ETH zürich

ETH Library

Purpose imputation for long-distance tours without personal information

Working Paper

Author(s):
Janzen, Maxim; Vanhoof, Maarten; Axhausen, Kay W.

Publication date:
2016-08

Permanent link:
https://doi.org/10.3929/ethz-b-000118790

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
Arbeitsberichte Verkehrs- und Raumplanung 1181

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

Purpose Imputation for Long-Distance Tours without Personal Information

Date of submission: 2016-08-1

Maxim Janzen
IVT, ETH Zurich, CH-8093 Zurich
phone: +41-44-633 33 40
fax: +41-44-633 10 57
maxim.janzen@ivt.baug.ethz.ch

Maarten Vanhoof
Laboratoire Sociology and Economics of Networks and Services, Orange Labs,
Issy-les-Moulineaux
phone: +33 1 45 29 60 78
maarten.vanhoof@orange.com

Kay W. Axhausen
IVT, ETH Zurich, CH-8093 Zurich
phone: +41-44-633 39 43
fax: +41-44-633 10 57
axhausen@ivt.baug.ethz.ch

Words: 5290 words + 3 figures + 5 tables = 7290 word equivalents

ABSTRACT

Analysis of long-distance travel demand has become more relevant in recent times. The reason is the growing share of traffic produced by journeys to remote activities, which are not part of daily life. In today's mobile world, these journeys are responsible for almost 50 percent of overall traffic. Traditionally, surveys have been used to gather data needed for the analysis of travel demand. Due to the high response burden and memory issues, respondents are known to underreport the number of journeys. Thus, alternative data sources are becoming more important. These sources collect the data passively, e.g. using GPS or GSM networks. The limitation of passively collected data is the lack of semantic information, especially trip purposes. Additionally, socio-demographic information is also missing making it difficult to impute the purpose. This paper shows how one can predict the tour purpose without the need of socio-demographic information. The solution extends the well known random forest approach. Attributes of the tours are used in order to overcome the lack of personal information. The training set for the algorithm was taken from a national travel survey.

INTRODUCTION

Investigation of long-distance travel behavior has become more important in recent years since the contribution of long-distance journeys to overall traffic is growing continuously. Therefore, its influence on planners of urban areas, highways, railroads etc. is becoming greater. Long-distance travel is usually defined by trips, which take place outside of a person's daily environment. In order to develop tools, which are able to provide reliable predictions, one needs data sources describing the current state of long-distance travel demand.

Data collection methods in the field of travel demand research were investigated in the past *(1, 2)*. The most frequently used data sources are surveys. In case of long-distance travel the number of these is limited (the main sources are national travel surveys). However, all long-distance travel surveys are facing similar problems. On one hand, due to the high response burden, these surveys have a low number of respondents. On the other hand, it is known that the number of journeys reported in surveys is too low *(3, 2)*. Both facts limit the explanatory power of studies and leave the question for the quality of the results unanswered *(4)*. To overcome these limitations alternative data sources are needed. Due to the high response burden, passive data collection methods are preferred, e.g. GPS or GSM data. Passive data collection, however, has an important drawback. The purpose of the tours is not part of the data. Additionally, the imputation of the purpose is difficult, because socio-demographic information is generally not available.

Nevertheless, knowledge of the tour purpose is crucial for the analysis of long-distance travel. In other words, there is a need for purpose imputation from passively collected data on long-distance travel. In this paper we focus on the case where personal information of the respondent is missing from the data, which is often the case in recent data sources. The main example of such data sources is mobile phone billing data as described by *(5)* or *(6)*. In this paper, we will elaborate an approach for purpose imputation based on a modified random forest algorithm. We validate the workings of our approach by applying it on a mobile phone billing data set.

RELEVANT LITERATURE

Data collection has always been an important issue in the field of travel demand research. Different methods of data collection were investigated in the past *(1, 2)*. The data sources mostly used are surveys, which have various forms *(7)* to suit the diverse requirements of the researchers.

In case of long-distance travel the number of surveys is limited. Focusing on Europe, the *Mobidrive* studies are available *(8–10)*. These studies focus on a six-week period, which is usually not sufficient for a deep analysis of long-distance travel behavior. Other sources are national travel surveys like the French *(11)*, British *(12)* or the Austrian *(13)*.

Due to the high response burden that is usually associated with long-distance surveys *(14, 15)*, it can be expected that the number of long-distance trips is usually underreported. Underreporting might be due to frequent travellers not responding, as well as travellers claiming not to travel, while answering other questions, so called soft-refusers *(3)*. Furthermore, there is a memory effect. Respondents tend to forget tours, which happened some time before the survey *(16–18)*. Additionally, the vehicle miles travelled are usually heavily underestimated as shown by *(19)*.

