_c(c') \) and \( z_c(c') \). \( s_y \) and \( s_z \) depend on the sign of \( y_c(c') \) and \( z_c(c') \) in order to model the asymmetry between connections; if there is a difference in terms of perceived journey time, the superior connection will exert a
stronger influence on the inferior one. These parameters have a strong influence on the value of $\text{IND}(c)$. However, the number of studies that use this formula is very limited and there is not much experience about the optimal value of $s_x$, $s_y$, and $s_z$. Therefore, based on the available data an experimental design is set up to derive the optimal values for these parameters in this specific choice context, the results of which are presented in the next paragraph.

**TESTING OF THE PARAMETER SETTING**

Several remarks should be made with regard to the setting of the parameters. First, the right part of the formula has to be smaller than 1 on average in order to influence the independence. Thus, the sum of $\frac{y_z(c')}{s_y}$ and $\frac{z_x(c')}{s_z}$ should be smaller than 1. Second, as already discussed, an asymmetry is modeled by using different values for $s_y^+$, $s_y^-$, $s_z^+$ and $s_z^-$. However, a certain asymmetry already exists in the dataset since the chosen alternative is on average cheaper than the other alternatives and/or its in-vehicle time is shorter. Finally, from the formula it can be seen that the parameter $\gamma$ weights the right part of the formula and therefore eventually influences the value of $\text{IND}(c)$.

For $s_x$ values between 120 and 720 minutes have been tested, whereas $s_y$ is varied between 60 and 780 minutes for $y_z(c') \leq 0$ and between 30 and 780 minutes for the remaining cases, with an average PJT of 90 minutes in the chosen alternatives. The range of tested values for $s_z$ goes from 15% to 170% percent of the fare for $z_z(c') \leq 0$ and from 10% to 170% for $z_z(c') > 0$. To evaluate the effect of these different parameter settings several combinations have been examined, some of which are presented in Table 2:
Table 2  Parameter settings for the independence measure

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symmetric setting</th>
<th>Extreme values</th>
<th>$\gamma = 0.25$</th>
<th>$\gamma = 0.75$</th>
<th>$\gamma = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_x$</td>
<td>720</td>
<td>360</td>
<td>720</td>
<td>720</td>
<td>720</td>
</tr>
<tr>
<td>$s_y^+$</td>
<td>780</td>
<td>180</td>
<td>780</td>
<td>780</td>
<td>780</td>
</tr>
<tr>
<td>$s_y^-$</td>
<td>780</td>
<td>120</td>
<td>540</td>
<td>540</td>
<td>540</td>
</tr>
<tr>
<td>$s_z^+$</td>
<td>1.7</td>
<td>1.1</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>$s_z^-$</td>
<td>1.7</td>
<td>0.9</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
<td>0.75</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The influence of the different parameter settings on the partial similarity components can be seen in Figure 1. The upper figure shows the average values of $\frac{x_c(c)}{s_x}$, $\frac{y_c(c')}{s_y}$ and $\frac{z_c(c')}{s_z}$ in the chosen and non-chosen alternatives. Generally speaking, they are larger for the chosen alternatives than for the non-chosen alternatives. Due to the above mentioned asymmetry $y_c(c')$ and $z_c(c')$ are usually positive for the comparison between the chosen and a non-chosen alternative. Thus, $s_y^-$ and $s_z^-$ are used resulting in higher partial similarity components. Furthermore, the values of $\left|y_c(c')\right|$ and $\left|z_c(c')\right|$ themselves may be larger for the comparison between the chosen and non-chosen alternatives than for those between the non-chosen alternatives. This is illustrated in the case where the parameters are set equal. The difference is most pronounced for the partial similarity component for fare.

The second parameter analysis examined the effect of varying $\gamma$ and is illustrated in the lower part of Figure 1, that shows the distribution of $IND(c)$ for chosen and non-chosen alternatives. It reduces or increases the influence of the differences in perceived journey time and fare on $IND(c)$. This leads to two effects: First, a higher $\gamma$ induces a higher total $IND(c)$. Second, the distribution of the independence measure becomes wider for higher $\gamma$. Since $\gamma$ weights the temporal similarity relative to the journey time and fare similarities, this
effects shows that the distribution of the temporal similarity in this dataset is wider than the
distribution of combined journey time and fare similarities.
MODELING RESULTS

Several models were estimated prior to the two final models that are presented in Table 3. Their formulation is identical except for the inclusion of the \( IND(c) \) factor in the second model. Subsequently, the MNL model without \( IND(c) \) will be discussed briefly before influence of the \( IND(c) \) factor is examined.

All parameters for carrier and flight attributes carry the expected signs. The estimated parameters for aircraft attributes follow the anticipated preference structure. Thereby, the dummy variables for individual carriers resulted in significantly better model results than those capturing specific carrier attributes. Obviously the perception of carriers is more complex than indicated by these attributes. However, from the modeler’s point of view it would be interesting to classify the carriers in a next step with respect to their attributes.

