



Trip purpose imputation using GPS trajectories with machine learning

Journal Article**Author(s):**

Gao, Qinggang; Molloy, Joseph ; Axhausen, Kay W. 

Publication date:

2021-11-13

Permanent link:

<https://doi.org/10.3929/ethz-b-000505634>


Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

ISPRS International Journal of Geo-Information 10(11), <https://doi.org/10.3390/ijgi10110775>

Trip Purpose Imputation using GPS Trajectories with Machine Learning

Qinggang Gao ^{1,*} , Joseph Molloy ¹  and Kay W. Axhausen ¹ 

¹ Institute for Transport Planning and Systems, ETH Zurich, 8093, Zurich, Switzerland; gaoqg111@gmail.com (Q.G.); joseph.molloy@ivt.baug.ethz.ch (J.M.); axhausen@ivt.baug.ethz.ch (K.W.A.)

* Correspondence: gaoqg111@gmail.com;

Abstract: We studied trip purpose imputation using data mining and machine learning techniques based on a dataset of GPS-based trajectories gathered in Switzerland. With a large number of labeled activities in 8 categories, we explored location information using hierarchical clustering and achieved a classification accuracy of 86.7% using a random forest approach as a baseline. The contribution of this study is summarized below. Firstly, using information from GPS trajectories exclusively without personal information shows a negligible decrease in accuracy (0.9%), which indicates the good performance of our data mining steps and the wide applicability of our imputation scheme in case of limited information availability. Secondly, the dependence of model performance on the geographical location, the number of participants, and the duration of the survey is investigated to provide a reference when comparing classification accuracy. Furthermore, we show the ensemble filter to be an excellent tool in this research field not only because of the increased accuracy (93.6%) especially for minority classes, but also the reduced uncertainties in blindly trusting the labeling of activities by participants, which is vulnerable to class noise due to the large survey response burden. Finally, the trip purpose derivation accuracy across participants reaches 74.8%, which is significant and suggests the possibility of effectively applying a model trained on GPS trajectories of a small subset of citizens to a larger GPS trajectory sample.

Keywords: class noise; data mining; ensemble filter; hierarchical clustering; machine learning; random forest; trip purpose

Citation: Gao, Q.; Molloy, J.; Axhausen, K. Trip Purpose Imputation using GPS Trajectories with Machine Learning. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 0. <https://doi.org/>

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted to *ISPRS Int. J. Geo-Inf.* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Trip purpose imputation is an important part of constructing travel diaries of individuals and has attracted the attention of many researchers due to its significance in understanding travel behavior, travel demand prediction, and transport planning. The prevalence of GPS-integrated devices provides a large amount of GPS trajectories consisting of a series of longitude-latitude pairs with abundant explicit information (such as travel timing, duration, and location). Nevertheless, the implicit information like travel modes and purposes needs to be imputed to enrich such data for better usage in transport management. While it triggered plenty of studies over past decades [1], most of them focused on mode detection. Although trip purposes can be reported by participants along GPS trajectories, this needs too much effort over a long study duration. In addition, such surveys might suffer from inaccuracy problems due to memory recall issues or the inattention of travelers, and their applicability is still limited to the collected travel diaries. As existing trip purpose imputation studies are mainly confined to small-scale case studies, how to generalize the results into a larger scale continues to be an important research topic and becomes the focus of our work. For comprehensive reviews of research status on trip purpose imputation, readers can consult the studies of Nguyen *et al.* [2], Ermagun *et al.* [3], and Gong *et al.* [4].

The classification performance of different studies depends on many factors, such as sample sizes, survey duration and methods, data sources, activity categories, and

39 data preparation and cleaning steps. For this reason, it is difficult to set a benchmark for
40 comparison across different papers, so we only emphasize innovative aspects of the most
41 recent articles instead of comparing their accuracy rate. In our preliminary research,
42 we found a striking similarity between trip mode and purpose derivation, which are
43 mostly considered separately in the existing literature. While we saw a comparable
44 model performance on these two tasks with similar techniques, this article engages in
45 trip purpose imputation for simplicity and also mentions some relevant progress in
46 mode derivation papers.

47 While point of interest (POI) information is considered useful in identifying possible
48 activities in a venue, it is not easy to efficiently incorporate such data into an imputation
49 scheme. As a solution, Meng *et al.* [5] employed social media data (Twitter) to determine
50 the popularities of POI in trip end areas for purpose inference with dynamic Bayesian
51 network models. Scholars in this field seldomly investigate the transferability of trained
52 models to other distinguishable datasets, while Gong *et al.* [6] did look into this aspect.
53 They adopted the Aslan & Zech's test and random forests to explore the effects of datasets
54 from different seasons on model performance and stressed the limited transferability of
55 models across datasets. To maximize the benefits of activity type detection, Ermagun
56 *et al.* [3] took up the challenge of real-time purpose derivation and advocated the use
57 of Google Places information. To impute trip modes, Yazdizadeh *et al.* [7] found that a
58 combination of ensemble convolutional neural networks (CNN) and a random forest as
59 a meta learner outperforms single learners like a decision tree, a random forest, or single
60 CNN models.

61 Although extensive studies have been devoted to the study of trip purpose imputa-
62 tion and there are several comprehensive reviews of this research field, most of them
63 are limited to small-scale case studies and do not consider the generalizability of their
64 imputation scheme. Consequently, the large-scale spatial-temporal characteristics of
65 trip purpose derivation and the problem of mislabeling by participants have not been
66 investigated. Albeit an inverse relationship between sample size and model performance
67 is expected due to the heterogeneity in diverse samples [2], there is a lack of quantitative
68 measures for such phenomenon, which can be used as a guide for comparison across
69 studies and future research design. Moreover, while geographic variables like land
70 use information or POIs can be of benefit, they show large differences among different
71 regions and thus the models using such information are less transferable. Similarly,
72 the benefits of participant-related features come with a survey burden and limited
73 transferability.