Consequently, there is a need of alternative data sources. Mobile phone network operators

produce mobile phone billing data that provides an enormous amount of data and has been already utilized in various fields *(20)* including the field of transportation *(21, 6, 22–24)*. Nevertheless, mobile phone billing data does not provide any semantic information, i.e. trip purpose or travel mode. Since this information is crucial for travel behavior analysis, imputation techniques have to be applied.

Several studies focussed on trip purpose imputation in the past. Firstly, it has been already an issue in traditional surveys *(25, 26)*. Secondly, the emerged GPS data source is usually in need of purpose imputation. Thus, several approaches have been presented to cover purpose imputation for GPS based data *(27–31)*.

Mobile phone billing data differs from the other data sources in terms of data richness. There is no information about socio-demographics. Additionally, the destinations of the identified trips are estimated at the resolution of cell towers. Consequently, it is impossible to identify the visited facilities and inherit the purpose from this information. Purpose identification techniques for mobile phone data are rule-based approaches and are limited to specific purposes, e.g. to tourism activities *(21, 6)* or work and home activities *(32)*.

METHODOLOGY

This work is motivated by mobile phone billing data that has become available recently to researchers and has been used to quantify long-distance tours. Thus, the methodology is based on assumptions that apply to this data source. These assumptions are the following:

1. The cleaned data provides home-based tours, i.e. a trip chain where the start point and the end point is the home location.
2. No socio-demographic information is available for the travelling persons (other than the home municipality).

Therefore, purpose imputation needs to use tour attributes rather than personal attributes. These attributes include tour destination or tour duration. Hence, the set of attributes is limited. We will show later which travel behavior attributes can be added to this set.

It has been shown that machine learning algorithms are reliable purpose predicters in case of GPS based data *(30, 31)*. Thus, this class of prediction algorithms is also chosen for mobile phone billing data. We adapted the random forest approach that has been already used in other cases of purpose imputation *(33, 34)*.

Random Forest

The random forest technique *(35)* belongs to the class of decision tree algorithms. It implements a set of rules that are learned by a machine and executed in a given order. It was already used successfully for trip purpose identification *(e.g. 33, 34)*. There are two advantages to the random forest approach in comparison to other decision tree methods. Firstly, random forests do not over-fit even if more trees are added *(35)*. Secondly, good results can be maintained even with missing data since they are estimated internally *(36)*.

The functionality of random forests can be described as follows *(36)*: A random forests grows many classification trees. To classify a new object (in our case: classify the purpose) from a vector of attributes, the vector is used for classification by each of the trees in the forest. Each tree gives a classification and thus a vote for a class. Eventually, the forest chooses the class having the most votes (over all the trees in the forest).

If N is the size of the full training set and M the number of attributes, each tree itself is grown as follows:

1. Randomly sample around $\frac{2}{3}N$ cases from the original data (with replacement). This sample will be the training set for growing the tree.
2. At each node $m \ll M$ variables are selected at random and the best split on these m is used to split the node. The value of m is a parameter and is held constant during the forest grow.
3. Each tree is grown to the largest extent possible. There is no pruning.

A pre-defined number of decision trees is grown and each one has a vote for a class.

Multi-Stage Random Forest

The traditional random forest approach has a disadvantage. It uses all available attributes to identify all classes, which are tour purposes in our case. Long-distance tour purposes are very diverse. Therefore, it is not advisable to use the classic random forest approach, because not all available tour attributes contribute to the classification of a tour purpose. For example, the month of the tour does not influence the commuting purpose class. As a solution we generate for each purpose class an individual random forest, where we just specify whether or not a tour belongs to the purpose class. An example of a decision tree that is part of a random forest identifying holiday tours is shown in Figure 1 (a). The idea to create specific random forests for each class has been investigated using a hierarchical approach *(37)*. A hierarchical approach is not applicable here since the purpose classes do not have a hierarchical structure.

Thus, a series of random forests has to be created for a classification of all tours. The length of the series is exactly the number of classes minus one. An illustration of such a series with five classes is shown in Figure 1 (b). The random forests are shown as red rectangles, while the purpose classes are the green ellipses. The five long-distance tour purposes shown in this example are the ones usually used and are also used in the application that is presented in the next section.

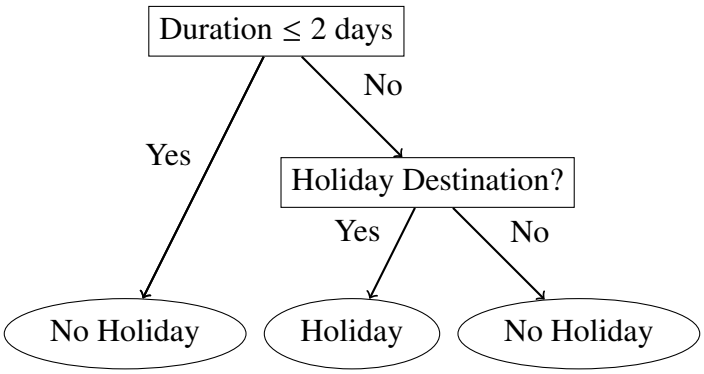
The multi-stage random forest as presented here maps each tour to a definite purpose class. Since there is also a random effect in this approach, we suggest to create and run the multi-stage random forest multiple times. Two advantages are gained by multiple runs. Firstly, the probability for misclassification is reduced. Secondly, one can identify those tours, which are not clearly classified, and treat them separately.

