Results of Discrete Choice Models for Long-Distance Travel based on the DATELINE Survey

Jeremy Hackney, IVT, ETH Zürich

Conference paper STRC 2006

STRC

6th Swiss Transport Research Conference
Monte Verità / Ascona, March 15-17, 2006

Title of paper

Jeremy Hackney
IVT, ETH Zürich
Zürich

Phone: 0446333325
Fax:
email:
hackney@ivt.baug.ethz.ch

author 2
institution
city

Phone:
Fax:
email:

author 3
institution
city

Phone:
Fax:
email:

March 2006

Abstract

This working paper describes the enrichment of the Dateline long-distance travel dataset with synthetic mode and destination choice sets and the results of logit modelling. Dateline is a revealed preference survey of travel of 55,000 residents of 16 European countries in 2001. It contains detailed trip stage and sociodemographic information. Synthetic mode choice sets were constructed by merging the attributes of non-chosen road, train, and air alternatives to the geocoded origins and destinations. The synthetic destination choice sets consist of the chosen destination plus 9 randomly chosen destination alternatives at NUTS3 level. The attributes of the destinations are taken from the Eurostat geographic databases GISCO and Regio/New Cronos. Other Pan-European discrete choice models at this resolution are not known. It is the first known application of discrete choice estimation to the Dateline dataset. This paper presents the construction of the dataset and its quality, the model specification, and the model results. The models show a significant interactive influence of gender and trip purpose on mode choice, with trip distance attenuating the travel time. Gross domestic product and purchasing power parity at regional scale, and the number of hotel beds, significantly explain destination choice.

Keywords

mode choice, joint estimation, destination choice, Dateline, long-distance travel, synthetic choice sets

1. Introduction

This paper describes a dataset and presents model results for jointly estimated disaggregate mode and destination choice models of long distance travel between NUTS3 (Nomenclature of Territorial Units for Statisics) zones within the EU15, EFTA and CEC countries, based on enrichments of the Dateline revealed preference travel survey (DATELINE Consortium 2003). This is the first application of Dateline to transportation behavior modelling. The survey includes mode choice for journey stages (="trips"), but no further information about the transportation alternatives or alternate destinations in the choice set of the decision makers. Enrichment with constructed choice sets was necessary.

The Dateline survey and dataset are described first. Then the steps are outlined for the construction of the choice set for the transportation modes for the chosen destinations at municpal resolution, followed by the construction of the regional attractiveness dataset for the chosen and non-chosen destinations at NUTS3 level. Descriptive statistics of the two enriched datasets are summarized. Two discrete choice models are then described. The mode choice model is a multinomial logit estimated on the municipal-level Dateline data, enriched with mode attributes for chosen and non-chosen modes. The joint destination-mode choice model is also a multinomial logit (MNL) model, estimated on the dataset that was enriched with location attributes at NUTS3 level, as well as alternative mode attributes. The models have significant coefficients of expected sign, with high explanatory power. A full investigation of the models' error terms and IIA behavior has not yet been carried out, and alternatives to MNL have not yet been proposed.

1.1 Studies of Long Distance Travel

Disaggregate models have not been used as much in long-distance travel as in everyday travel, primarily because of a lack of data of sufficient quality and detail. Highly skewed distributions of travel frequency, distance, duration, etc., are common in long-distance travel, and compound the bias effects of small sample sizes. For example, Last, Manz, Zumkeller (2003) found that half of the population in their sample produced over 90% of the long-distance journeys. But the most mobile one percent of the population made long-distance journeys ten times more frequently than the national average. It is therefore important to have sampled these frequent travellers.

The modelling of long-distance travel is further hampered by heterogeneous behavior among similar individuals in the population. Hubert and Potier (2003) point out, for example, that the ownership of a vacation home might be the most important determinant of a household's vacation travel, but this information would rarely appear in a transportation survey and is most likely independent of the level of service of transportation modes to the destination.

Reports with which to compare the aggregate long-distance travel statistics revealed in Dateline are not available for all of Europe, but exist in various countries (Last, Manz, Chlond, and Zumkeller 2004, Hubert and Potier 2003, Swiss Federal Statistical Office 1998).

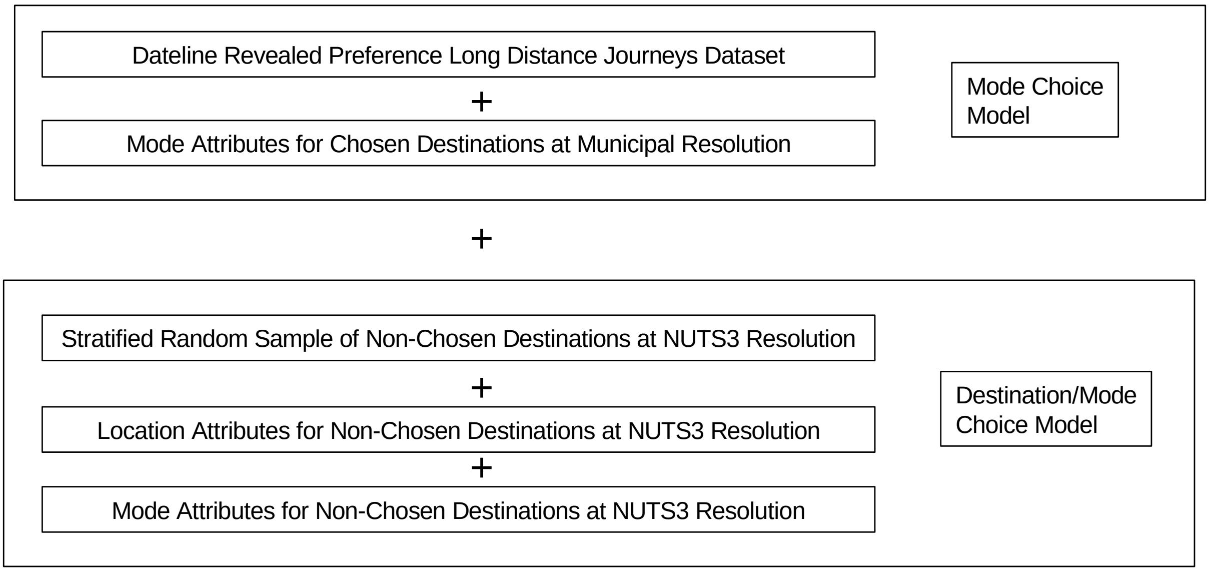
Previous work is also found for formulations of the mode- and mode/destination choice model. Discussions of the influence of socioeconomic variables are found in Limtanakool, Dijst, and Schwanen (2004), Eymann and Ronning (1997), Last and Manz (2005). The nested logit specification of long-distance mode/destination choice models has been presented in Koppelman and Sethi (2005), Eymann and Ronning (1997), Vrtic et. al, (2005).

Long-distance travel by train or airplane is often characterized by a choice of departure station or airport. The choice of departure airport has been determined by Hess (2005) and Kroes, Lierens, Kouwenhoven (2005) to be a complex interaction of accessibility, airport amenities, and the airlines available. Realism is sacrificed in this first analysis by only accounting for accessibility measures to the airports.

1.2 Project Overview and Contribution

The Dateline stated preference dataset (DATELINE Consortium 2003) is enriched with attributes of available transportation modes, alternative destinations, and attributes of the alternative destinations. Figure 1 is a schematic diagram showing the enrichment steps preceding the estimation of the two models presented here.

Figure 1 Construction of the Enriched Long-Distance Travel Dataset


Dateline Revealed Preference Long Distance Journeys Dataset
+
Mode Attributes for Chosen Destinations at Municipal Resolution
Mode Choice Model
+
Stratified Random Sample of Non-Chosen Destinations at NUTS3 Resolution
+
Location Attributes for Non-Chosen Destinations at NUTS3 Resolution
+
Mode Attributes for Non-Chosen Destinations at NUTS3 Resolution
Destination/Mode Choice Model

The enrichment process was preceded by collecting and correcting data and models of mode attributes for trans-european services and the attributes of regions.

Dateline is the first standardized pan-European long-distance travel survey. Davidson (2003) has performed a journey distribution analysis and re-weighting of the Dateline journeys OD matrix at NUTS1 level for the EU15 + Switzerland by comparison with the O/D matrices from the MYSTIC project.

This is the first attempt to model travel decisions based on the full resolution of the Dateline dataset at either the municipal or the NUTS3 level. The higher the NUTS number, the higher the resolution: for example, for Germany, there is 1 NUTS0 region, 16 NUTS1 regions (Länder), 40 NUTS2 regions (Bezirke), and 441 NUTS3 regions (Kreise). Municipalities are NUTS5.

The higher regional resolution allows more precise determination of travel impedance, because the beginning and end of the journey can be represented in more detail. The advantage of higher aggregation is that more data is typically available than for smaller regions. NUTS3 is the smallest practical regional unit for which attributes are formally compiled. Even so, data was inconsistently availabile.

The NUTS3 models and dataset are to be the basis for the estimation of a Pan-European O/D matrix on the three transportation modes. A route choice model is also desired from the data, though additional enrichment of the trips (journey stages) would have to be undertaken. A route choice model and an O/D matrix at NUTS2 level would be sufficient to describe the traffic on the Trans-European (TEN) network, meaning modelling at NUTS3 level, and justifying the effort given to constructing the dataset at this high resolution.

2. Description of the Dateline dataset

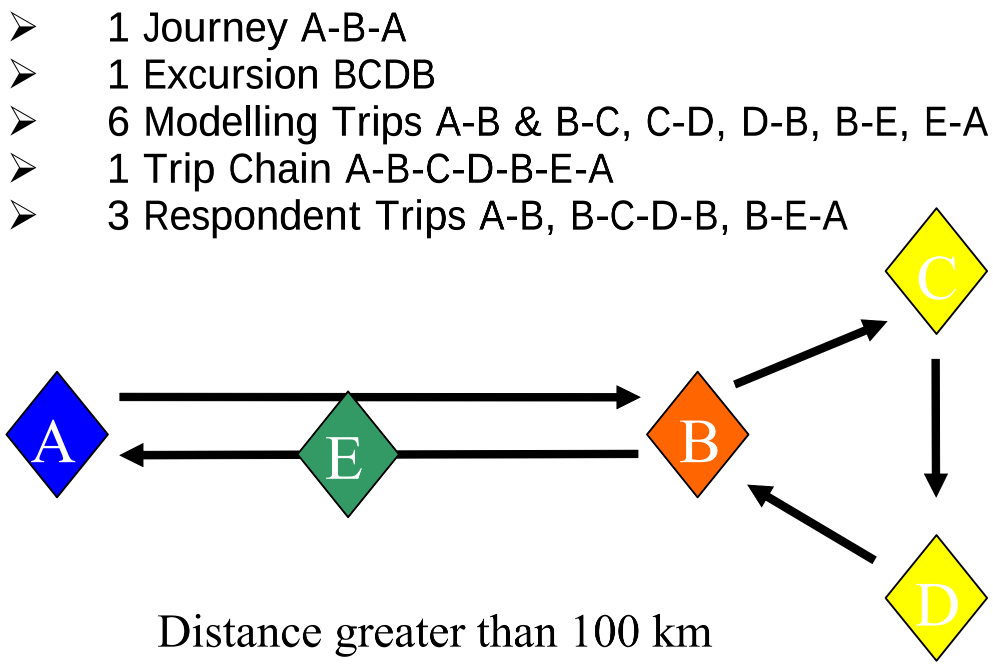
Dateline is a household-level survey of 86,000 residents of the EU 15 and Switzerland about their long-distance travel. Individuals over 15 years of age reported travel of over 100km crow-fly distance for the purposes of “holiday” in the previous 12 months, as well as “other private” and “business” in the previous 3 months, and “commuting” for the previous 4 weeks. The survey was carried out from October, 2001 through October, 2002. Along with socio-demographic variables, the dataset contains travel date, origin/destination, duration, and mode. Specific travel decisions like mode changes were recorded for travel between significant stops en route. Thus, a coarse record of route choice, in the form of stops en route, is also available, in particular for public transport based journeys. The data is available to browse or to download on the ETH Travel Data Archive (2004): http://129.132.96.89/nesstarlight/index.jsp

2.1 Trip-Chain Structure of Dateline

Travel in Dateline is coded into round-trip journeys and commutes in which the final destination is over 100km distant from the origin, and into excursions, which embark from and return to the journey destination. The stages of all these basic movements are recorded, yielding a series of trip chains (Figure 2, Davidson 2003), in which the end or beginning of a trip is determined by a stop of over 2 hours with a distinct purpose.