74 Accordingly, our study does not aim at achieving a superior performance to exist-
75 ing methods or improving classification accuracy, but intends to address the practical
76 problems mentioned above. To this end, we propose four research questions: 1) What
77 is the minimum set of data sources for a satisfactory model performance, so that the
78 applicability of the methods can be maximized even with limited data availability? 2)
79 How does the model performance depend on the geographical location, the number of
80 participants, and the duration of the survey? 3) How can we account for the mislabeled
81 activities by participants during the survey? 4) Can a model trained on a relatively small
82 set of data be applied to other data collected from a much larger number of individuals?
83 To the best of our knowledge, this is the first time that such problems are addressed in
84 the trip purpose detection context.

85 The rest of this paper is structured as follows: Section 2 covers the relevant literature.
86 The data and methods are presented in Section 3. Section 4 presents the results, with the
87 discussion and conclusions in Section 5.

88 2. Background

89 2.1. Data Sources

90 Besides information about location and time collected with GPS-integrated devices,
91 additional data sources are normally included in models to improve travel purpose

92 imputation precision. Generally, the socio-demographic characteristics of participants
93 are gathered together with GPS trajectories and are taken to be important supplementary
94 information [1]. Land use data and POI could be used to indicate possible activities for a
95 stopping point on GPS trajectories [8]. In addition, the popularity of POI inferred from
96 social media data (e.g. Twitter) [5], travel and tourism statistics [9], and mobile phone
97 billing data [10] have also been utilized to derive travel purpose.

98 2.2. Data Preparation

99 Data pre-processing, which has been intensively investigated in the data mining
100 field [11], receives much less discussion than it deserves in trip purpose imputation
101 research. Therefore, we discuss the issue in-depth below. García *et al.* [12] summarized
102 the three most influential data pre-processing requirements to improve data mining
103 efficiency and performance, i.e. imperfect data handling, data reduction, and imbalanced
104 data pre-processing.

105 An important aspect of imperfect data handling is noise filtering [13], which aims
106 at detecting the attribute noise and the more harmful class noise [14]. For class noise
107 removal, ensemble filters proposed by Brodley and Friedl [15,16] have been widely
108 applied as an excellent tool. Ensemble filters adopt an ensemble of classifiers to eliminate
109 the mislabeled training data that cannot be correctly classified by all or part of the
110 classifiers using n-fold cross-validation. To avoid treating an exception that is specific to
111 an algorithm as noise, multiple algorithms are used. Basically, there are two strategies
112 for implementing ensemble filters: majority vote filters, which mean the instances
113 that cannot be correctly classified by more than half of the algorithms are treated as
114 mislabeled; and conservative consensus filters, which mean only the instances that
115 cannot be correctly classified by all algorithms are treated as noise. Majority vote filters
116 are sometimes preferred to conservative consensus filters, as retaining bad data is more
117 harmful than discarding good data especially when there are ample training data [16].
118 Nevertheless, we chose conservative consensus filters, with the results of these two
119 strategies being similar.

120 Missing data is another typical problem in transport research that normally involves
121 survey processes. The first step to handle missing data should be understanding sources
122 of “unknownness” [17], which might be due to lost, uncollected, or unidentifiable in
123 existing categories. Besides omitting the instances or features with missing values, which
124 is usually not suggested, approaches for missing data inference can be classified into two
125 groups [18]: data-driven, e.g. mean or mode; and model-based, e.g. k-nearest neighbors
126 (kNN). kNN has gained popularity because of its simplicity and good performance in
127 dealing with both numerical and nominal values [19].

128 Attribute selection, as a classic part of data reduction, is conducive to generating
129 a simpler and more accurate model and avoiding over-fitting risks [12,20]. For feature
130 selection, feature importance measured by mean decrease in the Gini coefficient in
131 the random forest approach can be used as a reference [21]. However, such a rank-
132 based measure cannot take feature interactions into account and might suffer from
133 stochastic effects [22]. Conventionally, feature selection techniques can be grouped into
134 two categories: filter methods, i.e. variable ranking techniques; and wrapper methods,
135 which involve classifiers and become an NP-hard problem [20]. One of the most popular
136 algorithms for feature selection is minimum redundancy maximum relevance based on
137 mutual information [23], which is initially designed as a filter and then developed to
138 be a wrapper as well [12]. Another popular wrapper algorithm that is designed for the
139 random forest is provided in an R package Boruta [22], which aims at identifying all
140 relevant features rather than an optimal subset and is employed for our analysis.

141 An imbalanced distribution of categories might result in unbalanced accuracies
142 of classification. This problem also troubled the machine learning community, where
143 Ling and Li [24] suggested duplicating small-portion classes and Kubat and Matwin
144 [25] tried to downsize large-portion classes. One of the most prevalent ways to cope

145 with imbalanced data is the Synthetic Minority Over-Sampling Technique (SMOTE)
146 introduced by Chawla *et al.* [26], which suggests formulating new samples as randomized
147 interpolation of minority class samples. SMOTE is widely used because of its simplicity,
148 good performance, and compatibility with any machine learning algorithm [12]. As
149 a variation of SMOTE, Adaptive Synthetic Sampling Approach (ADASYN) proposed
150 by He *et al.* [27] puts more weight on minority samples that are harder to learn when
151 selecting samples for interpolation.

152 2.3. Classification Techniques

153 The methods used to derive trip purposes can be divided into two main categories
154 [28]: rule-based systems with an accuracy of around 70% [29], which rely predominantly
155 on land use and personal information, as well as timing, duration, and sequence of
156 activities; and machine learning approaches, which focus more on activities than position
157 and show varying accuracy between 70% and 96% depending on different algorithms,
158 data set, activity categories, and so on [8]. Although manual trip purpose derivation
159 approaches using rules give satisfactory results, there is no standard set of accepted rules
160 for mining travel information and thus it relies on researchers' experiences. Compared to
161 conventional deterministic approaches, machine learning algorithms like random forest
162 and dynamic Bayesian network models could even rank possible activities, which are
163 particularly helpful when activities are ambiguous [5]. Consequently, we opt for machine
164 learning approaches that have already been widely applied in this area, such as decision
165 trees [30], random forests [28], artificial neural networks [31], and dynamic Bayesian
166 network models [5]. Because of the good performance of random forests compared to
167 other methods demonstrated by numerous studies [32–34], we employed it as a starting
168 point for analysis. An introduction to random forests is given in Section 3.2.