APPLICATION

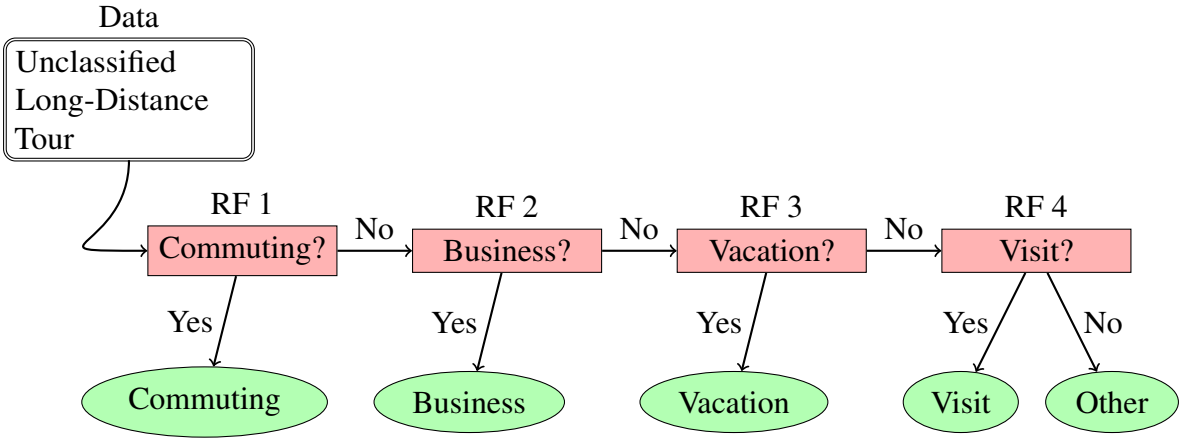
Data

The long-distance tour data described in this paper is based on an anonymised mobile phone billing datarecorded by Orange™France. It consists of **C**all **D**etail **R**ecords (CDRs) covering mobile phone usage of around 23 million users of the Orange™network in France during a period of 154 consecutive days (May 13, 2007 to October 14, 2007). Given a population estimate of 63.945 million inhabitants in 2007 [1], that is roughly 35.9% of the French population. The numbers match the estimates made at Orange™about mobile phone penetration in France anno 2007: 86% *(38)* and with a market share of Orange™in that year of 43.5%.

[1]This is the average of the monthly estimates for the period between Mai and October 2007 as obtained from the French National Statistics Website (www.insee.fr)



(a) Decision tree example: holiday identification



(b) Random forest sequence

FIGURE 1 Example for a single decision tree and the random forest sequence

Each CDR contains information about the action (outgoing/terminating call or sms) which took place in the network. Given location and time information for each action, a user can be traced and thus his or her movements can also be extracted. No information about phone calls made abroad is available in this data set. Even though it is known that France has one of the highest ratios of domestic trips to trips abroad within Europe *(39, 40)* this circumstance limits the scope to domestic long-distance tours.

The long-distance tour extraction algorithm is described in detail in *(5)*. In total, 578'614 domestic long-distance tours were reconstructed from the CDR data for the 79'866 mobile phone users that we tracked. The long-distance tour is defined as a trip chain that starts and ends at the home location of a person and at least one point in this trip chain being more than 80km away from this home location. This data can be very useful to transport planners since there are no traditional data sources for long-distance travel that provide as many observations as this data. Nevertheless, tour purposes are not part of the provided information. We will show in the following how the modified random forest approach presented in the previous section can be applied here.

Training Set

As mentioned, the CDR data described above does not provide any information about travel purposes. As a consequence we need an additional data source as training set for a purpose

classification algorithm. We rely on the **Enquête Nationale Transports et Déplacements (ENTD)**, the French National Travel Survey. The ENTD is conducted every 10-15 years (1967, 1974, 1982, 1994, 2007-08). Various actors are involved in the ENTD, including the Ministry of Transport, the INSEE (French National Institute of Statistics and Economic Studies) and IFSTTAR (French institute of science and technology of transport, planing and infrastructures).

The last ENTD was performed from April 2007 to April 2008 and most parts are publicly available *(41)*. Since the survey includes the time period covered by the CDR data described above, we use the ENTD 2008. One of the goals of this survey was the analysis of long-distance mobility. This fact is advantageous, because it ensures that we can compare the two data sources in terms of long-distance travel behaviour.