The modelling presented here does not use the trip information or the commutes and excursions. It focuses on the decisions made in planning journeys, thus the overall choice of travel mode and destination.

Figure 2 Journey Concepts



- 1 Journey A-B-A
- 1 Excursion BCDB
- 6 Modelling Trips A-B & B-C, C-D, D-B, B-E, E-A
- 1 Trip Chain A-B-C-D-B-E-A
- 3 Respondent Trips A-B, B-C-D-B, B-E-A

Distance greater than 100 km

Source: Peter Davidson Consultancy (2003)

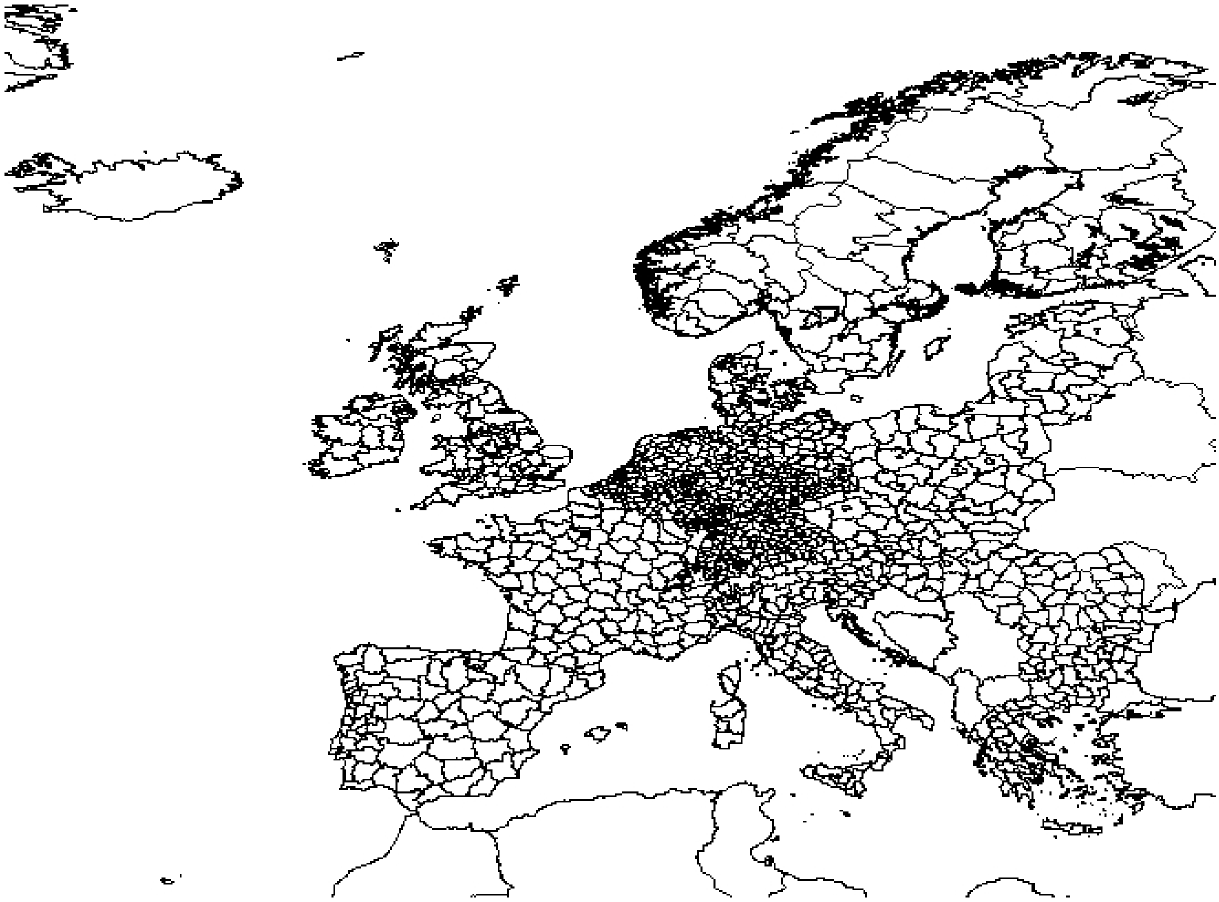
2.2 Geocoding of the Dateline Data

The endpoints of trips and journeys are coded to the name of the muncipality. Locations on continents other than Europe and European territories are denoted with an "X" and are not geocoded more precisely than that. Other journeys where the travel and the destination were identical (cruise ship) were also not usable in the analysis. A file of geocodes for municipalities is delivered with the Dateline dataset with over 2.9 million entries (including various spellings) so that longitude and latitude can be associated with the municipality name. If there are errors or missing data in any of the datasets, the place name cannot be matched to the coordinates, and the observation can not be geocoded. Thus, associating Dateline with geographic coordinates results in a loss of data.

2.3 Explanation of NUTS3 Geocodes

In addition to an X,Y geocode, Dateline points are associated with a NUTS 3 region which enables further associations to be made with regional attributes. The non-chosen destinations the transportation modes to these destinations are also coded to the level of NUTS3 regions. The locations of the NUTS3 centroids are available in the GISCO 2001 database (EUROSTAT 2001). The NUTS version used in the project is based on version 6.0 of 1998, plus the SABE database of seamless administrative boundaries for the CEC and EFTA countries. This set of regions is combined in the GISCO dataset, where it is called version 7. The codes used in Dateline, however, replace the two NUTS regions for Berlin (DE301 and DE302) with the newer, single zone DE3. The set of NUTS3 zones used here is therefore slightly non-standard. The zones in CEC countries which are included in various NUTS versions also have various names in various data sources. The set of NUTS3 zones used in this study contains 1355 zones. Identifying and adjusting the NUTS zones in the datasets of regional attributes is a challenge, and the approach used is described in detail in the extended version of this paper and to some extent in section 5.

Figure 3 NUTS3 Version 6 + CEC and EFTA regions



2.4 Summary Statistics

2.4.1 Volume of Available Data

Table 1 summarizes the volume of the geocoded journey data. The journeys that are not usable after geocoding to longitude and latitude consist of journeys to/from other continents, ocean cruises, or missing data.

Table 1 Volume of the Dateline Journey Data

Nationalities surveyed	EU 15 + CH = 16
Number of Journeys	97,196
Number usable after geocoding	84,126
Number of Households	55,546
Number of Travelling Individuals	109,495
Unique points (origin/destination)	5663 O/ 5430 D
Unique OD Pairs	16,063

2.4.2 Mode Share

The Dateline database lists 13 types of modes, which are aggregated here into groups ("main mode") for analysis and modelling in Table 2. Ferry transport has a significant mode share in European seas: Mediterranean, the English Channel, the North Sea, and the Baltic regions. In the models, ferries are not modes themselves, but are associated with the mode that the ferry carries (train = rail; motor vehicle = road), where the speed of the ferry leg is the speed of the ferry. Inland waterway travel and ocean cruises are not considered. Tour buses/coaches are disregarded, as described in section 3.3.

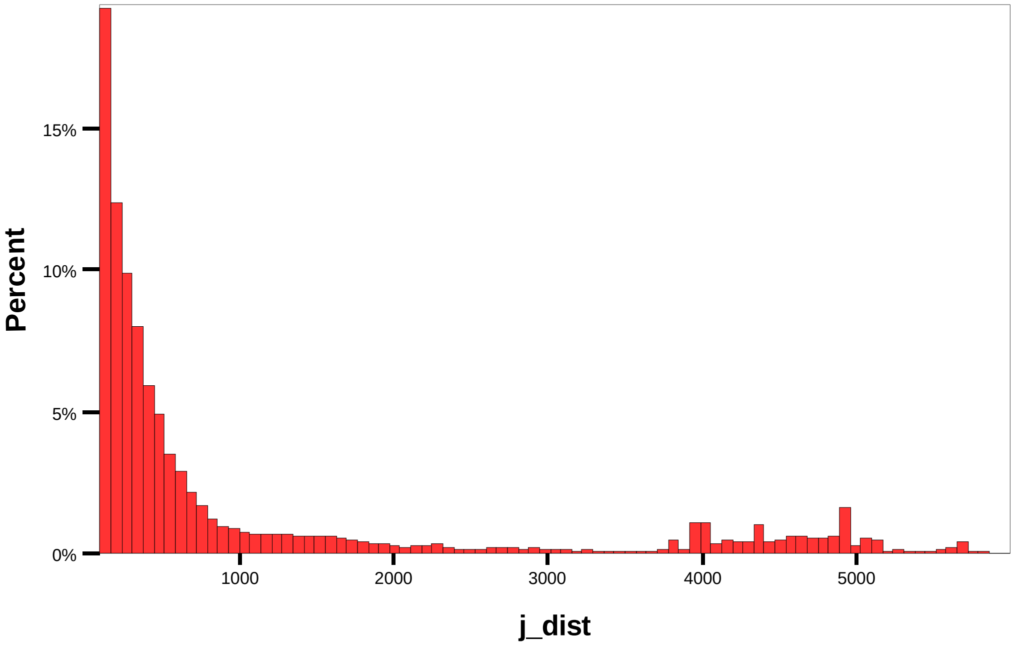
Table 2 Dateline Journey Mode Share (Aggregated Modes)

Main Mode	Share (%)	Km (%)
Road (Car as driver/passenger, motor home, truck, motorcycle)	60	28
Tour bus/coach	9	6
Air	20	60
Train (inc. Motorail)	10	5

2.4.3 Journey Distance Distribution

Dateline was structured in order to oversample long journeys, which are so infrequent that they would otherwise not be detected in such a small sample size (Davidson 2003). This corresponds to the rise in the histogram in Figure 4 at 4000-5000 km distance.

Straight line, one-way distance distribution (km) of journeys in Dateline



2.4.4 Journey Frequency

The dataset exhibits much lower journey rates (journeys/person/year) than national long-distance travel surveys have shown (compare with national surveys in Hubert and Potier 2003, Swiss Federal Statistical Office 1998, Last, Manz, Chlond and Zumkeller 2004). This inconsistency may have to do with the recall-oriented survey technique of Dateline, as compared to a travel diary. Dateline is therefore not useful for reconstructing OD volume matrices.

2.5 Variables used from Dateline

The individual observations from Dateline are used for the disaggregate modelling without weighting factors. The maximimum likelihood estimation of logit parameters based on non-random exogenous samples (enriched or double samples) simplifies to the estimation of parameters based on a random sample, and no correction to the sample is necessary (Ben Akiva and Lerman 1985).

Data from Dateline is extracted from the “journeys” records, which were first geocoded to latitude and longitude, the table of household characteristics, and the table of travellers (“persons”). The table called “participants” was used to key “households” to “persons”. All of the sociodemographic variables available in Dateline are initially retained for their potential to explain long-distance travel behavior. Variables from Dateline were extracted and kept in the dataset for potential applications in the future. Many peripherally relevant variables were subsequently removed to reduce the size of the dataset before estimation of the mode choice models (section 8.1). Individual or household income is not available in the survey. The discussion of the relevance of the variables to mode and destination choice is in section 7 on modelling.

3. Enrichment with Mode Alternatives

This section describes the data sources for transportation mode attributes of the time-shortest travel alternative for three modes: air, rail, and road. The attributes that are used to enrich the datasets of chosen and non-chosen destinations, respectively, come from different sources. The origin/destination pairs revealed in Dateline are recorded at municipal resolution. The non-chosen alternative destinations are at NUTS3 resolution (their generation is described in section 4, and the NUTS zones are described in section 2.3). Both OD sets are large for the chosen and non-chosen destination sets. Because the precise OD pairs for the chosen destinations are known from Dateline, there are ca. 16,000 relations. But because the set of non-chosen destinations are chosen stochastically, mode attributes are needed for all 2.5 million relations. The search for alternatives has to be automated for this large data volume, placing restrictions on the tools and data sources that could be used. The mode attributes are taken from batch-searchable timetables and network models (Table 3).