169 2.4. Model Performance Assessment

170 Model performance can be assessed in various ways, which act as an important
171 component of model development. Although reported trip information might suffer
172 from memory recall errors or other issues, it is probably the best candidate as ground-
173 truth for model validation and assessment [35]. Innovatively, Li *et al.* [36] used the
174 visualized spatial distribution of recognized trip purposes to validate simulation outputs.
175 Albeit classification models might be used to generate travel diaries for citizens that are
176 not in the training dataset, Montini *et al.* [32] found that the accuracy of trip purpose
177 detection is participant-dependent. As proportion and categories of trip purposes have a
178 significant influence on the accuracy of classification [9], high-frequency activities should
179 be treated with special care.

180 3. Materials and Methods

181 3.1. Materials

182 In this study, we analyzed GPS trajectories collected from 3689 Swiss participants
183 from September 2019 to September 2020 through the “Catch-my-day” GPS tracking
184 app, developed by Motion Tag. Considering solely the 91% of all activities that are
185 within Switzerland, it amounts to 1.82 million activities above a time threshold of 5
186 minutes, of which 43% is labeled by participants. Although a threshold of 5 minutes to
187 extract activities from GPS trajectories might ignore some short activities, we use it as a
188 simplification for the current study. As a GPS-integrated mobile phone has a position
189 error of 1 to 50 meters with a mean of 6.5 meters as shown by Garnett and Stewart [37],
190 this is taken into account when conducting spatial clustering of activities. More details
191 about the study design and research scope can be found in Molloy *et al.* [38] and Molloy
192 *et al.* [39].

193 Based on the “Mobility and Transport Microcensus 2015” in Switzerland, we
194 grouped activities into eight categories as shown in Table 1 with decreasing frequen-
195 cies of their occurrence. Following “Home” and “Work”, “Leisure” becomes the most

196 frequent activity and involves sophisticated characteristics that require special atten-
 197 tion [40,41]. The extracted features are shown in Table 2 and split into three types:
 198 personal-based, activity-based, and cluster-based information.

Table 1. Activity categories.

Category	Example activities	Count	Percent
Home	Any activities at home	293,129	16.1
Work	Any activities at work place	171,329	9.4
Leisure	Exercise, travel	123,735	6.8
Shopping	Food, clothing	64,071	3.5
Other	Transfer	46,413	2.5
Errand	Travel for business	40,119	2.2
Assistance	Pick up/drop off	28,189	1.5
Education	University, school	12,694	0.7
Unlabeled	-	1,041,409	57.2
Total	-	1,821,088	100

Table 2. Selected features for trip purpose imputation. The categorical features are indicated by *, while m() and std() denote "mean of" and "standard deviation of", respectively.

Personal-based	Activity-based	Cluster-based
Household size	Duration	m(duration)
Employment*	Start time	std(duration)
Age	End time	m(start time)
Annual income*	Day of week*	std(start time)
If a worker*	Activities per day	m(end time)
If a student*	-	std(end time)
-	-	Percentage of weekdays
-	-	Percentage of activities per cluster
-	-	Daily occurrence
-	-	Distance to most often visited cluster

199 Moreover, POI information from Google Places API as adopted in Ermagun *et al.* [3]
 200 was investigated for a pilot study and not considered further due to the large monetary
 201 cost for large datasets such as the one used here and its comparatively minor benefits.
 202 Residential zoning information in Switzerland as land use information is also tested
 203 with very little effect on trip purpose derivation accuracy and hence excluded from the
 204 final models.

205 3.2. Methods

206 As a classification method, kNN [42] is also shown to be a good missing value
 207 imputation technique [12,19]. Here we give a short introduction to the kNN algorithm.
 208 Given a training set $T = \{\mathbf{U}, \mathbf{V}\}$, where \mathbf{U} are predictors and \mathbf{V} are labels, we can
 209 estimate the distance between a test object $w_0 = \{u_0, v_0\}$ and all training objects $w =$
 210 $\{u, v\} \in \{\mathbf{U}, \mathbf{V}\}$ to find its k nearest neighbors. Then the label v_0 for this test object
 211 w_0 is determined as median of v of its k nearest neighbors in the case of numerical
 212 variables and mode in the case of categorical variables. The Gower distance computation
 213 between u_0 and u , which is applicable for both categorical and continuous variables,
 214 can be referred to in Kowarik and Templ [43]. Two issues might affect the performance
 215 of kNN: one is the choice of k , where a small value of k could be noise sensitive and a
 216 large value of k might include redundant information; another issue is that an arithmetic
 217 average might ignore the distance-dependent characteristics, where closer objects have
 218 higher similarities. These two issues can be addressed by weighting the vote of each

219 nearest neighbor for the final result by their distance, i.e. weighted kNN. Missing
 220 value imputation for personal-related information in this work is conducted using the R
 221 package “VIM” developed by Kowarik and Templ [43], which also provides weighted
 222 kNN methods for better performance.

223 To explore implicit information contained in the data, data mining techniques like
 224 clustering can be employed [28]. Using the hierarchical clustering method introduced by
 225 Ward Jr [44], we grouped the spatial location of activities for each participant to make use
 226 of repetitive patterns of human behaviors. Hierarchical clustering optimizes the route
 227 by which groups are obtained [45], so it might not give the best clustering result for a
 228 specified number of groups [44]. However, compared to another widely known k-means
 229 clustering technique, hierarchical clustering allows us to define the distance used for
 230 grouping rather than defining the number of groups. The basic steps for hierarchical
 231 clustering are illustrated below: 1) Treat initial x objects as individual clusters; 2) Group
 232 a pair of the most “similar” clusters; 3) Repeat step 2 until a single cluster containing all
 233 objects is obtained. To define the “similarity” between two clusters, [45] summarized six
 234 strategies, from which we selected the “Group-average” strategy as it is more reasonable
 235 and conservative than its alternatives. In our case, the similarity between two activities
 236 is defined as the Euclidean distance of their geographical location. Next, we use two
 237 general activity clusters X and Y to illustrate the estimation of their average distance.
 238 Assuming there are m and l activities in clusters X and Y , respectively, while i and j
 239 are single elements of the m and l activities, respectively. We use d_{ij} to represent the
 240 distance between activities i and j , d_{XY} the distance between clusters X and Y . Then we
 241 can calculate d_{XY} as:

$$d_{XY} = \frac{\sum_{i=1}^m \sum_{j=1}^l d_{ij}}{ml}. \quad (1)$$

242 Through the process of hierarchical clustering, d_{XY} will increase gradually. There-
 243 fore, we can define an appropriate threshold to stop the process and get intermediate
 244 clustering results. In our study, a threshold of 30 meters is chosen to restrict the size of
 245 each cluster considering the GPS accuracy [37] and results in a radius of fewer than 30
 246 meters for each cluster.