Nevertheless, the sample size of the ENTD 2008 is much smaller than the available CDR data. In total, 20'178 households and 44'958 individuals were surveyed. Just 18'632 (representative) persons were chosen for the long-distance travel module of the survey *(11)*. The latter were asked to report their long-distance travel practices within the preceding 4 weeks. In the ENTD 2008, a long-distance journey is either a journey with the furthest destination being more than 80km away from home (crow-fly distance) or a journey, which includes at least one overnight stay, or both. In order to adapt to the CDR data, tours shorter than 80km were removed as well as international tours.

Setup

We have described the data that has to be classified by purpose as well as the training set needed for the learning approach. However, the setup of the algorithm is not self-explanatory. The first choice that has to be made is the set of classes, i.e. one has to name the travel purpose classes that are assigned to the unclassified long-distance tours. The following classes were chosen in this application:

1. Commuting: Regular (daily or weekly) tours to a fixed workplace.
2. Business: Any tour with a professional background, besides commuting.
3. Vacation: Irregular, long-duration tours with a leisure purpose.
4. Visit: A tour with a purpose to meet somebody at the tour's destination (family or friends)
5. Other private reasons: e.g. shopping or leisure activities.

These purpose classes are most common for long-distance travel studies (e.g. in *(42)*) The order of classes in the multi-stage approach plays a role. The order chosen here is the one in the list (see also Fig. 1 (b)), because experiments have shown that this order ensures the best results (in terms of error rate for self-prediction).

Finally, the attributes used for classification have to be chosen. The choice set is very limited for the application described here, because no personal socio-demographic attributes are available. In other words, there is no information about the persons travelling other than the city of their home locations. Consequently, we focus on tour attributes, which can be extracted from the underlying data set. There are eight different attributes that are useful for a purpose classification. These attributes and their definitions are listed in Table 1. Five attributes are related to a specific tour (distance, duration, destination, month and weekend-share), two attributes take into account the overall long-distance travel behavior of the person (deviation and frequency) and one attribute accounts for the size of the person's home city (residence).

The destination attribute was defined by clustering French departements. Especially, coastal locations, mountain regions or the Ile de France were clustered, because the tours with destina-

TABLE 1 Attributes Used for Purpose Classification

Name	Range	Description
Destination		see Figure 2
Distance	80-1500 [km]	Distance between home location and furthest point on the tour
Duration	1-102[days]	Days between start and end of the tour
Month	$\{Jan, \cdots, Dec\}$	Month of the tour start
WE-Share	0.0-100.0 [%]	Share of the tour duration that is weekend (or public holiday)
Deviation	Yes/No	Is tour distance close to average distance of that persons tours
Frequency	0.0-31.0	Average number of domestic long-distance tours per month
Residence	0-8	Size of the home city: rural(0), <5k(1), 5k-10k(2), 10k-20k(3), 20k-50k(4), 50k-100k(5), 100k-200k(6), 200k-800k(7), Paris(8)

tions in these regions are likely to have similar purposes. For example, tours to the Cote d'Azur are likely to be vacation tours. The clusters defined here are 'Alpes', 'Ile de France' (without Paris), 'Mediterranian Area', 'North Atlantic Coast', 'North-East', 'Paris', 'Pyrennes', 'South Atlantic Coast' and 'Other Regions'. You can find a map of the clusters in Figure 2. All the other tour based and travel behavior based attributes are directly accessible in the underlying data set. The place of residence can be extracted, because the tours in the data set are home based. Therefore, an estimate of the home location is available.

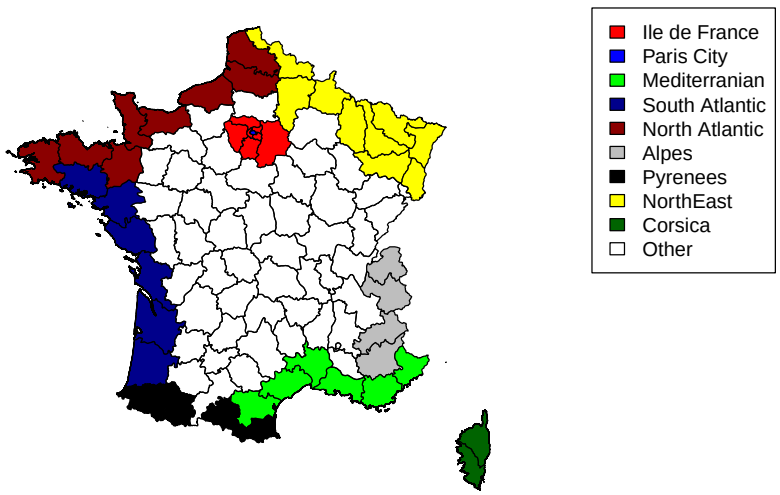


FIGURE 2 Destination Classes Used for Purpose Classification

The advantage of the multi-stage approach is that we can exclude specific attributes in the classification process of particular purposes. As an example, the destination or the month of a tour might be important to identify vacation tours, but is rather irrelevant for a classification of

commuting tours. Since irrelevant attributes can cause the random forest to learn dependancies that are not existent, it is beneficial to exclude them. Based on several experiments following attributes were excluded:

- Stage 1, Commuting: Destination, Month, Residence.
- Stage 2, Business: Duration, Deviation.
- Stage 3, Vacation: Deviation, Weekend-Share.
- Stage 4, Visits/Other: None.