Table 3 Data sources for transportation mode attributes

Data	Dateline (chosen destinations)	Dateline + non-chosen destinations
Dateline (chosen)	Municipal	Municipal (1) + NUTS3 (9)
Origins/Destinations in choice set	5663 O/ 5430 D	1355/1355
OD relations in choice set	16,063	2.5 million
Road attributes	PTV High Resolution Road Model	IVT Trans-European Road Model
Rail attributes	IVT Trans-European Rail Model 2002	HaCon Deutsche Bahn Timetable 2004
Air attributes	IVT European Air Travel Model	IVT European Air Travel Model

It is assumed that the traveller has complete knowledge about the shortest-time services to the journey destination on all modes. It is also assumed that only one of the main modes will be used for the entire journey.

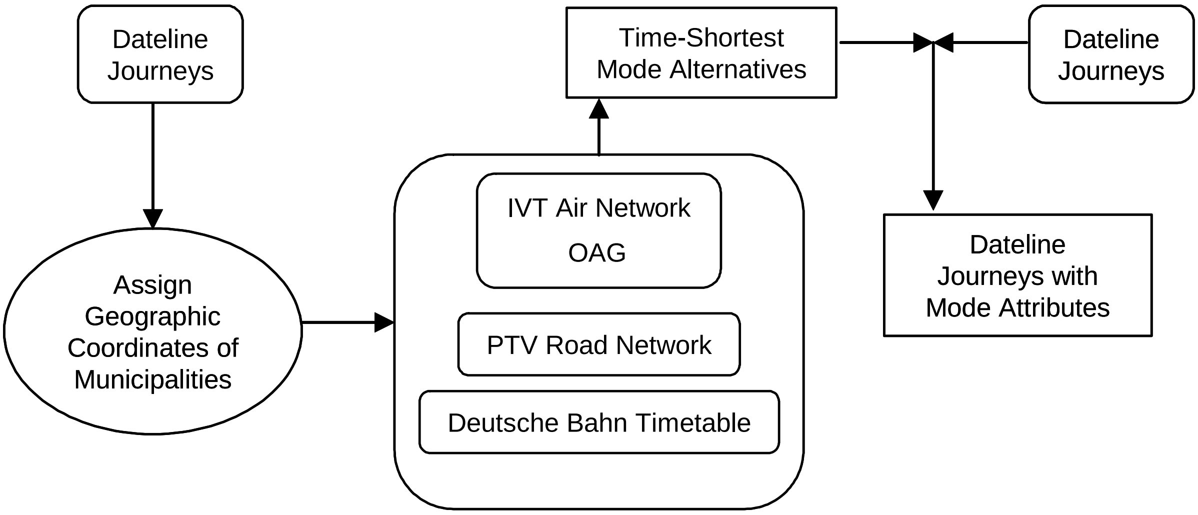
The travel impedance models for the air, rail, and road modes at IVT were improved in the context of the project, with the aim of supporting destination and route choice, as well as inter-modal models in the future. However, the road and rail models were only useful for the non-chosen destination alternatives, at the coarser NUTS3 level.

The IVT air travel model was sufficient for representing both municipal and NUTS3 interzonal travel after minimal improvement to the zones and connectors. This is because the its accuracy depends only on accurate representation of access to the airports by the road and rail modes.

Other data sources were used to establish the set of rail and road alternatives for OD pairs at the municipal level (chosen destinations). For the rail mode, a batch-search was carried out with the Hafas server (HaCon AG) on one sample day in the 2004 Deutsche Bahn timetable. The dataset for the base year, 2002, was not available for batch search. This dataset also provides a standard of comparison for the rail model that can be used for calibration in the future. For the inter-municipal modelling of the road mode, a special high-resolution network was used from PTV AG. This road model also provided the high-resolution impedance measures between municipalities and airports for the air travel model.

Figure 5 depicts the enrichment of Dateline chosen destinations with mode attributes. The enrichment for the non-chosen destinations is similar. A detailed schematic is in the extended version of this paper. The description of the data sources and the enrichment procedure follows in this chapter.

Figure 5 Schematic Diagram of the Enrichment of Dateline Chosen Destinations with Mode Alternatives


Dateline Journeys
Assign Geographic Coordinates of Municipalities
IVT Air Network
OAG
PTV Road Network
Deutsche Bahn Timetable
Time-Shortest Mode Alternatives
Dateline Journeys
Dateline Journeys with Mode Attributes

3.1 A Note on the Monetary Cost of Travel

The general impedance for mode attributes consists of measures of convenience, travel time and out of pocket cost. However, the Dateline survey did not ask respondents to recall the price of their travel. Regression models are available to reconstruct the likely costs for rail and road modes, but no reliable estimates for air ticket prices were available. However, the modelling presented here did not use prices. Description of the regression models appears in the longer version of this paper.

3.2 Rail

The rail impedance for the chosen destinations was taken from the representative-day search of the Deutsche Bahn database using the Hafas server. This sample was extracted for IVT by an out of house partner which has developed a program to perform batch searches using the server version of the Hafas program. The data is a mySQL database. The OD matrix for the search is the set of major train stations used in ETIS Base, one major train station per NUTS3 region. The database contains the 1355 x 1355 matrix of the time-shortest connections between all train stations, independent of accessibility to the station, and the rail connection to all airports within 700 km straight line distance of each train station. All information contained in the database is included: all stations en route, all stops, station ID numbers, station name and geocode, transfers, travel time between stations, departure time, arrival time, waiting times, train type and number, etc. The dataset comprises 33 GB.

The time-shortest total route, including access to/from the train station, is the basis for the mode choice in this study. The 3 nearest stations to the Dateline origins and destinations were first determined in ArcMap. The attributes of the rail connection between the resulting 9 combinations of stations per person-journey were extracted from the mySQL database with a batch query. The 9 time-shortest rail connections were merged with the Dateline dataset. The straight-line access distance is written out by ArcMap. After the road and air connections are added to the dataset, this access distance to/from the train station is converted to time by assuming a speed of 40 km/h, and the rail connection with the lowest total time is retained for each person-journey.

The IVT rail network was improved with the long-term intent of being able to use the model for a route choice dataset and to calculate rail costs. It was not finished and checked by the time the unchosen mode alternatives were added to the Dateline dataset. An additional complication is that there are too many Dateline points to be able to use VISUM for shortest-path searches without breaking the dataset up into at least 5 parts. However, the IVT model was used for the rail connections to the non-chosen destinations, which are only defined at the NUTS3 level. This was done in order to 1) test the improvements to the rail model and 2) because it was less labor-intensive than repeating the calculations done in ArcMap and mySQL for nine sets of NUTS3 centroids. Time-shortest paths were found on a complete OD matrix of 1355 x 1355 NUTS3 centroids. A comparison with the Deutsche Bahn timetable would still be desirable, however.

The IVT model based on the Thomas Cook Timetable of September 2002 in the VISUM environment is orginally described in Bleisch and Fröhlich (2003). It was improved in ETIS Base by the addition of stations, lines, and line routes by Hackney and Ripka (Hackney 2004). Altogether, 739 new stations and 1394 time profiles (services) were added to the BAK model, increases of 37% and 26% over the initial state. The new model has 2736 served stations and 6537 time profiles (services). The improvements were made primarily in eastern and northern Europe, with important updates in Spain and the UK. The distribution of the lines in the European countries is summarized in Table 4.

Table 4 Number of Trains (Time Profiles/Services) to each Country

Country	Lines	Country	Lines	Country	Lines
Albania	19	Greece	73	Slovenia	6
Austria	211	Holland	90	Spain	277
Belarus	4	Hungary	107	Sweden	193
Belgium	76	Ireland	77	Switzerland	668
Bulgaria	46	Italy	773	Turkey	18
Croatia	8	Macedonia	2	Ukraine	8
Czech Republic	106	Norway	61	Slovenia	6
Denmark	180	Poland	153		
Finland	119	Portugal	63		
France	738	Romania	93	International*	781
Germany	976	Russia	9	Airport Connections	173
Great Britain	517	Slovakia	105		

* International lines are those which cross several countries, especially Eurocity trains

As more serviced train stations are added to the model, the choice of stations to attach to demand centers (NUTS3 centroids) requires special care. Using the nearest 3 stations to the centroid, for example, will often result in having only 3 very small and marginally useful train stations attached to the NUTS3 centroid. This was less of a problem when only major train stations were carrying traffic (and therefore connected to the demand zones). The stations associated with the NUTS3 regional centroids were carefully chosen on the basis of size (number of lines), accessibility, and national or regional identity (Ripka, personal communication 2005). As in the Hafas search, a fixed access speed of 40km/h in a straight line was used for the access to and from the stations. For a valid connection, a maximum of 6 train changes is allowed, and the waiting time for train changes my also not exceed six hours. This treatment sets a minimum measure for level of comfort. In order not to constrain the search too tightly, it is conducted over a departure period of 24 hours, and a completed connection is permitted to take up to 7 days.

The detailed procedure for constructing the transportation alternatives for the non-chosen (NUTS3) destinations is in section 6.2.

3.3 Road

Tour buses have significant mode share (Table 2), but they are left out of the mode set due to a lack of summarized data about the bus schedules, routes and stops, restrictions in countries of operation (such as laws restricting competition with national train systems), price, special package offers, and other service variables. The researcher would need to compile this information from individual tour bus agencies, a high time investment.

The large number of origins and destinations in the Dateline sample exceeds the license of the VISUM software at IVT. The OD matrix cannot be broken in to parts, as it could be for the rail model, since road service depends on the flow results of an assignment calculation. So the assignment of Dateline to the road network was outsourced to PTV, the makers of the software. Because it is the intent to use the Dateline database for route choice, the trip geocodes at stage level, rather than simply the journey endpoints, were sent for processing. In addition to relieving the limitations of the software license, having PTV carry out the work enabled the use of the high resolution PTV basis network (PTV, 2004), as an improvement over the IVT road network. Two files of road travel times were calculated: the road travel times between Dateline trip origins and destinations, and the road travel time between the Dateline points and the 3 nearest train stations and 3 nearest airports.

The point-to-point file provides the road travel time for the chosen journey destinations. In order to calculate the OD travel time for the journeys, the trip (stage) travel times have to be added together. This gives not only the OD travel time, but the time specific to the route chosen. There are inconsistencies in the place names between the files of trips and the file of journeys in Dateline, though, with the result that the aggregation of the PTV calculation over trips does not always give a road travel time that matches the endpoints of the journey. The journeys affected in this way are not usable without detailed work on each case to locate the inconsistency. This was not done for this study, contributing to the reduction in data available for modelling.

PTV was not asked to find the shortest-path road travel times between NUTS3 regions. The free-flow time from the IVT European Road Model (Bleisch and Fröhlich 2003) was used for the non-chosen destinations at NUTS3 resolution (Section 6.2).

3.4 Air

The air travel attributes are taken from the IVT european air travel model with the OAG flight schedule for September 2002 (Bleisch and Fröhlich 2003). This sample of flights does not include charter flights or the newer low-cost carriers. The set of 3 nearest airports to the Dateline points is used in the search for time-shortest flights. The access times to and from the airports are provided by the high-resolution road travel file from PTV, described above. As for the train connections, a set of 9 time shortest air connections are assembled for each person-journey. Likewise, the single connection with the shortest total travel time is chosen in a final data-processing step based on total shortest time.

For the non-chosen destinations, the access time from all NUTS3 centroids to all airports was provided by the IVT European road network, which includes airports as zones in the network. The air connection with the shortest total time is then found by assignment (Section 6.2).

There are certainly more realistic ways to sample the airports for each of the municipal-level journey endpoints. Hess (2005) and Kroes, et. al (2005) have represented airport choice with complex discrete choice models, for example. More simply, choosing all airports within 700 km, as was done for the rail connections, may have been desirable for the road times, as well, for consistency with the rail choice set on the one hand, and in order not to leave out realistic but distant airport choices on the other. The rail sample gives the rail travel attributes from the NUTS3 centroids to the airports. But this search was not carried out for municipalities on the high-resolution road network.