247 A random forest is an ensemble of classification and regression trees [46]. Since
 248 its introduction, classification and regression tree (CART) has been an important tool
 249 and received lots of attention in different research fields [42]. A detailed description of
 250 CART can be found in Song and Ying [47]. As a further development of CART, Breiman
 251 [21] developed the random forest with detailed proofs and experiments based on prior
 252 studies.

253 The process to develop a forest comprises three stages: 1) Bootstrap N sets of
 254 samples from and with the same size as training data; 2) Build a decision tree for each
 255 sample, and at each node choose the best feature from randomly selected M features;
 256 3) Obtain classification results as the mode of outputs of all trees. As classification
 257 algorithms are unstable, this bagging (bootstrap aggregating) process could improve the
 258 accuracy of model results [48]. The ensemble method with sampling techniques has also
 259 the advantage of more accurate imputation in case of imbalanced distribution across
 260 different activities [5]. The classification power and generalization errors of random
 261 forests depend on the accuracy and interdependence of each tree, which can be measured
 262 by out-of-bag (OOB) errors [49] with two steps: 1) For each tree, predict the data that
 263 are not in its bootstrap samples (also called OOB data, about 37% of the training set); 2)
 264 Aggregate predictions and calculate error rates.

265 The advantages of the random forest are multifaceted. Firstly, the generalization
 266 error converges with the increase of the number of trees N , so there is no over-fitting
 267 problem based on the strong law of large numbers even when N gets large, which allows
 268 us to select a large N as long as it is computationally efficient. OOB estimates can not only
 269 reveal generalization errors, variable importance, strength and correlation of trees, but
 270 also replace a test set as it is as accurate as using a test set of the same size as the training

271 set. Also, OOB estimates are unbiased in contrast to cross-validation with unknown bias.
272 In addition, it is robust in the case of unbalanced class population, missing data, and
273 noise, which often exists in labels of objects [50]. The significance of forest parameters,
274 e.g. N , M , and the maximum final node size of trees, as well as multiple extensions of
275 random forests, are well summarized by Biau and Scornet [51]. Khoshgoftaar *et al.* [52]
276 suggested default values of $N = 100$ and $M = \log_2 m + 1$ through extensive experiments,
277 where m is the number of features. While many efforts have been devoted to improving
278 the original random forest approach [51,53], the implementation of random forests in
279 this paper is based on a classic R package “randomForest” developed by Liaw and
280 Wiener [54].

281 In addition to the above-mentioned algorithms, C5.0 [55] - an extension of a well-
282 known classification algorithm C4.5 [56], naive Bayes classifier [57], and multivariate
283 adaptive regression splines (MARS) [58] are adopted in the framework of ensemble
284 filters. In our preliminary analysis, principal component analysis for numerical features
285 transformation, support vector machine [59], which is time-consuming ($O(N^3)$) for high
286 dimensional data, and ADASYN [60] were tested but excluded from further analysis
287 because of limited contributions and high computational requirements. Furthermore,
288 Janzen *et al.* [10] proposed a Multi-Stage Random Forest method as a modification to
289 account for the independence of certain trip purposes on specific tour attributes, but this
290 complicated method did not improve the model performance in our re-implementation.

291 4. Results

292 4.1. Initial Analysis using Random Forests

293 The performance of random forests can be measured through OOB error rates
294 without splitting the training and test dataset and implementing cross-validation, so we
295 use only labeled data as training data in this subsection for supervised machine learning.
296 Table 3 presents the confusion matrix of labeled versus predicted trip purposes using
297 random forests. We set $N = 100$ and $M = \log_2 m + 1$ as suggested in Khoshgoftaar *et al.*
298 [52], which approaches the best possible performance in reasonable computation time.

299 Several important patterns can be observed in Table 3: Firstly, an overall accuracy
300 of 86.7% indicates a satisfactory performance of random forests as already demonstrated
301 by numerous studies. Secondly, the accuracy for each activity category decreases approx-
302 imately in sync with their occurrence frequency except for “Education”. Two reasons
303 might explain this phenomenon: One is that “Home”, “Work”, and “Education” have
304 more regular spatial and temporal patterns, so it is easier to correctly classify them;
305 Another reason is that the imbalanced distribution of these categories makes random
306 forests prefer labeling ambiguous objects as majority classes, as has been discussed
307 by del Río *et al.* [61]. Another interesting phenomenon in Table 3 is that all categories
308 except “Leisure” are most likely to be mislabeled as “Leisure”, which involves more
309 complicated characteristics that often make it hard to be distinguished from other cate-
310 gories. In addition, the difference in precision and accuracy might result from specific
311 characteristics of each category: For “Home” and “Leisure”, accuracy is slightly higher
312 than precision as it is safer for the model to classify ambiguous objects as these major-
313 ity classes, while accuracies of “Errand”, “Other”, and “Assistance” are lower for the
314 same reason. To better understand the strengths and possible improvement of random
315 forests, we investigate the importance of feature selection, the number of participants
316 and duration of the survey, and spatial characteristics of the accuracy.

Table 3. Confusion matrix of labeled versus predicted trip purposes using random forests (Overall accuracy: 86.7%).