Each of these exclusions is reasonable. The destination attribute and the month were mentioned above. Another example is the weekend-share attribute. While it plays an important role in the definition of commuting and business classes (namely being close to zero), it does not contribute to the classification of a vacation tour. In the last stage no attribute is excluded. This is due to the fact that we are separating *visits* from *other private* tours. The latter has no clear structure. Thus, all attributes have to be included.

Tuning and Stability Analysis

There are several parameters that can be used to tune the (multi-stage) random forest algorithm. The order of the classes within the series of random forests has been mentioned above. Additionally, there are two main parameters that can be used to tune each random forest individually. Firstly, the number of attributes that should be randomly selected for each split decision, referred to as m, can be chosen. Secondly, the number of grown decision trees, i.e. the size of the random forest, has to be set.

As suggested by Breiman *(35, 36)* a random forest can be validated by the Out-Of-Bag (OOB) Error. The OOB error is estimated during the random forest construction as follows. Each decision tree in the random forest is constructed by leaving out a subset of the training set (about one third). The observations left out are classified afterwards and the OOB error reports the share of misclassified cases.

Reasonable values for m are $1, 2, 3$ and 4. In order to evaluate the best value, the algorithm has been run ten times for each of the $4^4 = 256$ combinations of m for 4 stages. It turns out that a fixed $m = 2$ produces the lowest OOB error rates. Thus, $m = 2$ was chosen in this application. This is consistent with Breimans suggestion for m, namely the square-root of the number of used attributes. Here, the number of attributes is between 5 (stage 1) and 8 (stage 4).

The choice of the number of grown decision trees is a trade-off between computing time and stability. Figure 3 illustrates the OOB error for each stage in dependence of the number of trees. It can be seen that the error is stable after the 100th tree grown. Thus, any value above 100 is sufficient in this case. The figure shows that the OOB error is growing from stage to stage. This is not surprising due the order of purpose classes chosen here. While commuting (stage 1) has unique attribute patterns (e.g. frequency attribute), the differences between visits and other private tours are not obvious.

Training Set Prediction

Since there is no possibility to validate the presented methodology for the CDR data, we want to measure the quality of the random forests that were computed. Therefore, we report for each stage two different error measures (see table 2). On the one hand, the out-of-bag (OOB) error mentioned above is shown. On the other hand, the full data error was calculated as the share of miss-classified observations, when the purpose for the full training set is predicted by the whole