4. Data Sources for Location Attributes

Both chosen and non-chosen destinations must be enriched with location attributes in order to model the destination choice of the travellers. Though the Dateline dataset provides geocodes of the municipalities of the origin and destination, collecting statistics about tens of thousands of specific municipalities and ensuring their methodological comparability was not feasible within the project budget. NUTS3 regions are a compromise resolution between the higher precision of the available survey and the relatively low precision of the available structure data.

The NUTS3 regions are characterized for the base year 2001-2 for the estimation of the destination choice model. The typical proxies for the attractiveness of destinations that have been used in the literature on long-distance travel behavior include indicators of population, economic productivity, tourist activity, and the environment or climate (e.g. Eymann et. al 1997). Regional variables for at least 16 countries were required for this study, corresponding to the countries of residence of the Dateline participants. However, the ETIS Base project of the EU, with which this modelling is associated, extended the region of interest to 27 countries. Data from as many of these as possible was collected.

4.1 Availability of regional data: Spatial coverage

Databases of the European Union, either EuroStat or Regio/New Cronos, were used because of the uniformity of the definitions of the variables, the compatibility of the data with the NUTS scheme chosen for Dateline, the availability of specially prepared new datasets from the ETIS Base project, and availability of the variable descriptions in English. While the use of EU data assures standardization and geographic consistency, it also limits the specificity of the available variables, since only information relevant to macroscopic concerns at the EU is collected and stored in EU databases. Local micro-drivers of tourism, for example, are not yet systematically surveyed and compiled at EU level. Additionally, not all EU databases are available for the desired reference year.

Cleaning the data consists of geocoding certain data to match that of the mode choice model, adjusting the year of the data if it is from other than 2001-2, and filling data holes. Missing data must be filled by imputation, or else the cases or the variable must be dropped from consideration because a utility cannot be calculated.

4.2 Availability of regional data: Variables

The relevant attributes of the regions which explain the observed travel decisions depend on journey purpose. Work (business) journeys depend on the level of business activities in various branches, and on the available infrastructure for doing business, such as holding meetings, accessibility to fast transportation modes, saturation of telecommunications, etc.(e.g. Last et. al, 2005) Holiday (leisure) journeys depend on local prices, seasonal climate, natural surroundings (beach, etc.), local accessibility measures, cultural features, landmarks, etc. Personal journeys depend on information about individuals and households that is not available to modellers, for example migration trends and trips to visit family in the country of emigration. No special effort is made in this study to find regional variables that describe these types of personal journeys.

Variables describing subtle regional characteristics are not easy to define and it is not expected that the proxies that are found will fully describe the attractiveness of the regions for business and holiday journeys. Likewise, the proxies will be so highly aggregated as to incorporate the effects of multiple attributes that influence travel behavior, resulting in loss of specificity in explanatory power. Finally, the choice of proxy is conditioned on what variables are available in the datasets to which we have access within a reasonable time frame and budget.

4.3 Description of Data Sources

Regional data are available to us from Regio/New Cronos (continually updated, Eurostat 2005), GISCO (Eurostat 2001) and country statistical offices (CH, NO).

4.3.1 Eurostat Regio Data from 2001a

The web portal at Eurostat for Regio data from New Cronos:

http://epp.eurostat.cec.eu.int/

Data from certain eastern countries (BG, RO, PL, etc.) and NO, CH are seldom available in Eurostat datasets and must be found in other sources or imputed (see below). NO data can be found in English at Statistics Norway:

www.ssb.no/english/

and Swiss statistics are found at the Swiss Statistical Office:

http://www.bfs.admin.ch/bfs/portal/de/index/infothek/lexikon/bienvenue___login/blank/zugang_lexikon.topic.1.html

The basis on which the statistics are calculated outside of the EU15 may not be the same as those used within the EU.

The variables used that come from or are based on New Cronos, and the equivalent variables for non-EU15 countries are listed in Table 5. The variables in *italics* have not been used in the final structure file.

The data was downloaded in *.dbf or other ASCII format. Scripts were written to import the data into SAS and to merge the variables together according to the NUTS3 region. This process is detailed in section 5.

Table 5 Regional Structure Data from Eurostat Regio

File	SAS name	Eurostat Description	Units	Detail
eht_reg	employed	Annual data on employment in technology and knowledge-intensive sectors at the regional level	1000 employees	NUTS2
eht_reg_pct	emppct	Annual data on employment in technology and knowledge-intensive sectors at the regional level	Percent	NUTS2
hh2inc xhh2inc	income	Income of households at NUTS level 2 & Non-EU25 Countries	mio_eur	NUTS2
s2sbs x_sbs	business	Structural business statistics by economic activity & Non-EU25 Countries	Number of enterprises	NUTS2
d3area xd3area	area	Total area of the regions & Non-EU25 Countries	km2	NUTS3
d3avg	pop	Average population by sex and age	1000 People	NUTS3
d3dens xd3dens	dens	Population density & Non-EU25 Countries	People/km2	NUTS3
pppsna95	ppp	Purchasing Power Parity	Indexed to EU25	NUTS3
e3gdp95 gdpBgRo	gdp	Gross domestic product (GDP), market prices at NUTS level 3	mio_eur	NUTS3
t_3r	beds	Number of beds	Number/1000	NUTS3

4.3.2 GISCO Land Use Coverages

GISCO is the Geographic Information System for the European Commission The GISCO dataset for 2001 was acquired during ETIS Base on CD. A detailed description of the steps to extract the GISCO data is in the extended version of this paper.

The files used for the destination attributes are listed in Table 6.

Table 6 Regional Structure Data from GISCO

File	SAS name	Description	Units	Detail
lstype	lstype	Landscape type (7 Categories) GIS Intersect	Binary	NUTS3
tempppt	tempppt	Average July maximum and January minimum temperatures, precipitation amount	Degree C, mm/month	NUTS3

5. Building the Dataset of Location Attributes

The files of the regional attributes are constructed by standardizing the NUTS regions of the structure datasets to the Dateline version of NUTS 1999, cleaning the source data by hand (removal of footnotes, standardization of MISSING variable, etc.); harmonizing the NUTS version within the file (if version 2003); merging the files together from highest to lowest resolution (NUTS3-NUTS2-NUTS0); converting all NUTS3 to the 1999 versions; and data imputation, if applicable. Detailed steps are described in the extended version of this report.

5.1 Adjusting NUTS Zones to 1999 Standard

A portion of the data must be translated from NUTS3 (2003) into NUTS3 (1999). Variables must be split and reallocated to regions in some cases, and in others they must be aggregated. The list of adjustments follows:

Table 7 Required NUTS Code Conversions

Code 2003	Code 1999	NUTS Level
ITC	*IT1+IT2*	1
ITD	*IT3+IT4*	1
ITE	*IT5+IT6*	1
ITF	*IT7+IT8+IT9*	1
ITG	*ITA+ITB*	1
PT16	*PT12+PT13 (part)*	2
DE41	*DE4 (part)*	2
DE42	*DE4 (part)*	2
ES63	ES63 (part)	2
ES64	ES63 (part)	2
FI18	*FI17 (part)+FI16*	2
FI19	*FI14 (part)+FI17 (part)*	2
FI1A	*FI14 (part)+FI15*	2
ITD1	IT31 (part)	2
ITD2	IT31 (part)	2
PT17	*PT13 (part)*	2
PT18	*PT14+PT13 (part)*	2
DE300	*DE301+DE302*	3
DE929	*DE921+DE924*	3

Following this scheme, converting the NUTS3 regions to the 1999 standard is simple. Only the Regio/New Cronos data used the 2003 geocodes. The 1999 NUTS3 regions for Berlin, DE301 and DE302, were not used in Dateline; the more modern DE3 was used, instead, and no adjustment to the location attributes needed to be made. The 1999 regions DE921 and DE924, both in Hannover, are disaggregated from the 2003 region DE929 by allocating the GDP, hotel beds, purchasing power parity, and population by the land area. This assumes a constant distribution of productivity, wealth, and population across both parts of the city.

Attributes of NUTS 2/1 regions are not currently used in the modelling, and are dropped from the dataset. If they are to be used for later imputation of values at NUTS3, they will also need to be adjusted from 2003 geocodes to 1999 geocodes.

5.2 Missing Data for Base Year

Filling in missing data for non-EU25 countries, particularly CH and NO, requires work by hand (e.g. Excel) to derive values comparable to those collected by the EU at NUTS3 level. Population is missing for many regions in BG and RO. The latter 2 are calculated by multiplying population density by land area.

Too much information is lacking at the NUTS3 level for “employment by branch” (eht_reg) and “enterprises by branch” (s2sbs) for these to be useful in the estimation of destination choice. The missing values may be imputable, but for the moment they are not used.

Information at the desired level of detail is not available for the Baltic states or Turkey, and much of the Balkans.

The other variables used to describe the alternative destinations are also incomplete to varying degrees. Automated imputation techniques to fill large areas of missing data have not been applied, and the data are used in this incomplete state. An alternative with missing data cannot be used in the discrete choice modelling, so it is desirable to fill as many holes as possible.

A detailed account of data holes and the treatments for them is in the extended version of this report. A coarse summary follows.

ES701, ES702, FR91, FR92, FR93, FR94, PT2, PT3, and UKM46 are islands whose airports are not represented in the available model and which do not have train stations linked to the rest of Europe. They are therefore dropped from consideration in constructing the choice sets for these modes.

Rail connections are additionally missing in 567 NUTS3 OD relations because the regions are reachable only via connectors. They are too close together in the model to be served by the high quality trains in the models, and the corresponding local train service is not represented. These are flagged “not available” in the discrete choice estimation.

In addition to the islands listed above, air connections are missing for 50351 OD relations over short distances where regions connect to each other only by model connectors and not by air services. These are flagged “unavailable” in the discrete choice modelling.

The UK islands are reachable by ferry in the road model. Aside from the distant islands listed above, there are no missing values in the road connections between NUTS3 regions.

The missing regional attribute data is more difficult to compensate for. It either has to be filled by imputation or from other years and adjusted to the base year.

Purchasing Power Parity is missing for ES630, ES640, MT001, MT002, PL330 (PL076), PL341 (PL077), PL431 (PL0C4), PL432 (PL0C5), PL511 (PL0C6), PL512 (PL0C7) (the zone in parentheses is the 1999 code) and is taken from the subsequent year and adjusted to the base year.

The number of BEDS is missing for 83 NUTS3 regions in CH, DE, LT, NL, PL, UKM, UKN. These are filled based on the number of beds at NUTS2, which is available, scaled by the ratios of the GDP. This assumes that the number of hotel beds is instrumental in producing local income.

Population is estimated from population densities for 103 NUTS3 regions in IT where it is missing in the 2001 files.

5.3 Constructing Final File of Location Attributes

The following variables at NUTS3 level were chosen for the trial dataset for joint modelling of mode and destination choice:

- Land Area
- Population
- Max Temperature Summer
- Min Temperature Winter
- Mean Precipitation Summer
- Mean Precipitation Winter
- Number of Hotel Beds
- Purchasing Power Parity (relative to EU25)
- Gross Domestic Product

6. Addition of Non-Chosen Destinations

Nine non-chosen destinations per journey are desired for each observed journey in order to build a decision set. The work steps first limit the set of non-chosen alternatives that will be used in the modelling by choosing them randomly based on a sampling criterion. These NUTS3 regions are then merged to the enriched file of Dateline plus non-chosen travel modes. Finally, the journey attributes are added to the non-chosen destinations.

6.1 A sample of 9 non-chosen destinations per journey

The 1355 NUTS 3 destination alternatives in the European region have to be sampled to establish an alternative set. Vrtic, et. al (2005) chooses alternative destinations using a sampling method based on straight line distance from the origin. The choice set of destinations is all the NUTS3 zones in Europe. An alternative method would be to calculate this distance based not on physical space, but on a Euclidean similarity of the alternatives, based on normalized values of common attributes. A Java program was written to perform this sampling, based on a “distance” variable which can be defined as desired. The results shown here are samples based on crowfly distance.