Labeled \ Predicted										Accuracy	Precision
	Home	Work	Leisure	Shopping	Errand	Other	Assistance	Education			
Home	286,000	1,980	2,980	937	720	558	384	20		97.4%	95.8%
Work	3,670	157,000	5,400	1,680	1,490	1,270	362	131		91.8%	92.0%
Leisure	3,430	3,850	105,000	4,670	2,720	2,930	870	190		84.9%	74.0%
Shopping	1,340	1,960	7,770	47,600	2,690	2,130	511	71		74.3%	74.7%
Errand	2,020	2,630	7,700	4,400	27,000	1,960	560	107		58.3%	72.7%
Other	1,090	1,680	6,870	2,710	1,540	25,600	429	192		63.9%	71.7%
Assistance	913	1,110	5,430	1,590	861	1,030	17,200	50		61.1%	84.5%
Education	64	448	737	149	145	267	39	10,800		85.4%	93.4%

317 An advantage of the random forest is that it provides an inherent measure of
 318 feature importance using Gini impurity as shown in Fig. 1, which provides an important
 319 reference on feature selection. Among the 21 features, the most important six features
 320 are more useful in classification, whereas the personal-based attributes are less relevant:
 321 except for "Age", all personal information belongs to the least relevant 7 features. To
 322 assess the importance of three sets of features grouped in Table 2, we conduct three
 323 additional experiments by leaving one set of features out and present the results in Fig.
 324 2. When leaving all the personal information unused, the overall accuracy decreased
 325 around 0.9%. Although the Boruta method [22] shows that all features are relevant,
 326 which indicates a good result of our preliminary feature selection, we omit the personal
 327 information from further analysis for the following reasons: This could indicate the
 328 strength and applicability of our method even when no personal information is available,
 329 i.e., we can undertake trip information enrichment at high accuracy using only GPS
 330 trajectories; The inclusion of socio-demographic data might lead to overfitting of models
 331 to current participants and limit the applicability of models on GPS trajectories of other
 332 users. While the elimination of activity information gives similar results, the removal of
 333 cluster-based information leads to a dramatic decrease in model performance, which
 334 strongly suggests the effectiveness of our usage of hierarchical clustering algorithms.

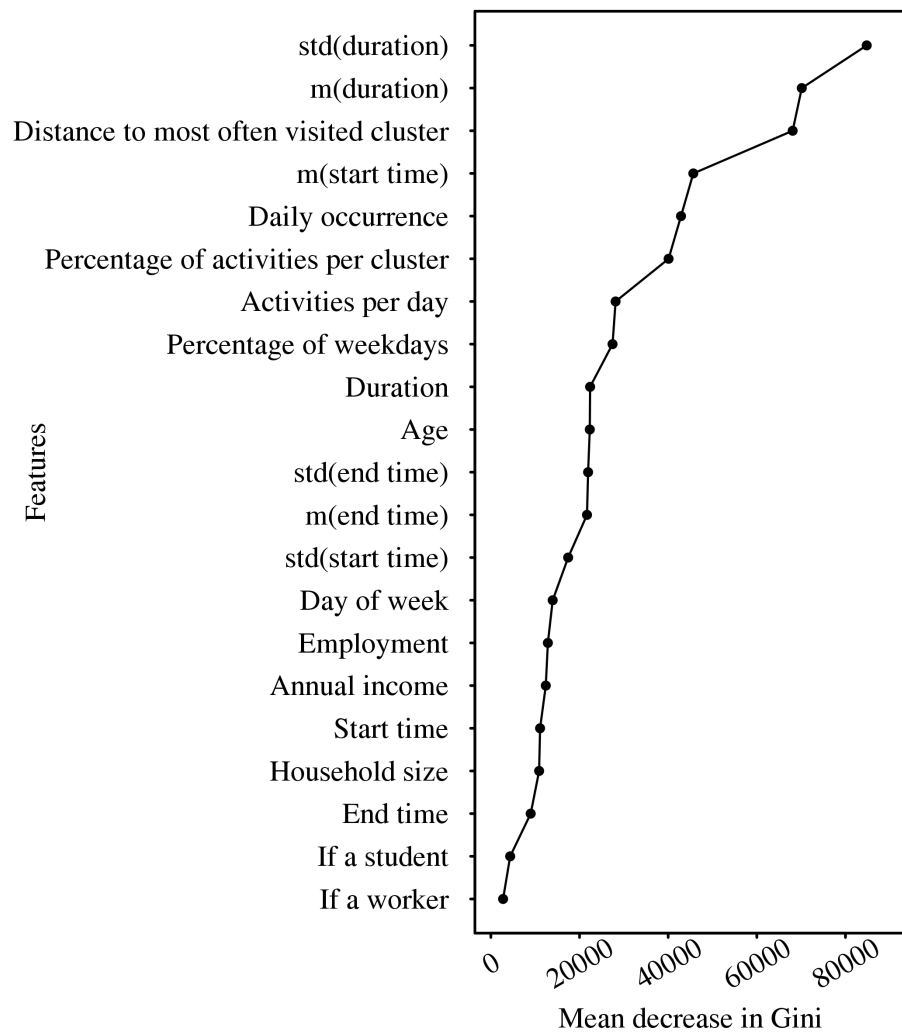


Figure 1. Feature importance in trip purpose imputation measured with mean decrease in Gini in random forests.