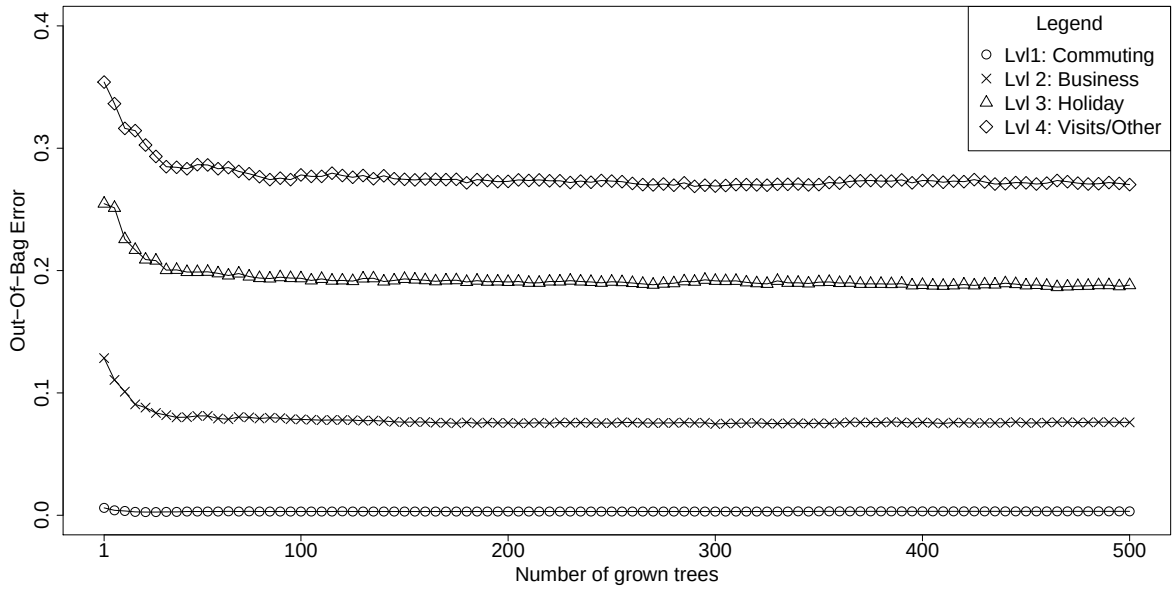


FIGURE 3 Out-Of-Bag error for different levels

random forest. In other words, the self-prediction of the training set was tested. One can see that the error levels are fairly low. Though, the error grows stage by stage. Consequently, the highest error rate can be found in the last stage, where we try to distinguish visits from other private tour purposes.

TABLE 2 Error levels of random forests per stage

	Stage 1 Commuting	Stage 2 Business	Stage 3 Vacation	Stage 4 Visits/Other
OOB Error	0.34%	7.58%	18.96 %	26.79%
Full Data Error	0.09%	0.02 %	0.07 %	0.19%

RESULTS

The outcomes of the purpose imputation algorithm described above are presented in this section. The multi-stage random forest algorithm was applied to the unclassified long-distance tour data 1'000 times. Before the final prediction results are presented, the focus is put on the attribute importance and the measurable uncertainty of the predictions.

Attribute Importance

It is clear that some attributes are more important than others for a classification with a random forest. Since the presented algorithm consists of several stages, the importance may also vary between the stages. Table 3 shows the different importance levels per stage and attribute. The importance can be measured by the increase of the OOB error after a permutation of the corresponding attribute's values. One can see that few attributes are very important for the classification, e.g. long duration is the main attribute for identification of holidays.

TABLE 3 Importance of the attributes used

Attribute	Stage 1 Commuting	Stage 2 Business	Stage 3 Vacation	Stage 4 Visits/Other
Distance	Average	High	Average	Average
Duration	Average		Very High	High
Destination		Average	High	Average
Month		Average	Average	Low
Weekend-Share	Low	Very High		Average
Deviation	Average			Low
Frequency	Very High	Very High	Average	Low
Residence		Average	Average	Low

Uncertain Predictions

Uncertain predictions can be identified since the algorithm was run for 1'000 times. If the predicted purpose of a tour is consistent in more 75% of the runs, is can be assumed the prediction is certain. These cases are called well classified in the following. The purpose prediction of the other cases is uncertain.

First of all, 402'908 of the 578'614 tours (69%) were classified with the same purpose by all of 1'000 generated random forests. More than 93% were well classified. This shows that the prediction method is robust. Table 4 shows the confusion matrix of not well classified tours. The matrix describes which purpose classes could not be clearly separated from other classes in the prediction process. One can see that a clear separation between visits and other private tours is difficult. These are the two purpose classes that interfere with each other the most during the prediction. However, the total number of not well classified tours is relatively small.

TABLE 4 Number of tours that were not well classified

	Confused with				
Main Purpose	Commuting	Business	Holidays	Visits	Other
Commuting	–	549	2	13	51
Business	390	–	460	1650	2988
Vacation	5	348	–	4382	1037
Visits	6	1429	4566	–	11309
Other	27	2874	723	10951	–

Purpose Prediction Results

Finally, the purpose classification is presented here. Table 5 shows the total number of predicted purposes, their shares and the corresponding shares from the training set. One can see that the predicted shares fit quite well the shares from the survey. The share of the vacation purpose is lower in the classified data set in comparison to the training set. Two reasons may influence this difference. On the one hand, vacations may be over-reported in the training set, which is a survey. Survey respondents tend to report longer tours (e.g. vacations) and tend to forget shorter tours

(e.g. business tours or visits). On the other hand, the vacation purpose is not clearly defined in the training set. One can find a substantial amount of vacation tours with short distances and short durations in the data set. This fact limits the random forests learning from the data.

TABLE 5 Predicted Purposes for the CDR data

	Commuting	Business	Vacation	Visits	Other	Uncertain
Total	18'733	87'418	64'738	231'800	136'968	38'957
Share	3.2%	15.1%	11.2%	40.1%	23.7%	6.8%
Share in ENTD 2008	3.5%	13.8%	22.0%	38.5%	22.2%	0.0%

DISCUSSION

We have presented an algorithm for purpose imputation in case of long-distance tours. The algorithm overcomes the lack of socio-demographic information by generating tour based attributes from the tour set. The presented approach is an adaption of a classic machine learning approach that was already utilized for GPS based travel data.

The presented random forest approach provides robust predictions for purposes long-distance travel. The question of the lower number of vacations in the classified data set was tackled. As explained, the reason might be in the training set. Nevertheless, the prediction results are reasonable in the end.