First, the complete set of NUTS3 regions is read in, as used in the ETIS rail model (IVT European Rail Model), which are saved in an “attribute” file from VISUM. These NUTS3 regions correspond to the NUTS version 6(7) used in the ETIS Base and Dateline projects. A program calculates the distances between the NUTS3 centroids in meters. This program also sorts and exports the file of distances between NUTS3 regions, as well as the relevant variables from the enriched Dateline file, to tab-delimited text files for calculations in Java. The Dateline file is used in case the modeller wants to use an alternative socioeconomic distance between socio-economic variables as a basis for selecting the sample of non-chosen destinations, instead of straight line distance.

The choice of alternative set is done in a Java program. It chooses 3 alternative destinations from the set closer to the origin zone than 70% of the distance to the chosen destination; 3 from 70%-130% of this distance; and 3 from farther than 130% away. There are cases where there are no closer or farther NUTS3 zones. In these cases, the algorithm seeks another destination in the neighboring distance class, until 9 non-chosen alternatives have been chosen. No destination may appear twice for a given journey.

6.2 Adding mode attributes to non-chosen destinations

The data sources for the mode-specific characteristics of travel to the non-chosen destinations is described in sections 3.2, 3.3, and 3.4. The extended version of this paper details the steps to associate these attributes with the destinations. The 9 alternative destinations are merged with the attributes of the rail, air, and road modes are then merged for each of the 9 destination alternatives.

6.3 Final Enrichment of Dateline with Destination and Mode Alternatives

The file of generalized cost attributes of travelling to the non-chosen destinationsis merged to the socio-economic and geographic characteristics of the destinations, and then to the Dateline dataset which already has the enrichments of mode alternatives to the chosen destinations. This file is then cleaned for use in the discrete choice modelling environment, BIOGEME. All data lines containing a “missing” data flag and all character strings have to be flagged or removed. Certain variables are re-coded into numbers. The destination attraction attributes are merged to the alternatives and saved in a format for BIOGEME.

7. Plan for Model Estimation

The enriched Dateline data contains personal, household, and journey information, and mode and destination attributes. Two kinds of models are attempted: mode choice alone and joint mode/destination choice.

Prior work into mode choice for long distance journeys has identified key observable traveller, household, journey, and destination characteristics that influence travel behavior. Limtanakool, Dijst, and Schwanen (2004) present a summary of long-distance travel modelling and discuss the influence of travel time and destination characteristics on mode choice with respect to the socioeconomic characteristics of the traveller. Increased density and diversity of land use types tends to discourage car use. Good access from the urban center to train stations increases the use of trains, particularly for business journeys. The most important characteristic for train use by vacationers is the presence/absence of a train station. Education and car availability, as in everyday traffic, are consistent influences on mode choice for all journey purposes. Including the variable travel time is crucial in estimating the model, provided the socioeconomic corrections are correct.

Hubert and Potier (2003) find in their study of national long-distance travel surveys that the number of long-distance journeys per year increases with income, with a saturation effect for high incomes. The effect of the size of the origin and destination towns is found to be a relevant factor in determining journey rates, for towns of certain size, as is the regional urbanization of area around the towns. The journey purpose confounds the explanatory power of urbanization on journey rate. People who make international journeys are described by other characteristics than people who make generic long-distance journeys. Journeys across borders might therefore be best treated by a geographic dummy (0,1) linked to destination so that the parameters of these individuals are estimated correctly. Macroeconomic trends, such as the productivity of industry and services, reduced working time, growth in international business relations, or people going into retirement, have large influence on travel rates on a macro scale.

Last, Manz, Zumkeller (2003) show that frequent journeys are strongly tied to characteristics that are associated in our society with success: middle age, high education, a job with a high income. The respondents need to be divided into groups of journey purpose however, in order to describe behavior such as mode or destination choice, or to identify which socioeconomic characteristics and which mode attributes are important. 90% of the highly mobile people either travel extensively on business and little privately, or vice-versa. The latter group (those who travel a lot, but much more for leisure than for business) is four times as large as the former. Meanwhile, only 10% of highly mobile people divide their journeys equally between work and non-work. This finding accents the importance of segmenting the respondents by the proportion of journey purposes that the person makes because sensitivities to supply characteristics like cost, time, and comfort will be correspondingly differentiated. Indeed, for travelers who make more work journeys than non-work journeys, the mode choice, which is distance-dependent, tends toward higher cost public modes with longer distance quicker than for people who travel more frequently for non-work purposes than for work purposes. The former group is averse to using more time than necessary in travel, and the latter prefers to save money.

The number of modes used by an individual in the year can also be used to classify the person as a “multi-“ or “mono-“modal personality, to help quantify the person’s awareness of other modes. The people who made the most journeys in their dataset als used the most diverse modes of transport.

Many of these idealized model formulations are possible with the available data. The analysis is begun with simple models corresponding to intuitive tradeoffs of the inconvenience of travel and socioeconomic variables. Slightly more complex, nonlinear interactions between these variables are then incorporated into the utility functions. More subtle models involving grouping or typing of the decision makers has not been pursued.

8. Mode Choice Model

Dateline has 84,126 person-journeys after geocoding. The mode-enriched Dateline dataset for the mode choice model has 46,185 lines of data. The loss is a result of incomplete data records in Dateline (missing household, person, or journey information), incomplete mode alternative data about the connections between the origin and destination municipalities, and the deletion of the points that lie outside the EU. Irrelevant Dateline variables were deleted and all variables were converted to numeric values and formats compatible with the estimation software. The set of variables is listed in Table 8.

Table 8 Variable list for file mode001.dat

J_DUR	Duration of the journey in number of nights
J_PPID	Main purpose of journey (see list of purposes)
J_DIST	longer distance of the J_DIST_O and J_DIST_R
J_DNUTS	NUTS3 code of chosen destination
HH_PERS	Total number of persons in household
HH_PC	Number of privately owned cars
HH_CC	Number of company cars
HH_PERSC	Number of persons per household in classes (1=1 / 2=2 / 3=3 / 4=4 / >4=5) (5 classes)
HH_TOTHJ	The number of holiday journeys in total of the person's household
HH_TOTBJ	The number of business journeys in total of the person's household
HH_TOTPJ	The number of other private journeys in total of the person's household
HH_G	Household level weight
P_AGE	Age of the person
P_AGEC	Age class of the person (-24=1 / 25-44=2 / 45-64=3 / 65- =4) (4 classes)
P_TOTHJ	The number of holiday journeys in total of the person
P_TOTBJ	The number of business journeys in total of the person
P_TOTPJ	The number of other private journeys in total of the person
P_DISH	The total distance of all holiday journeys performed by the person
P_DURH	The total duration of all holiday journeys performed by the person
P_DISP	The total distance of all other private journeys performed by the person
P_DURP	The total duration of all other private journeys performed by the person
P_DISB	The total distance of all business journeys performed by the person
P_DURB	The total duration of all business journeys performed by the person
P_G	Person level weight
RTT0	Road Travel Time (minutes)
TFZ0	Rail Ride Time: first station to last station (minutes)
TUH0	Number of Transfers Rail
TBH0	Service Frequency Rail
TZD0	Distance from Origin to Start Rail Station (km)
TAD0	Distance from End Station to Destination (km)
AIRZZ0	Time from Origin to Start Airport (minutes)
AIRAZ0	Time from End Airport to Destination (minutes)
ATT0	Air Travel Time (minutes)

AIRUH0	Number of Transfers Air
AIRBH0	Service Frequency this OD Air
PURPOSEP	Journey Purpose Personal (binary)
PURPOSEH	Journey Purpose Holiday (binary)
PURPOSEB	Journey Purpose Business (binary)
GENDER	Gender "F"=0
LICENSE	Driver's license "Y"=1
DISCOUNT	Reduced fare transit ticket "Y"=1;
PASS	Transit pass "Y"=1
WWW	Internet access "Y"=1
MODE	Road=10; Rail=20; Air=30

8.1 Descriptive Analysis of the Mode Choice Dataset

The descriptive analysis checks for correlations and qualitative trends in the data to help estimate the regressions. This analysis describes the first-order influence of mode attributes, journey characteristics, mobility tools, sociodemographics of people and households, and geography on mode choice.

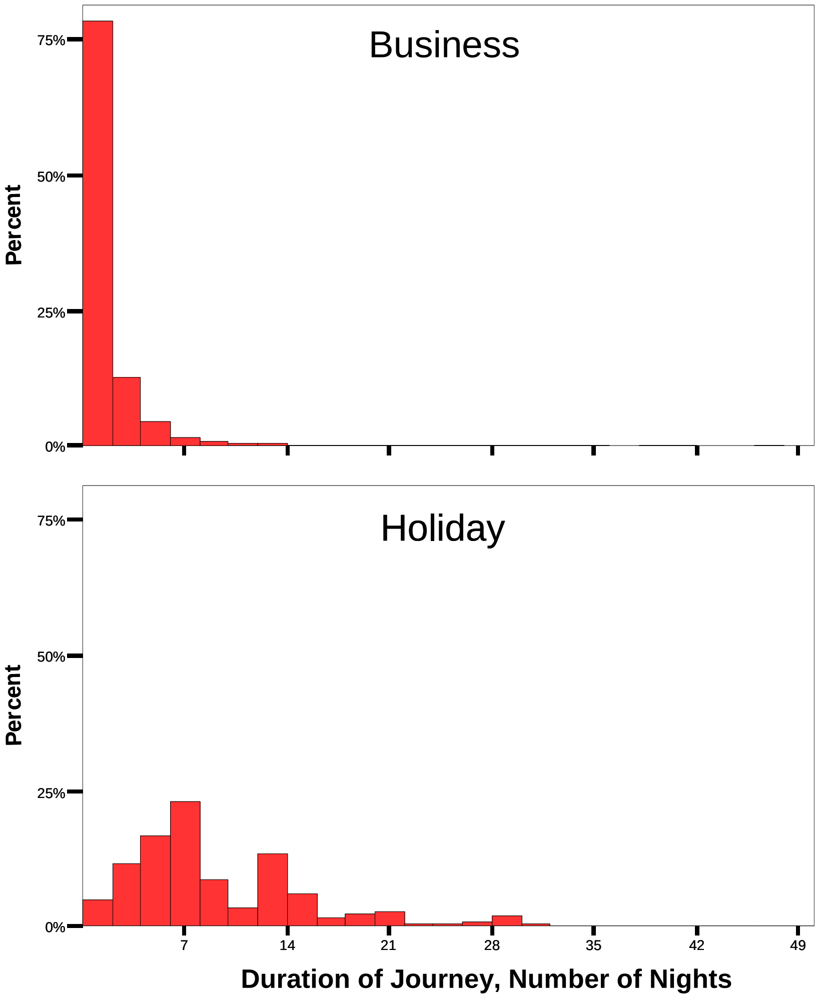
8.1.1 Travel Times and Transfers

Air travel times are bimodally distributed with peaks at 50 minutes and 150 minutes and a rapidly decaying right tail. The number of air transfers is highly positively correlated with the air travel time (0.88) and negatively correlated with air service frequency (-0.43), indicating that either or both of the latter variables should be left out of the regression. The road travel times have a very thick right tail and a peak at 250 minutes. The rail travel times have a peak at 200 minutes and a less thick right tail than the road travel times. The road and rail travel times correlate highly with distance (0.82, 0.83), probably due to a fairly uniform average speed on major service corridors. Rail travel time correlates with the number of train transfers. The latter is an indicator of the directness of the route and would be a valuable descriptor of qualitative service if its correlation does not degrade the model fit.

8.1.2 Journey Duration by Purpose

Holidays make 68% of journeys, business 25%, and personal journeys 7%. The business purpose (B) shows the tendency for single-day or overnight journeys, while holiday (H) travel is grouped into peaks every 7 days, showing multiples of weeks. An average vacation lasts 7.8 days, while an average business journey lasts 1.8 days.