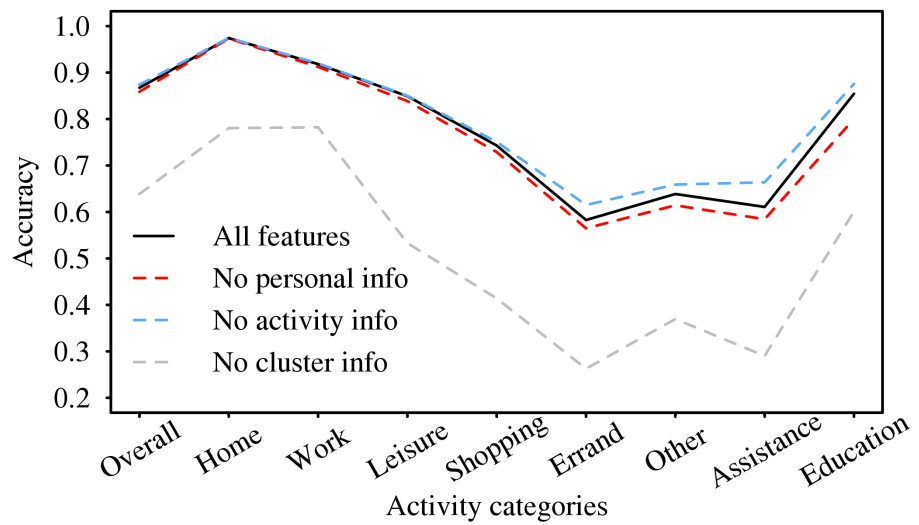


Figure 2. The model performance for each activity categories and the overall accuracy in four experiments, where we use all features or leave one set of features unused to measure the significance of each set of features.

335 Figure 3 shows the spatial distribution of labeled activities and accuracy rate using
 336 grids with an area of 4 km^2 in Switzerland. As these two fields have a small correlation
 337 coefficient of 0.23, we cannot conclude that higher spatial activity density, which normally
 338 means an urban area, will result in a higher accuracy rate. However, the five big cities in
 339 Switzerland with higher activity density seem to correspond to a more homogeneous
 340 accuracy rate. We can also observe that low activity density areas show large fluctuations
 341 in the accuracy rate.

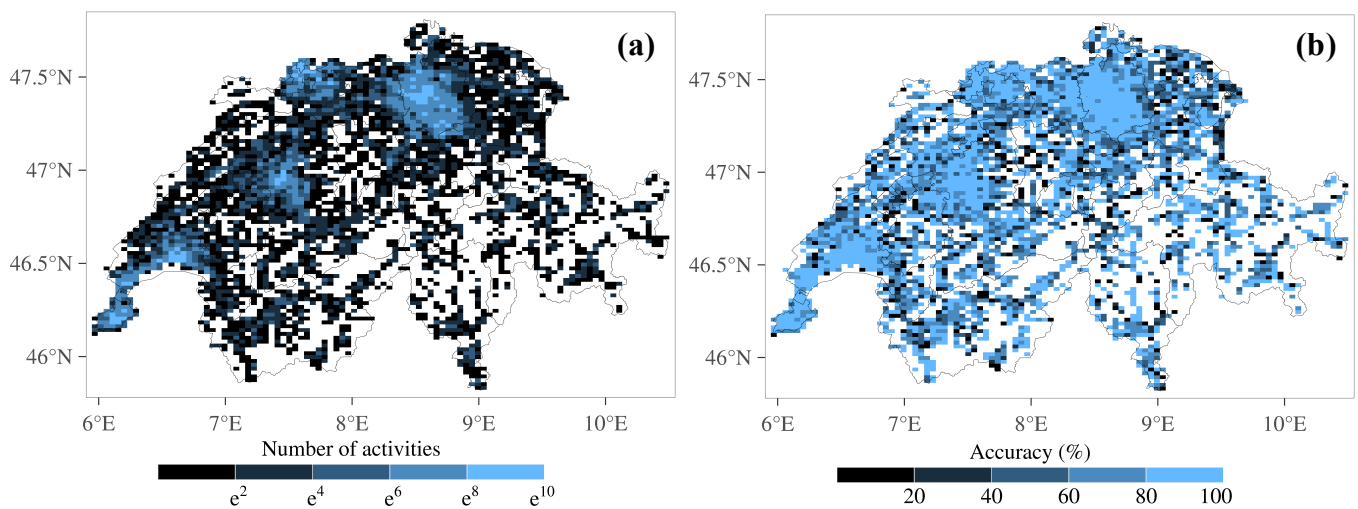


Figure 3. The spatial distribution of the number of labeled activities (a) and accuracy rate (b) using grids with an area of 4 km^2 in Switzerland. The exponential scale in (a) is used to account for the unevenly distributed activities.

342 To investigate the dependence of classification performance on the number of partic-
 343 ipants and the duration of the survey, we extract five groups of participants with different
 344 duration of the survey - from 60 days to 300 days - during which all activities are labeled
 345 as shown in Fig. 4. Several interesting patterns can be observed and could provide
 346 a reference when comparing results in existing literature with different datasets and
 347 designing similar research: Longer survey duration leads to higher accuracy, whereas
 348 increasing the number of participants deteriorates the accuracy due to more heteroge-
 349 neous data; When there are only 8 participants, the model performance undergoes some
 350 fluctuations at a short survey duration; Moreover, there seems to be an upper bound
 351 at around 90%. Further research is required to determine whether this upper bound
 352 is due to stochastic human behaviors, model ability, incomplete information, or class
 353 noise. In the next subsection, we focus on class noise, which has not been discussed in
 354 the existing transport literature, due to the smaller datasets available in this research
 355 field. It is, however, an essential consideration when dealing with large data sets like
 356 ours. We also propose a new criterion in exploring additional features and improving
 357 model performance in the next subsection.

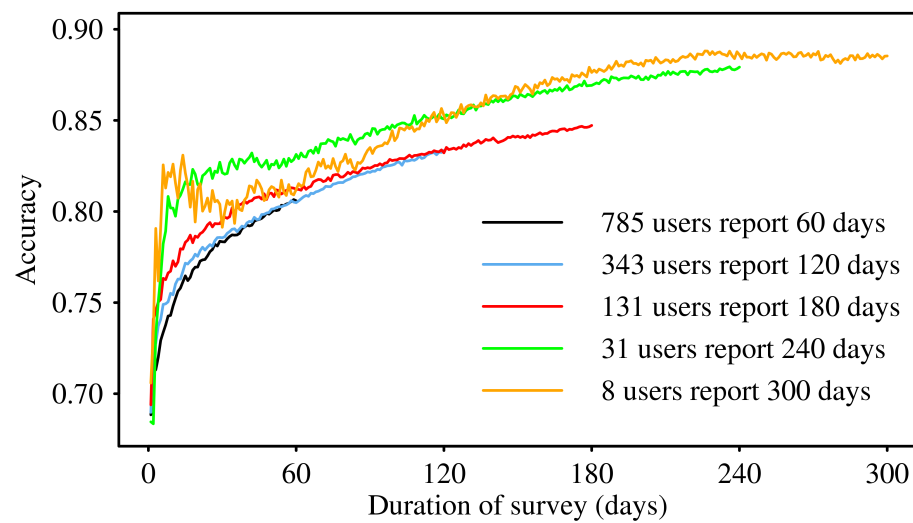


Figure 4. The impact of the number of participants and the duration of the survey on model performance.