Travel data without a purpose classification is of little value for an analysis of (long-distance) travel demand. Consequently, the purpose classification technique presented in this paper is beneficial. It makes mobile phone billing data usable for transport planners with a focus on long-distance travel behavior.

CONCLUSION

This paper proposed a purpose imputation technique for long-distance tours that can be applied to data sources without socio-demographic information. The approach was applied to a mobile phone based data set with an enormous amount of unclassified long-distance tours.

There is no opportunity to validate the algorithm in combination with the data set directly. The validity of the proposed approach is supported by two facts. Firstly, self-validation of the used training set was succesful. Secondly, the purpose prediction provides results similar to the national travel survey statistics. Thus, the multi-stage random forest algorithm is the first applicable purpose imputation technique for this data type. Therefore, it is a valuable contribution to transport planners working with long-distance travel without personal information.

ACKNOWLEDGEMENTS

We are grateful for the CDR data, which was kindly made available by the Orange™Labs, FR. We are also thankful to Aare Puussaar for assisting us with his database knowledge and programming skills.
We would also like to acknowledge the Swiss National Science Foundation (SNSF) for providing funds to the authors under grant 205120_165900 *Long Distance Travel Demand Simulation*.

REFERENCES

1. Axhausen, K. W., J.-L. Madre, J. W. Polak and P. L. Toint (eds.) (2002) *Capturing Long-Distance Travel*, Research Studies Press, Hertfordshire.

2. Armoogum, J. and J.-L. Madre (2002) Weighting and correcting long-distance travel surveys, in K. W. Axhausen, J.-L. Madre, J. W. Polak and P. L. Toint (eds.) *Capturing Long-Distance Travel*, 152–170, Research Studies Press, Hertfordshire.

3. Madre, J.-L., K. W. Axhausen and W. Brög (2007) Immobility in travel diary surveys, *Transportation*, **34** (1) 107–128.

4. Kuhnimhof, T. and J. Last (2009) The Path to Better Long-distance Travel Data in Europe – The Potential of Combining Established Household Survey Instruments and Methodological Innovations Long Distance Mobility in Central Europe: Status Quo and Current Trends, paper presented at the *First International Conference on the Measurement and Economics Analysis of Regional Tourism*, San Sebastian, October 2009.

5. Janzen, M. and M. Vanhoof (2015) Report for the Collaboration between Orange Labs France and IVT, ETH Zurich, *Arbeitsberichte Verkehrs- und Raumplanung*, **1104**, Orange Labs, France and IVT, ETH Zurich, Zurich.

6. Ahas, R., A. Aasa, Ü. Mark, T. Pae and A. Kull (2007) Seasonal tourism spaces in Estonia: Case study with mobile positioning data, *Tourism Management*, **28** (3) 898–910.

7. Dillman, D. A. (2000) *Mail and Internet Surveys. The Tailored Design Method*, John Wiley & Sons, New York.

8. Zimmermann, A., K. W. Axhausen, J. Beckmann, K. J. Beckmann, M. Düsterwald, E. Fraschini, T. Haupt, A. König, A. Kübel, G. Rindsfüser, R. Schlich, S. Schönfelder, A. Simma and T. Wehmeier (2001) Mobidrive: Dynamik und Routinen im Verkehrsverhalten: Pilotstude Rhythmik, *Final Report*, Federal Ministry of Education and Research, PTV and IVT, ETH Zurich and Institut für Stadtbauwesen und Strassenverkehr (ISB), RWTH Aachen, Karlsruhe, Zurich and Aachen.

9. Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser and T. Haupt (2002) Observing the rhythms of daily life: A six-week travel diary, *Transportation*, **29** (2) 95–124.

10. Chalasani, V. S. and K. W. Axhausen (2004) Mobidrive: a six weeks travel diary, *Travel Survey Metadata Series*, **2**, IVT, ETH Zurich, Zurich.

11. Armoogum, J., J.-P. Hubert, D. Francois, B. Roumier, M. Robin and S. Roux (2008) Enquête nationale transport et déplacements 2007-2008, *Technical Report*, IFSTTAR, Paris.

12. Department for Transport (2016) National travel survey: 2014, webpage, `http://www.gov.uk/government/statistics/national-travel-survey-2014`. Accessed on 17/02/2016.

13. BMVI (2012) Oesterreich Unterwegs, webpage, `http://www.oesterreich-unterwegs.at`. Accessed on 26/04/2016.

14. Axhausen, K. W., B. Schmid and C. Weis (2015) Predicting response rates updated, *Arbeitsberichte Verkehrs- und Raumplanung*, **1063**, IVT, ETH Zurich, Zurich.

15. Axhausen, K. W. and C. Weis (2010) Predicting response rate: A natural experiment, *Survey Practice*, **3** (2).

16. Smith, R. S. and J. E. A. Wood (1977) Memory–Its Reliability in the Recall of Long Distance Business Travel, *Final Report*, Transport and Road Research Laboratory, Transport and Road Research Laboratory, Berkshire.