Regarding the treatment of the relationship between journey time and transfers, those models accounting for the total travel time and the number of transfers delivered better results. The other variant would be more intuitive for European flights, for which the waiting time represents a significant part of the overall journey time if a transfer is necessary. However, the number of chosen alternatives with a transfer is rather low in this sample. This leads to a better performance of models accounting for the number of transfers and not the actual waiting time.

Adding fare to the models increased their explanatory power significantly, as did the inclusion of outbound period of day and inbound period of day, which show an expected preference structure. Passengers who opted for a stay of a fortnight do not show a clear preference for the departure period of their inbound flight. Departing in the morning is preferred by all types of passengers. However, the inclusion of the latter variables influenced the sign and value of carrier variables. Seen in the light of passenger preferences for a certain departure time and the high number of observations of passengers staying at their destination for a short period of time, this is reasonable.
Table 3  Model results

<table>
<thead>
<tr>
<th>Model</th>
<th>MNL</th>
<th>MNL with IND(c) &amp; $\gamma = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter</td>
<td>Robust t-test</td>
</tr>
<tr>
<td><strong>Carrier constants</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flight attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code share</td>
<td>-0.7465</td>
<td>-12.81</td>
</tr>
<tr>
<td>Total travel time out (minutes)</td>
<td>-0.0006</td>
<td>-5.86</td>
</tr>
<tr>
<td>Number of transfers</td>
<td>-6.2404</td>
<td>-40.42</td>
</tr>
<tr>
<td><strong>Aircraft type attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional aircraft</td>
<td>-0.2170</td>
<td>-8.55</td>
</tr>
<tr>
<td>Propeller aircraft</td>
<td>-1.6427</td>
<td>-19.04</td>
</tr>
<tr>
<td><strong>Fare attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fare (€)</td>
<td>-0.0057</td>
<td>-73.71</td>
</tr>
<tr>
<td><strong>Departure time period attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Duration of stay 0 days</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outbound afternoon flight</td>
<td>-1.6057</td>
<td>-63.39</td>
</tr>
<tr>
<td>Outbound evening flight</td>
<td>-2.4960</td>
<td>-18.59</td>
</tr>
<tr>
<td><strong>Duration of stay 1 day</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outbound afternoon flight</td>
<td>-0.7807</td>
<td>-24.92</td>
</tr>
<tr>
<td>Outbound evening flight</td>
<td>-0.8573</td>
<td>-26.15</td>
</tr>
<tr>
<td><strong>Duration of stay 14 days</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outbound afternoon flight</td>
<td>-0.2829</td>
<td>-2.75</td>
</tr>
<tr>
<td>Outbound evening flight</td>
<td>Non-significant</td>
<td></td>
</tr>
<tr>
<td><strong>Duration of stay 0 days</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inbound afternoon flight</td>
<td>Non-significant</td>
<td></td>
</tr>
<tr>
<td>Inbound evening flight</td>
<td>0.1907</td>
<td>2.04</td>
</tr>
<tr>
<td><strong>Duration of stay 1 day</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inbound afternoon flight</td>
<td>1.1967</td>
<td>12.90</td>
</tr>
<tr>
<td>Inbound evening flight</td>
<td>1.5915</td>
<td>17.28</td>
</tr>
<tr>
<td><strong>Duration of stay 14 days inbound flights</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-significant</td>
<td></td>
</tr>
<tr>
<td><strong>Similarity attribute</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity measure</td>
<td>0.0000</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Summary statistics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of observations</td>
<td>18479</td>
<td>18479</td>
</tr>
<tr>
<td>Initial LL</td>
<td>-69279</td>
<td>-69279</td>
</tr>
<tr>
<td>Final LL</td>
<td>-48816</td>
<td>-48712</td>
</tr>
<tr>
<td>R-square</td>
<td>0.2954</td>
<td>0.2969</td>
</tr>
</tbody>
</table>
By adding $IND(c)$ to the model, the parameters for carrier attributes, aircraft attributes, fare and inbound period of day attributes remain constant. The parameters for outbound period of day change slightly: the parameter for outbound evening flights becomes smaller for passengers returning the same day. The opposite is the case for passengers returning the next day. Here the parameter for the outbound afternoon flight becomes larger. The estimated parameter for total travel time becomes insignificant. These changes can be explained by the inclusion of $IND(c)$, which includes differences in arrival and departure time and the perceived journey time.

Most interesting is the sign and value of the parameter of $IND(c)$ itself. Recalling that this attributes enters the utility function in logarithmic form and is between the 0 and 1, its negative sign implies that the decision-maker perceive alternatives similar to the chosen alternative as positive. Though this is a contradiction to the implicit availability/perception theory by Cascetta et al. (1996), it is in accordance with earlier findings by Hoogendorn-Lanser and Bovy (2007) or Frejinger and Bierlaire (2007). They showed that similarity can have a positive influence on the utility of ground-based transport alternatives. Their interpretation refer this to the possibility to switch routes or connections while the passenger is traveling. However, this is not likely to be decisive for air transport connections. It rather appears that decision-makers have a strong preference for certain alternative attributes such as a specific departure time, travel time or fare. Thus, they prefer alternatives that are similar with respect to these attributes and choose one of them. However, the influence of the individual similarity components needs further investigation.