358 4.2. Ensemble Filter with Multiple Classification Algorithms

359 A large data set is more vulnerable to class noise than smaller ones because of the
 360 heavier and longer survey response burden of participants. It is a challenging topic that
 361 has not been considered in the context of trip purpose imputation. Although it has been
 362 intensively studied in the machine learning field [13,15,62], no perfect solution exists
 363 due to a lack of validation information from real data. For our research, we investigate a
 364 very popular solution – the ensemble filter - proposed by Brodley and Friedl [15]. The
 365 main idea behind the ensemble filter is to identify mislabeled instances that cannot be
 366 correctly classified by a set of classifiers. We employ four classifiers with satisfactory
 367 performance – random forests, C5.0, Naive Bayes classifiers, and MARS – based on a
 368 preliminary test on a pool of algorithms. In this case, we use 10-fold cross-validation to
 369 assess model performance. For cross-validation, we also split training and test datasets
 370 based on participants, i.e. we test the model performance across participants.

371 The results are shown in Table 4. For the original labeled data, the random forest
 372 gives the best results with an overall accuracy rate of 85.8% and is followed by C5.0 with
 373 84.7%. Naive Bayes classifiers have the lowest accuracy of 57.2%, which is still higher
 374 than the suggested threshold (50%) for classifiers in the ensemble filter [16]. Using the
 375 strategy of conservative consensus filters in ensemble filters, we removed 8.5% of the
 376 labeled data. The model performance improved significantly on these ensemble filtered
 377 data - 93.6% with random forests and 93.2% with C5.0. These results are promising
 378 because of not only the increased accuracy, but also the reduced uncertainties in blindly
 379 trusting labels recorded by participants. The minority classes benefit more from this
 380 technique as shown in Fig. 5, where the accuracy of “Errand”, “Other”, and “Assistance”
 381 increased by 23%, 17%, and 29%, respectively. When the model is applied across
 382 participants, the accuracy of random forests and C5.0 decreased by about 20% as in
 383 Table 4, whereas Naive Bayes classifiers and MARS show nearly no deterioration. The
 384 classification accuracy of random forests (74.8%), which is applied across participants on
 385 the ensemble filtered data, is an acceptable baseline considering the limited information
 386 and inherent difficulties of the across-participants classification. The behavior of Naive
 387 Bayes classifiers and MARS in this example might require further exploration in a future
 388 study. When one plans to improve the model performance through incorporating more
 389 features or investigating new algorithms, considering the model performance across
 390 participants should be an essential part to avoid overfitting to a training dataset, which
 391 has inherent differences with a test set.

392 The results in this paper indicate some directions for future research. The trip
 393 purpose imputation across participants or across data that are inherently different should

394 still be improved for wider applicability in transport management, where the possibility
 395 might exist in including other data sources. While the division of activity categories
 396 is primarily subject to practical applications, its effects on model performance could
 397 be quantified in further analysis. In addition, the complexity of specific activities like
 398 "Leisure" can be considered, such as dividing it into several categories.

Table 4. Classification accuracy of multiple algorithms with ensemble filter and across participants imputation.

	Random forest	C5.0	Naive Bayes	MARS
Original data	85.8%	84.7%	57.2%	66.7%
Ensemble filtered (8.5%) data	93.6%	93.2%	61.8%	73.0%
Original data, across participants imputation	68.0%	65.4%	57.2%	66.6%
Ensemble filtered (8.5%) data, across participants imputation	74.8%	72.3%	62.0%	72.7%

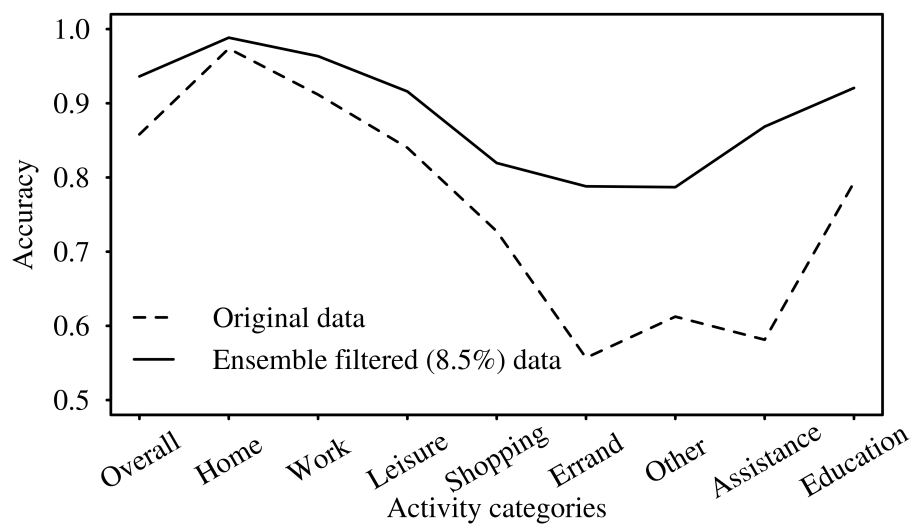


Figure 5. The model performance on the original data and the ensemble filtered data through four classification algorithms.

399 5. Discussion and Conclusions

400 To summarize, this paper investigated multiple classification algorithms including
 401 the random forest for trip purpose imputation to enrich GPS trajectories with data
 402 mining techniques like hierarchical clustering and ensemble filters.