Figure 6 Dateline Journey Duration (days) by Purpose



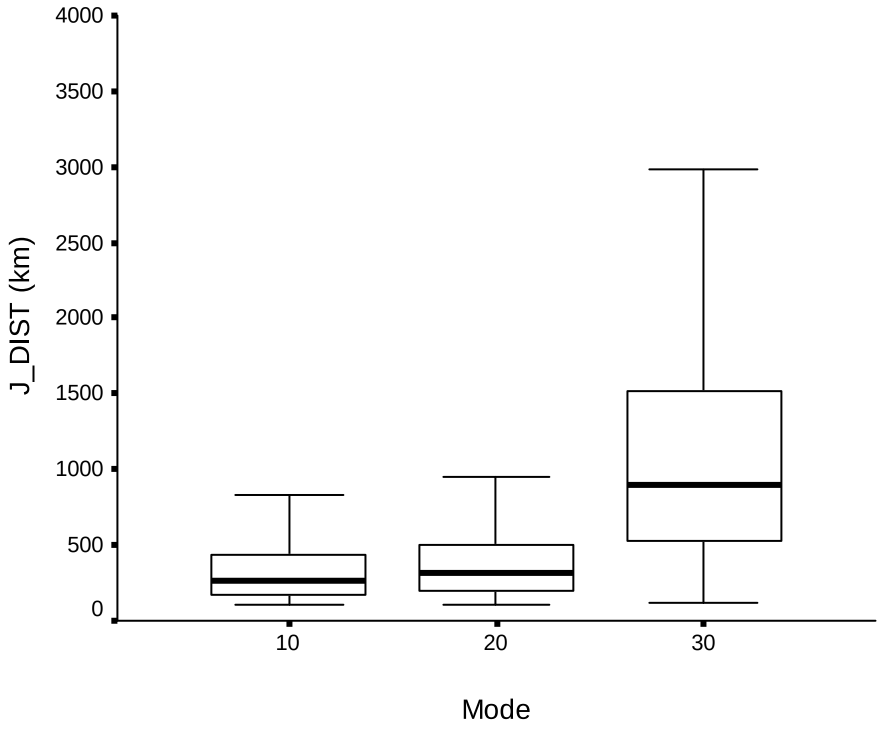
The distribution of journey duration by mode (not shown) has a thicker right tail for air than for the other modes, but the mean does not vary significantly across modes.

8.1.3 Journey Distance

The boxplot in Figure 7 shows that the air mode is clearly preferred for longer distance journeys. The cutoff at 500km where mode preference shifts from road (and rail) to air corroborates the observations made in the MYSTIC project that preceded Dateline (Dateline Deliverable 2).

Figure 7 Dateline One-Way Journey Distance (km) by Mode (10 Road, 20 Train, 30 Air).

Quartiles and extreme values, excluding outliers.



8.1.4 Mobility Tools

Mobility preferences with respect to mobility tools are evident in the dataset. Train travellers stand out with the highest proportion of transit discounts (26% compared to sample mean 11%), transit passes (29% compared to 23% sample mean), and a lower rate of licensed drivers (61% compared to 72% in the sample mean). Internet connectivity is highest for air travellers (57% versus 50% in the sample mean).

8.1.5 Sociodemographics

There is a parabolic distribution of the number of journeys with age for all modes, indicating that an age * age variable might have descriptive power. Men and women are equally represented in the dataset. Males make significantly more business journeys than females (2702/778). More men (59%) than women take the train, but there is no difference on other modes. The household size of persons using the air and road modes is nearly identically distributed. Train travel is undertaken by people from smaller households. The number of business or holiday journeys taken by the person in a year is an important determinant of journey generation. But this variable has a similar distribution for all modes and is therefore not expected to be a strong descriptor of mode choice.

8.1.6 Origin Characteristics

It might seem reasonable to expect centrally-located countries to have better train and road access than edge-located countries, and therefore slightly more preference for air travel in countries at the edge of Europe. The mode choice does not vary significantly across origin country, however. The effect of origin on mode choice at NUT3 level was not explicitly examined. At this finer level, it would be expected that attractivity attributes of the origin and destinations would be correlated with geographic location, which would confound analysis.

8.2 MNL for Mode Choice

Multinomial logit models with 3 choices (Road, Train, Air) were specified for a randomly chosen subset of 5000 person-journeys from the enhanced dataset and compared. Travel time is the service variable that enables comparison of utility across modes. Other variables which might be desirable in a conceptual model, but which correlate with travel time, therefore cannot be used if their introduction changes the other model coefficients; the number of air transfers, for example. Beginning with a simple model based on travel time, variables were added one at a time, in order to maintain control of the model, according to the indications in the descriptive statistics. The MNL was estimated using BIOGEME version 1.3.

The coefficients for the access and egress times to and from the airport, as well as air service frequency, are not significant but are retained in all models as important conceptual components of generalized cost. A quadratic function of AGE, as well as the SEX coefficient, were significant at 5% depending on other included variables, indicating correlations with other explanatory variables. The availability of the WWW did not yield a significant coefficient. Car ownership, household size, and owning a transit discount card obtained coefficients as expected from the descriptive analysis.

Between what is known about long-distance travel and the observation that coefficients of several variables changed depending on the inclusion of other variables, it is clear that the interaction of key explanatory variables must be taken into account. Two sets of interactions were identified which improve the model. The rho square increases slightly and the coefficient for SEX becomes significant, with the expected sign, when mode-specific travel time is scaled by relative distance (distance divided by mean distance) raised to an exponent (Mackie,

Wardman, Fowkes, Whelan, Nellthorp und Bates (2003)).

$\boldsymbol{\beta_{TT}}$* TT*(JDIST / 408.6)^$\boldsymbol{\beta_{JDIST}}$, where the average one-way journey distance is 408.6 km.. The signs of the exponent and the the travel time coefficients are negative, as expected.

The interaction between journey PURPOSE and SEX is based on the observation in the dataset that more men than women take business journeys, and that a higher percentage of business journeys are taken on the train and in the air than non-business journeys. The interaction takes the form:

$\boldsymbol{\beta}_{Business}$ * (1 + $\boldsymbol{\beta}_{SEX}$*SEX) * PURPOSEB.

The AGE + AGE * AGE coefficients to simulate the peak travel undertaken at middle age is just insignificant (t = 1.89, -1.86).

Several plausible models incorporating these variables and interactions, of roughly equivalent rho square measures, could be accepted as final, depending on what features of the system are to be emphasized. Two sample MNL specifications are presented in Table 9. Only the utility for ROAD mode was different in Model 2. The set of coefficients follows in Table 10.

Table 9 Dateline Utility functions for 2 MNL mode choice models based on the Enriched dataset

30 AIR Models 1 and 2	CA * one + HHSIZE * HH_PERS + DCARD * DISCOUNT + ATA * AIRZZ0 + ETA * AIRAZ0 + AIRD * J_DIST + FREQA * AIRBH0 + DURA * J_DUR + ((TTA * (PCTD ^ EPSILONA)) * ATT0) + ((PBUSA * (1 + (SEX * GENDER))) * PURPOSEB)
20 RAIL Models 1 and 2	CT * one + ATT * RAT + ETT * RET + TRSFRT * TUH0 + HHSIZE * HH_PERS + DCARD * DISCOUNT + FREQT * TBH0 + ((TTT * (PCTD ^ EPSILONT)) * TFZ0) + ((PBUST * (1 + (SEX * GENDER))) * PURPOSEB)
10 ROAD Model 1	CR * one + TTR * RTT0 + CARR * NCARS
10 ROAD Model 2	CR * one + CARR * NCARS + (TTR * (PCTD ^ EPSILONR)) * RTT0

where CR = 0, RAT and RET are the access and egress times at 40km/h average speed to and from the train stations, based on the distances TAD0 and TZD0 in Table 8, and NCARS =

HH_TC + HH_PC, the total number of cars available to the household, and journey duration (number of nights) and distance are specific to the AIR mode.

The coefficients and fit indicators, including non-significant coefficients, follow in Table 10.

Table 10 Coefficient Values for MNL in Table 9

Parameter Name	Description	Model 1	Model 2
AIRD	Distance (air-mode specific) (km)	2.48e-03	3.20e-03
ATA	Time origin to airport (min)	-8.95e-04 *	-1.24e-03 *
ATT	Time origin to train station (min)	-3.21e-03	-3.58e-03
CA	Air-modeconstant	-2.98e+00	-3.52e+00
CARR	Car ownership (road-mode specific)	4.16e-01	4.16e-01
CR	Road-mode constant	0.00e+00	0.00e+00
CT	Train-mode constant	-1.41e+00	-1.73e+00
DCARD	Transit discount card dummy	1.07e+00	1.09e+00
DURA	Duration (air-mode specific) (N nights)	-5.72e-02	-5.79e-02
EPSILONA	Distance scale factor (air mode)	-1.32e+00	-1.28e+00
EPSILONR	Distance scale factor (road mode)		-5.05e-01
EPSILONT	Distance scale factor (train mode)	-1.00e+00	-1.47e+00
ETA	Time airport to destination (min)	2.43e-04 *	5.79e-05 *
ETT	Time train station to destination (min)	-3.94e-03 *	-4.11e-03 *
FREQA	Service frequency air mode (/day)	9.99e-03	8.16e-03
FREQT	Service frequency train mode (/day)	7.25e-02	7.68e-02
HHSIZE	Household size	-8.53e-02	-8.25e-02
PBUSA	Interaction term purpose-sex (air mode)	2.13e+00	2.18e+00
PBUST	Interaction term purpose-sex (train mode)	1.39e+00	1.45e+00
SEX	Sex dummy	-2.93e-01	-2.92e-01
TRSFRT	Number of transfers train mode	-8.93e-02	-6.63e-02 *
TTA	In-vehicle travel time air (min)	-3.94e-03	-3.92e-03
TTR	In-vehicle travel time road (min)	-8.53e-04	-1.19e-03
TTT	In-vehicle travel time train (min)	-1.22e-03	-9.82e-04
Sample size:		5014	5014
Null log-likelihood:		-5508	-5508
Final log-likelihood:		-2270	-2251
Adjusted rho square		0.5838	0.5872

* Parameters are not significant at 5% and are kept in the models for comparison across modes and models and for conceptual consistency.

8.3 Discussion of Mode Choice Model

The access and egress times for train stations and the egress time from the 3 nearest airports are not significant in either model. As noted earlier, it is likely that this is correct and that access time is not the important variable determining the origin or destination train station or airport. Unaccounted for variables like ticket price, airlines serving the airport, opportunity for nearby shopping, and available parking, as well as accounted-for variables like frequency of service at the station or airport, probably have more influence on the choice (Hess 2005). Thus missing variables inhibit the estimation of a significant value of time parameter for the access legs of the journey. A trip-level analysis of the journey stages could help determine realistic decision variables for airport choice in Dateline.

All coefficient signs are as expected, and, aside from the indicated coefficients, all t-statistics in both models show strong evidence that the parameters are not equal to zero and are useful in describing mode choice. The two models yield very similar parameters and similar ρ^2 (improvements of fit relative to all parameters = 0), despite the linear versus nonlinear treatment of road travel time. The change in the significance of train transfers and the value of the train travel time coefficient with the addition of the nonlinear correction to road travel times indicates that the correlations in these variables (train transfers, train travel time, and road travel time) are confounding the parameter estimation. This reflects the similarity of the included basic level of service variables between train and road modes over long distances. Decoupling the two is difficult without additional alternative-specific variables like parking cost or ticket prices, or activity-specific information like whether the activity is tied to the automobile (e.g. auto touring or camping). The coefficients of the distance correction and of travel time for the air mode in model 1 are not influenced by the change to the road utility made in model 2 because the attributes of and journeys taken with the road and air modes are fundamentally different (less correlated). Both models are presented to illustrate that it is not clear from the statistics how to handle the interdependence of road and rail mode attributes, but that several models can be useful, depending on the questions posed about the travel behavior.