As already discussed, it is possible to influence the weight of the similarity in perceived journey time and fare against the temporal similarity of alternatives by varying the value of $\gamma$. The estimation of different settings for $\gamma$ results in a changed parameter for $IND(c)$, while other parameters and the explanatory power remaining constant. However, the different parameter values correct for the change of $IND(c)$ by increasing when $IND(c)$ is
decreasing and vice versa. Thus shows that varying $\gamma$ is of less importance, what is also illustrated in Table 4. In this table, the estimated parameters for the $IND(c)$ factor are shown, together with the average value of $IND(c)$ in the chosen alternatives, the logarithm of $IND(c)$ and the partial utility based on its average value. It can be seen that the setting of the parameters influences the partial utility, especially the setting of the parameter for the temporal similarity $s_x$. A low $s_x$ results in a high independence of air connections. Implicitly, this is a penalty for departing and/or arriving too late or too early and results in a lower utility compared to the other settings, but still a positive one. Thus, it can be concluded that the traveler’s perception of similarity between connections is indeed multi-dimensional.

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimated parameter for $IND(c)$</th>
<th>Average value for $IND(c)$ chosen</th>
<th>ln($IND(c)$) Utility</th>
<th>final log-likelihood</th>
<th>r-square</th>
<th>Difference in r-square compared to MNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.2948</td>
<td>-</td>
</tr>
<tr>
<td>$s_x = 720, \gamma = 0.25$</td>
<td>-1.1966</td>
<td>0.0386</td>
<td>-3.2545</td>
<td>3.8943</td>
<td>-48816</td>
<td>0.5%</td>
</tr>
<tr>
<td>$s_x = 720, \gamma = 0.5$</td>
<td>-1.0362</td>
<td>0.0473</td>
<td>-3.0512</td>
<td>3.1617</td>
<td>-48684</td>
<td>0.6%</td>
</tr>
<tr>
<td>$s_x = 720, \gamma = 0.75$</td>
<td>-0.7210</td>
<td>0.0621</td>
<td>-2.7790</td>
<td>2.0037</td>
<td>-48752</td>
<td>0.3%</td>
</tr>
<tr>
<td>$s_x = 120, \gamma = 0.5$</td>
<td>-0.1835</td>
<td>0.1650</td>
<td>-1.8018</td>
<td>0.3306</td>
<td>-48798</td>
<td>0.1%</td>
</tr>
<tr>
<td>$s_x = 240, \gamma = 0.5$</td>
<td>-0.6776</td>
<td>0.0984</td>
<td>-2.3187</td>
<td>1.5712</td>
<td>-48622</td>
<td>0.9%</td>
</tr>
<tr>
<td>$s_x = 360, \gamma = 0.5$</td>
<td>-0.7666</td>
<td>0.0734</td>
<td>-2.0118</td>
<td>2.0222</td>
<td>-48640</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

**CONCLUSION**

This paper explores the way to account for similarities/correlations between air connections. For this purpose, an MNL model for air connections in Europe is estimated. The model includes carrier attributes, aircraft attributes and a fare attribute and is based on revealed preference data, such as actually booked tickets through CRS systems and fares as observed on the web-based booking system Expedia.
In order to account for similarities and to overcome the IIA property of the classical MNL model, an independence factor is added to the utility function, which captures the similarity of a connection with the other routes in its choice set. The main difference compared to earlier presented overlap measures is that the degree of similarity is calculated across multiple dimensions, namely a two temporal (departure time and journey time), and one monetary dimension. These three dimensions correspond to differences between air connections in particular and to public transport connections in general, where spatial overlap may be unknown or unimportant to the decision-maker.

For the calculation of the independence factor, several parameters are required which set the range of influence. Correct setting of these parameters is required for the independence measure to work properly. Experiments are made with several combinations of parameter settings. These experiments lead to better insights into the degree of similarity between alternatives in a choice set. Perhaps even more important, the experiments show that the independence or similarity of an alternative can be quantified and the relative similarity of the several dimensions can be determined. Especially with revealed preference data, where not much knowledge of the decision-maker’s preferences is available, the inclusion of multiple dimensions can lead to more insight in preferences without penalizing other alternatives too strong.

In order to evaluate the effect of the independence measure on the utility of an air connection, the measure is added to the utility function. Estimated parameters for the independence measure are highly significant and the model performance increases. Thus, it can be shown that the independence measure captures the similarity between alternatives at least to a certain extent without increasing estimation time.

For all models the sign of the parameter indicates that passengers perceive similar alternatives as positive. Thus, they prefer alternatives that are similar with respect to certain attributes and choose one of them.
However, the influence of the individual similarity components needs further investigation. This remains to be done, together with a comparison with other model structures, such as (cross)-nested model structures. With regard to the used data, other application fields can be imagined, such as a potential application in the field of revenue management. A closer look to the formation of the choice sets needs to be given then as well.
LITERATURE