17. Bradburn, N. M., L. J. Rips and S. K. Shevell (1987) Answering autobiographical questions: the impact of memory and inference on surveys, *Science*, **236** (4798) 157–161.

18. Tourangeau, R. (1999) Remembering What Happened: Memory Errors and Survey Reports, in A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman and V. S. Cain (eds.) *The Science of Self-report: Implications for Research and Practice*, chap. 3, 29–47, Psychology Press, East Sussex.

19. Wolf, J., M. Oliveira and M. Thompson (2003) The impact of trip underreporting on vmt and travel time estimates: Preliminary findings from the california statewide household travel survey gps study, paper presented at the *82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2003.

20. Blondel, V. D., A. Decuyper and G. Krings (2015) A survey of results on mobile phone datasets analysis, *The European Physical Journal - Data Science*, **4** (1), February 2015.

21. Ahas, R., A. Aasa, A. Roose, Ü. Mark and S. Silm (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study, *Tourism Management*, **29** (3) 469–486.

22. Friedrich, M., K. Immisch, P. Jehlicka, T. Otterstätter and J. Schlaich (2010) Generating OD matrices from mobile phone trajectories, paper presented at the *89th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2010.

23. Pan, C., J. Lu, S. Di and B. Ran (2006) Cellular-Based Data-Extracting Method for Trip Distribution, paper presented at the *85th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2006.

24. Song, C., Z. Qu, N. Blumm and A.-L. Barabási (2010) Limits of predictability in human mobility, *Science*, **327** (5968) 1018–1021.

25. Fessant, F. and S. Midenet (2002) Self-Organising Map for Data Imputation and Correction in Surveys, *Neural Computing Applications*, **10** (4) 300–310.

26. Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

27. Shen, L. and P. R. Stopher (2013) A process for trip purpose imputation from Global Positioning System data , *Transportation Research Part C: Emerging Technologies*, **36**, 261–267.

28. Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira and K. W. Axhausen (2004) Eighty weeks of Global Positioning System traces, *Transportation Research Record*, **1870**, 46–54.

29. Wolf, J., R. Guensler and W. Bachman (2001) Elimination of the travel diary - experiment to derive trip purpose from Global Positioning System travel data, *Transportation Research Record*, **1768**, 125–134.

30. Lu, Y., S. Zhu and L. Zhang (2012) A machine learning approach to trip purpose imputation in GPS-based travel surveys, paper presented at the *4th Conference on Innovations in Travel Modeling*, Tampa, May 2012.

31. Montini, L., N. Rieser-Schüssler, A. Horni and K. W. Axhausen (2014) Trip purpose identification from GPS tracks, *Transportation Research Record*, **2405**, 16–23.

32. Alexander, L., S. Jiang, M. Murga and M. C. González (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data, *Transportation Research Part C: Emerging Technologies*, **58**, 240–250.

33. Griffin, T. and Y. Huang (2005) A decision tree classification model to automate trip purpose derivation, paper presented at the *18th International Conference on Computer Applications in Industry and Engineering*, Honolulu, November 2005.

34. Deng, Z. and M. Ji (2010) Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach, paper presented at the *7th International Conference on Traffic and Transportation Studies*, Kunming, August 2010.

35. Breiman, L. (2001) Random forests, *Machine Learning*, **45** (1) 5–32.

36. Breiman, L. and A. Cutler (2013) Random forests, webpage, `http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm`. Accessed on 29/07/2013.

37. Ham, J., Y. Chen, M. M. Crawford and J. Ghosh (2005) Investigation of the Random Forest Framework for Classification of Hyperspectral Data, *IEEE Transactions on Geoscience and Remote Sensing*, **43** (3) 492–501.

38. ARE (2016) Le Suivi des Indicateurs Mobiles - les chiffres au 31 décembre 2007 (publication le 4 février 2008), webpage, March 2016, `http://www.arcep.fr/index.php?id=9545`.

39. OECD (2012) OECD Tourism Trends and Policies 2012, webpage, `http://www.oecd-ilibrary.org/industry-and-services/oecd-tourism-trends-and-policies-2012_tour-2012-en`. Accessed on 26/04/2016.

40. Eurostat (2016) Tourism statistics - top destinations, webpage, `http://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics_-_top_destinations`. Accessed on 26/04/2016.

41. IFSTTAR (2016) French national travel survey: 2008, webpage, `http://http://www.statistiques.developpement-durable.gouv.fr/transports/s/transport-voyageurs-deplacements.html`. Accessed on 17/02/2016.

42. Outwater, M., M. A. Bradley, N. Ferdous, S. Trevino and H. Lin (2015) Foundational knowledge to support a long-distance passenger travel demand modeling framework, *Technical Report*, Transport Studies Unit, University of Oxford.