Journey distance enters the models as an air-specific variable. The positive sign is consistent with the observation that longer distance journeys are taken with the air mode. Journey duration is also air-specific because it was not significant for road and train. Its negative sign indicates a prefence to take air for shorter duration journeys, perhaps an indication of willingness to pay for the higher level of service for important meetings. The two variables together may be a way to capture air ticket prices (given an OD relation or journey distance, longer stays are generally a cheaper rate than shorter stays, certainly in the middle range of 2 days to 2 weeks). It might be worthwhile to attempt a specification with an interaction between these two variables.

The journey purpose enters the model specific to mode in an interaction with the gender of the person. This is intended to capture the observed tendency for business travellers to be men, and to differentiate between the air mode and the train mode for business travel. The values of PBUSA and PBUST, the business purpose coefficient for air and train, respectively, both reflect the observed higher preference for air and train modes (relative to road) for long-distance business travel versus non-business travel. Their value is significantly different, confirming the importance of mode-specific treatment. The negative coefficient for gender captures the tendency for men (SEX=1) to take more business journeys.

The distance/time interaction results in negative EPSILON coefficients which amount to a correction of travel time with the speed of conveyance. If a longer distance than average is covered, meaning the mode is faster than average, then the travel time is weighted less in the regression. If progress is slower, the time is weighted more. In other words, the traveller is willing to sit longer in order to get a little farther, if progress is evident. In this way, higher-value services are not penalized by the longer distances that travellers use them for, despite the long travel times. This distance devaluation for the road mode in Model 2 is one-third to half as strong as for air and train and indicates that long road journeys “feel” relatively longer than long train or air journeys.

9. Joint Mode/Destination Choice Model

The Dateline dataset with chosen alternatives and mode attributes used has 46,185 observations (see section 8). As a result of missing data associated with the non-chosen destinations, compounded by associating 9 of these to each observed destination choice, the loss of data for the joint mode-destination dataset was large. For this analysis, only the observations were used for which all attributes of the enriched destination and mode choice set were valid. The availability flags were not implemented. The entire dataset of 6326 lines is used to estimate the joint mode/destination model.

9.1 Variables

The dataset has the regional attractiveness variables for chosen and non-chosen destinations listed in section 5.3 in addition to the socio-demographic characteristics and mode attributes listed in Table 8.

9.2 Descriptive Analysis

The distributions of the location attractiveness variables in the synthetic dataset appear to be consistent (this was not formally tested).

The annual minimum and maximum temperatures of NUTS3 regions have bell-shaped curves which could be qualitatively described as approximately normal or negative (winter)- or positive(summer)-skewed normal distributions. The summer precipitation distributions are bimodal with obvious separation of dry and wet regions. Winter precipitation approximates a positive-skewed normal, Weibull, or similar distribution. Purchasing power parity is scaled to the normal value of the EU25 and is nearly normal. GDP is a strongly skewed log-normal type distribution, a shape which also describes the number of hotel beds, the land area, and the population.

There are few obvious differences in the distributions by mode in the attractivity characteristics of the non-chosen destinations. However, the distributions of geographic and economic characteristics associated with chosen destinations reached by the air mode tend to differ from the other two modes, indicating that users of this mode can be more selective of their destination. The populations reached by the air mode have a much thicker right tail, as do the number of beds in the region and the regional GDP. The summer and winter temperature distributions have lower variance for these destinations and they have a stronger tendency to be warmer and wetter than destinations reached by the other modes.

The mode attributes for non-chosen destinations maintain the main statistical characteristics as for the chosen destinations, with differences as a result of the inhomogeneous distances to the non-chosen destinations and the strong dependence on distance of travel time and other level-of-service variables. The series of figures, Figure 8, Figure 9, and Figure 10 illustrate the varying distributions of travel time for the rail mode, for example.

Figure 8 Dateline Train Travel Time to Chosen Destination (min) by Mode (10 Road, 20 Train, 30 Air).

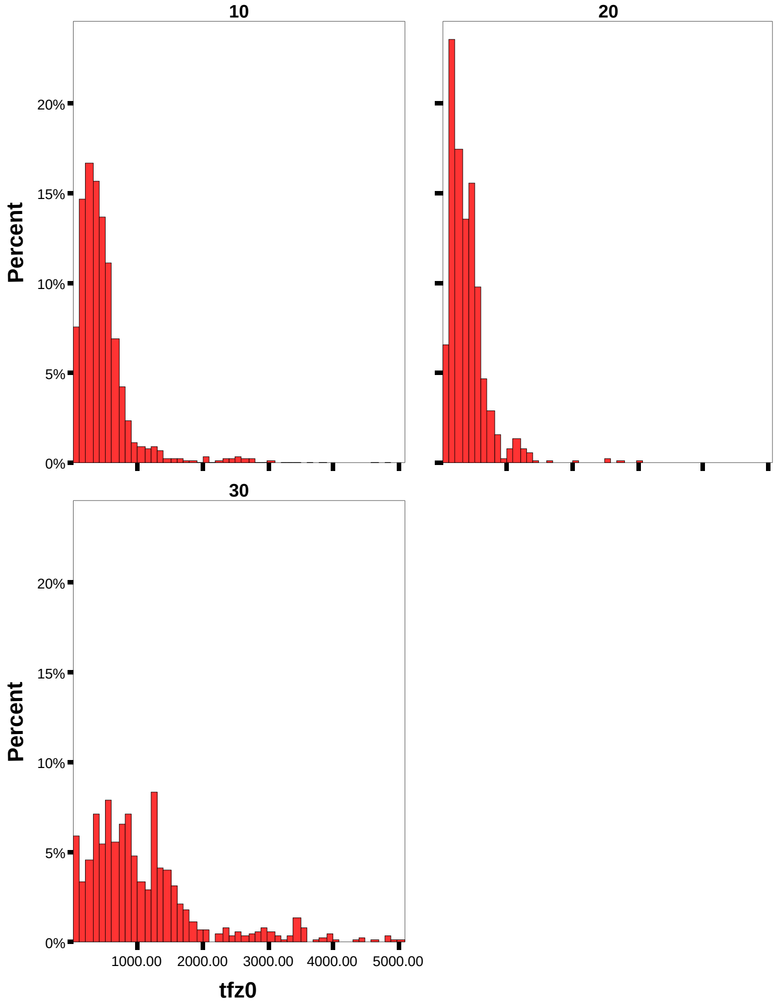


Figure 9 Dateline Train Travel Time to First Non-Chosen Destination (min) by Mode (10 Road, 20 Train, 30 Air).

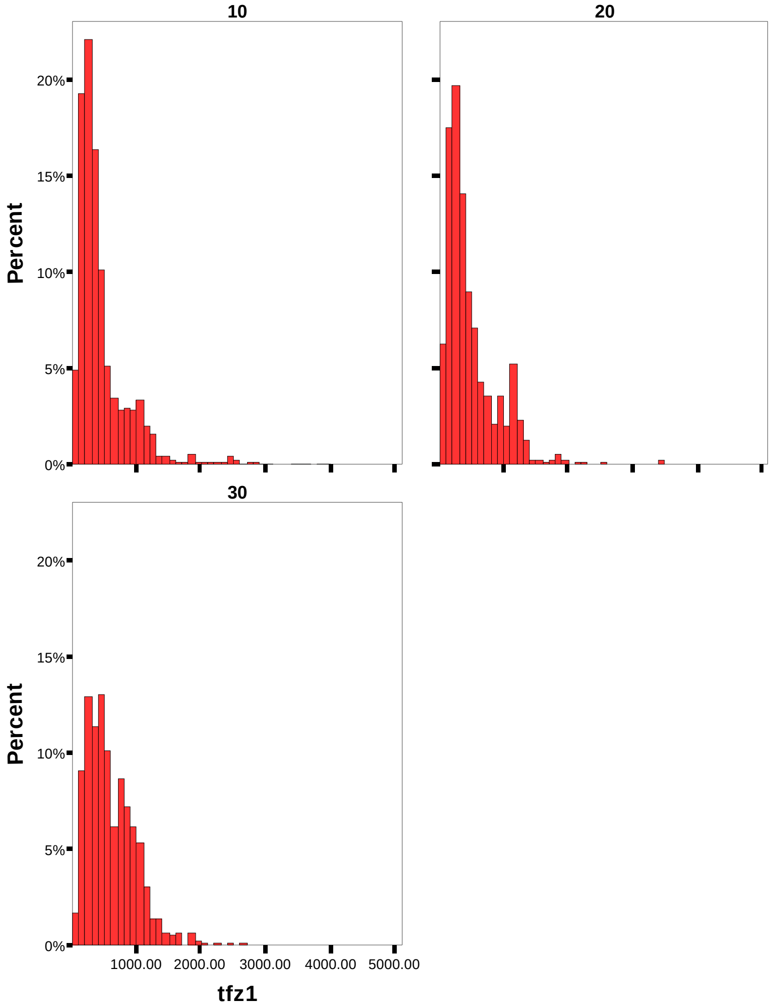
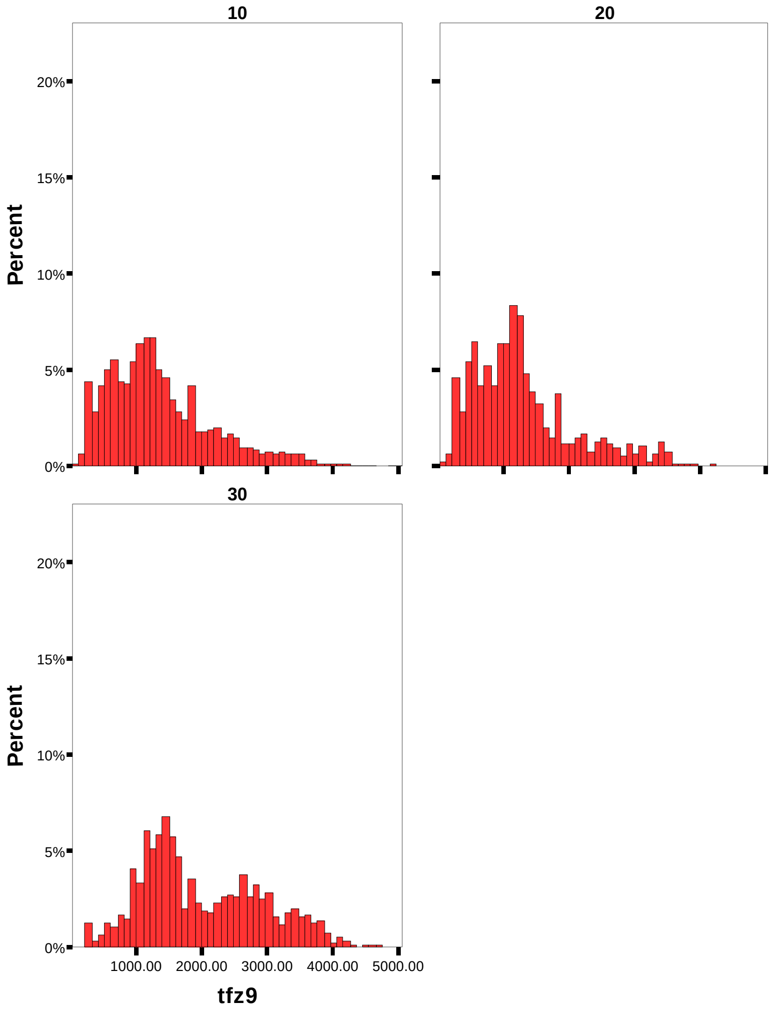


Figure 10 Dateline Train Travel Time to Ninth Non-Chosen Destination (min) by Mode (10 Road, 20 Train, 30 Air).



Qualitatively comparing the distributions, the artificially generated set appears to have a realistic distribution of rail travel times, compared to the observations.

9.3 Fitting the Joint Mode/Destination Model

There is one utility equation for each combination of mode and destination, or 30 equations in all. The non-chosen destinations are numbered 1-9 chosen destination is numbered "0" for each mode, and the modes are numbered 1-3. The combined mode-destination alternative is numbered with mode in the tens digit and destination in the ones digit. Thus, alternative "32" means "AIR" mode, "second" non-chosen destination.

The regression equations from Table 9 using alternative-specific travel times and 30 alternatives, result in positive coefficients for travel time and distance. This would indicate a desire to sit in a vehicle for the sake of doing so, clearly not the economically expected result (Hess, et. al 2004). The lack of prices in the dataset may be causing this inconsistency.

9.4 Final Mode/Destination Model

To constrain the sensitivity of travellers to travel time equally across alternatives, the MNL described here uses a generic travel time variable and a mode-specific distance variable. The set of economic indicators for the destinations is used, but the geographic indicators have not been included in the model, yet. Alternative 10, the ROAD mode to the chosen destination 0, is the basis of comparison for generic variables. Table 11 shows the equations for the chosen mode and one non-chosen mode (all others are identical to the utility for the chosen mode, except that the destination subscript corresponds to another destination).

Table 11 Utility functions for MNL destination and mode choice model based on the Enriched Dateline dataset

30 AIR0	CA * one + HHSIZE * HH_PERS + DCARD * DISCOUNT + ATA * AIRZZ0 + ETA * AIRAZ0 + FREQA * AIRBH0 + DURB * J_DURB + GDP * G0 + TT * ATT0 + PPP * P0 + POP * LPOP0 + BEDS * LBEDS0 + DURN * J_DURN
20 RAIL0	CT * one + ATT * RAT + ETT * RET + TRSFRT * TUH0 + HHSIZE * HH_PERS + DCARD * DISCOUNT + FREQT * TBH0 + GDP * G0 + TT * TFZ0 + PPP * P0 + POP * LPOP0 + BEDS * LBEDS0 + DISTT * J_DIST + DURB * J_DURB + DURN * J_DURN
10 ROAD0	CR * one + CARR * NCARS + GDP * G0 + TT * RTT0 + PPP * P0 + POP * LPOP0 + BEDS * LBEDS0 + DISTR * J_DIST
11 ROAD1	CR * one + CARR * NCARS + GDP * G1 + TT * RTT1 + PPP * P1 + POP * LPOP1 + BEDS * LBEDS1 + DISTR * ALTDIST1
...	

In this model, CR=0 and other variables are defined as in Table 9. G0 is LN(GDP0) * PURPOSEB and LBEDS0 is LN(BEDS0) to represent quantities which are additive across regions. Purchasing power parity (P0 * (1-PURPOSEB)) is already normalized so it is not necessary to take the logarithm. DURN and DURB are the purpose-specific journey durations relative to the road mode. Note that this model uses mode-specific distance coefficients to simulate the distance-dependent attributes, apart from travel time, of being in the vehicle. This is intended as a proxy to out of pocket cost, for example the cost of gasoline to drive a certain distance.

Table 12 Coefficient Values for MNL in Table 11

Parameter Name	Description	Value	t-test
ATA	Time origin to airport (min)	-1.99e-02	-14.20
ATT	Time origin to train station (min)	2.10e-03	2.56
BEDS	Thousands of hotel beds	7.86e-01	37.00
CA	Air mode constant	-3.57e-01	-2.83
CARR	Car ownership (road-mode specific)	5.44e-01	11.60
CR	Road-mode constant	0.00e+00	Fixed
CT	Train-mode constant	-8.42e-01	-6.46
DCARD	Transit discount card dummy	1.04e+00	10.10
DISTR	Distance (road-mode specific) (km)	-3.10e-03	-24.50
DISTT	Distance (train-mode specific) (km)	-8.36e-04	-6.86
DURB	Duration Business Journey (N nights)	-6.23e-01	-12.70
DURN	Duration Non-Business Journey (N nights)	-3.81e-01	-61.80
ETA	Time airport to destination (min)	-7.88e-04	-2.29
ETT	Time train station to destination (min)	9.46e-03	14.80
FREQA	Service frequency air mode (/day)	5.30e-02	6.37
FREQT	Service frequency train mode (/day)	3.62e-02	5.03
GDP	GDP million EUR	4.34e-01	6.34
HHSIZE	Household size	-1.54e-01	-5.45
POP	Population, 1000 people	1.53e-01	5.82
PPP	Purchasing power parity indexed to EU25	-1.18e-03	-2.68
TRSFRT	Number of transfers train mode	-1.13e-01	-3.59
TT	In-vehicle travel time (min)	-8.18e-04	-14.50
Sample size:		6326	
Null log-likelihood:		-21516	
Final log-likelihood:		-10100	
Adjusted rho square		0.5296	

9.5 Discussion of Destination and Mode Choice Model

The coefficients are all significantly different from zero and the improvement over a model with all coefficients = 0 (ρ^2) is high at 0.53, indicating significant explanatory capability for joint mode and destination choice.

The mode-specific coefficients have the correct sign. ATA/ETA (<0) and ATT/ETT (>0) show that travellers prefer closer airports and farther train stations, perhaps because of better connections available at larger main stations. Public modes with higher frequency are preferred. The number of train transfers has a significant coefficient in this model and indicates disutility. The number of cars and ownership of a discount card have the correct mode-specific signs, both encouraging use of the mode they serve. Household size has a negative coefficient as in the mode choice model, for the same reason, because larger groups tend to use a car and smaller groups tend to fly or take the train. The generic travel time coefficient is negative for this unweighted variable, and the coefficient of linear distance is negative for road and train (it was negative but not significant for the air mode). The duration for business and non-business journeys is negative relative to the road mode/chosen destination alternative.

The destination characteristics GDP, BEDS, and POP all indicate positive influence on trip-making. GDP represents the business activity or opportunity of a region and was expected to be a strong indicator for destination choice for business journeys. Population increases the pool of individuals one could productively meet with and should increase both business and non-business attractiveness. The stock of hotel beds indicates the development of tourist infrastructure and interregional business relationships. Purchasing power parity was included as an indicator of attractiveness for tourism, because the choice of vacation spot often has to do with local prices. Its coefficient is negative, showing that high prices indeed discourage choice of the destination.

9.6 Improving the Model

The interactions between journey purpose/attractiveness variables, such as purchasing power parity and the holiday purpose, or GDP and business, were achieved by defining new variables. The resulting coefficients are strongly significant. Interactions between other variables were not attempted in this model, nor were nonlinear relationships.

The inclusion of climate data would enhance the destination choice model, based on the findings of the descriptive analysis of the dataset, which suggests descriptive power in destination temperature, especially for the air mode. The interaction of the travel month with the climate data, or the origin climate/NUTS3 zone and the destination climate would be important. This information is available in Dateline but would have to be reintroduced into the final dataset that was used for modelling. Whether an origin or destination region containing the capitol city has significance, and whether trips cross formal borders or language boundaries are also testable hypotheses that should be investigated with this data.

Tests for violations of the IIA property have not been performed to evaluate the appropriateness another discrete choice form allowing correlated errors (for example Eymann and Ronning 1997). The complete evaluation of the discrete choice model is a primary priority.

The number of cases (person-journeys) could be improved for the mode-destination modelling by utilizing the availability flags in BIOGEME.

10. Conclusions and Future Work

An attempt should be made to obtain or model the prices of the modes. A model for rail is available. Road prices and out of pocket cost for automobile travel will be available soon from the ETIS Base project. Airline prices may also be available from ETIS Base.

Finally, for route choice modelling, the work must be repeated with the “trips” file from Dateline instead of the “journeys” file.

11. References

Axhausen, K. W.and M. Frick (2004) Nutzungen, Strukturen, Verkehr, *Arbeitsberichte Verkehr- und Raumplanung*, **205**, Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, Zürich.

Ben-Akiva, M.E. und S.R. Lerman (1985*) Discrete Choice Analysis*, MIT Press, Cambridge.

Bleisch, A. and Ph. Fröhlich (2003) Die Erreichbarkeit von Regionen, IBC Modul Erreichbarkeit Phase 1, Schlussbericht, BAK Basel Economics, Basel.

BOKU-ITS (2003) DATELINE Deliverable 10b Weighting and Grossing up Report, Socialdata, Munich.

Cook's Timetable September, 2002.

DATELINE consortium homepage (March, 2003): http://www.ncl.ac.uk/dateline/home_page.htm, downloaded March, 2006.

ETIS Base D6 (2005) Annex report WP 7: ETIS-Database methodology development and database user manual – passenger transport supply. European Union.

EUROSTAT homepage (2006): http://epp.eurostat.cec.eu.int/

Eymann, A. and G. Ronning (1997). Microeconometric models of tourists' destination choice, Regional Science and Urban Economics 27 (1997) 735-761.

Hess, S. (2005) Analysing air-travel choice behaviour in the greater london area, Paper presented at the 45th Congress of the European Regional Science Association, August 2005, Amsterdam.

Hess, S., M. Bierlaire, J. W. Polak (2004) Estimation of value of travel-time savings using Mixed Logit models, Working Paper, Centre for Transportation STudies, Imperial College London.

Hubert, J. P. and F. Potier (2003) What is known? In K. W. Axhausen, J.-L. Madre, J. W. Polak and Ph. L. Toint (eds.) *Capturing Long Distance Travel*, 45-70, Research Studies Press, Baldock.

Institute for Transport Studies, University of Leeds und John Bates Services, Leeds.

Koppleman, F. S. and V. Sethi (2005) Incorporating variance and covariance heterogenetiy in the Generalized Nested Logit model: an application to modeling long distance travel choice behavior, Transportation Research Part B 39 (2005) 825-853.

Kroes, E., Lierens, A. Kouwenhoven, M. (2005) The airport Network and Catchment area Competition Model: A comprehensive airport demand forecasting system using a partially observed database, Paper presented at the 45th Congress of the European Regional Science Association, August 2005, Amsterdam.

Last, J, W. Manz and D. Zumkeller (2003) Heterogenität im Fernverkehr – wie wenige Reisen wie viel? *Internationales Verkehrswesen*; **55,** Heft 6, Deutscher Verkehrsverlag, Hamburg, Germany.

Last, J. and W. Manz (2003) Unselected mode alternatives: What drives modal choice in long-distance passenger transport? Conference Paper, 10th International Conference on Travel Behavior Research, Lucerne, Switzerland.

Last, J., W. Manz, B. Chlond, D. Zumkeller (2004) Eckwerte des Personenfernverkehrs in Deutschland. *Internationales Verkehrswesen***,** **56**, Heft 10, Deutscher Verkehrsverlag, Hamburg, Germany.

Limtanakool, N., M. Dijst, and T. Schwanen (2004) The Influence of Socio-Economic Characteristics, Land Use and Travel Time Considerations on Mode Choice for Long-Distance Trips, 83rd Annual Meeting of the Transportation Research Board, January, 2004, Washington, D.C.

Mackie, P.J., M. Wardman, A.S. Fowkes, G. Whelan, J. Nellthorp und J.J. Bates (2003),

Neumann, Alex (2003) Korrekturverfahren für Stichproben von Verkehrsverhaltenserhebungen des Personenfernreiseverkehrs, Doktorarbeit, Institut für Verkehrswesen, Universität für Bodenkultur Wien.

Peter Davidson Consultancy (2000), Methodology for Statistical Analyses, Modelling and Data Collection (MYSTIC) http://www.cordis.lu/transport/src/mystic.htm.

Peter Davidson Consultancy (2003) DATELINE Deliverable 11 O-D Matrices, Socialdata, Munich.

PTV (2000): Benutzerhandbuch VISUM 7.5. Planung Transport Verkehr AG, Karlsruhe.

PTV (2004) Europäisches Basismodell Strasse, Planung Transport Verkehr AG, Karlsruhe.

Schürmann, Carsten (2001): Networks.txt. Institut für Raumplanung, Universität Dortumund.

Swiss Federal Statistical Office (2002) Methodologies and principal results, *Travel Behaviour of People Living in Switzerland in 1998*, SFSO, Neuchatel.

TIS.pt + UNEW (2003) DATELINE Deliverable 7 Data Analysis and Macro Results, Socialdata, Munich.

Vrtic, M., P. Fröhlich, N. Schüssler, K.W. Axhausen, S. Dasen, S. Erne, B. Singer, D. Lohse, C. Schiller (2005), Erzeugung neuer Quell-/Zielmatrizen im Personenverkehr, Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation / Bundesamt für Raumentwicklung, Bundesamt für Strassen und Bundesamt für Verkehr, May 2005.