403 As a baseline, we achieved an overall accuracy rate of 86.7% for eight activity
 404 categories using the highly heterogeneous data (3689 participants) with random forests.
 405 Through feature importance analysis using the inherent measure of the mean decrease
 406 in Gini of random forests and the Boruta method, we verified that current features are of
 407 high relevance and the features extracted with hierarchical clustering are crucial in model
 408 performance. Additional experiments that leave out a set of personal-related features
 409 reveal the possibility of trip purpose imputation with only GPS trajectories. Thanks
 410 to the innovative application of hierarchical clustering in extracting relevant features,
 411 the answer to the first research question becomes obvious: the required data sources
 412 for a satisfactory model performance are minimized to GPS trajectories. Although
 413 many researchers managed to achieve better performance by incorporating various
 414 data sources, we advocate considering limited data availability on a larger scale, where
 415 collecting personal information along with GPS trajectories is impossible or the quality
 416 of data sources varies considerably, is vital to generalize our results.

417 In this context, it is important to note, this is misleading to compare accuracy rates
 418 among papers due to the different sample sizes (persons and length of observation
 419 periods), activity categories, and data sources. To alleviate this circumstance and pro-

vide a reference for the design of similar research, we investigated the dependence of classification performance on the geographical location, the number of participants, and the duration of the survey. Taking advantage of the abundant (780,000) user-labeled activities along GPS trajectories, the spatial distribution of the accuracy rate over Switzerland is visualized. This result is meaningful for densely populated regions where better transport management is required. We show that the model performance in these regions undergoes fewer fluctuations and is more reliable. Furthermore, a longer survey duration helps to improve performance except for the fluctuations when only limited participants are involved over a short period, whereas a larger number of participants results in a higher diversity of the data that is detrimental to model performance but helps improve representativeness.

The employment of the ensemble filter improves model performance significantly from 85.8% to 93.6% with random forests, particularly for minority classes that have lower accuracy due to imbalanced class distribution and complicated characteristics of instances. Besides improving accuracy, the ensemble filter is also effective in reducing errors caused by mislabeling, to which our dataset is vulnerable due to the large response burden imposed on participants.

Another aspect that has not been studied in the existing literature is the trip purpose derivation across participants, where we obtained an accuracy rate of 74.8% using random forests. This result is quite promising, as it indicates that we can apply our trained model using only GPS trajectories and user labels to other GPS trajectories at a much larger scale, without the need to collect additional personal information. As the collection of GPS trajectories exclusively involves much less effort and monetary costs, the applicability of our imputation scheme could be readily expanded, which is significant for transport demand prediction and transport planning at a large scale.

Author Contributions: Conceptualization, Kay W. Axhausen and Joseph Molloy; methodology, Qinggang Gao; software, Qinggang Gao; validation, Qinggang Gao and Joseph Molloy; formal analysis, Qinggang Gao; investigation, Qinggang Gao; resources, Kay W. Axhausen and Joseph Molloy; data curation, Kay W. Axhausen and Joseph Molloy; writing—original draft preparation, Qinggang Gao; writing—review and editing, Kay W. Axhausen and Joseph Molloy; visualization, Qinggang Gao; supervision, Kay W. Axhausen and Joseph Molloy; project administration, Kay W. Axhausen and Joseph Molloy; funding acquisition, Kay W. Axhausen and Joseph Molloy. All authors have read and agreed to the published version of the manuscript.

Funding: The project is funded by the Swiss Innovation Agency (Innosuisse) and the Federal Department of the Environment, Transport, Energy and Communications (DETEC).

Data Availability Statement: The data presented in this study are available on request. The data are not publicly available due to project restrictions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results

References

1. Axhausen, K.; Schönfelder, S.; Wolf, J.; Oliveira, M.; Samaga, U. 80 weeks of GPS-traces: Approaches to enriching the trip information. *Transportation Research Record* **2003**, *1870*, 46–54. doi:10.3929/ethz-a-004570614.
2. Nguyen, M.H.; Armoogum, J.; Madre, J.L.; Garcia, C. Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering (English Edition)* **2020**, *7*, 395–412. doi:10.1016/J.JTTE.2020.05.004.
3. Ermagun, A.; Fan, Y.; Wolfson, J.; Adomavicius, G.; Das, K. Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies* **2017**, *77*, 96–112. doi:10.1016/j.trc.2017.01.020.
4. Gong, L.; Morikawa, T.; Yamamoto, T.; Sato, H. Deriving personal trip data from GPS data: A literature review on the existing methodologies. *Procedia - Social and Behavioral Sciences* **2014**, *138*, 557–565. doi:10.1016/j.sbspro.2014.07.239.
5. Meng, C.; Cui, Y.; He, Q.; Su, L.; Gao, J. Towards the Inference of Travel Purpose with Heterogeneous Urban Data. *IEEE Transactions on Big Data* **2019**, pp. 1–1. doi:10.1109/tbdata.2019.2921823.
6. Gong, L.; Kanamori, R.; Yamamoto, T. Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. *Travel Behaviour and Society* **2018**, *11*, 131–140. doi:10.1016/j.tbs.2017.03.004.

7. Yazdizadeh, A.; Patterson, Z.; Farooq, B. Ensemble Convolutional Neural Networks for Mode Inference in Smartphone Travel Survey. *IEEE Transactions on Intelligent Transportation Systems* **2020**, *21*, 2232–2239. doi:10.1109/TITS.2019.2918923.
8. Deng, Z.; Ji, M. Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach. 7th International Conference on Traffic and Transportation Studies; , 2010. doi:10.1061/41123(383)73.
9. Lu, Y.; Zhang, L. Imputing trip purposes for long-distance travel. *Transportation* **2015**, *42*, 581–595. doi:10.1007/s11116-015-9595-0.
10. Janzen, M.; Vanhoof, M.; Axhausen, K.W. Purpose imputation for long-distance tours without personal information. *Arbeitsberichte Verkehrs- und Raumplanung* **2016**, *1181*, IVT, ETH Zurich, Zurich. doi:10.3929/ethz-b-000118790.
11. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer: Berlin, 2015.
12. García, S.; Luengo, J.; Herrera, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems* **2016**, *98*, 1–29. doi:10.1016/j.knosys.2015.12.006.
13. Frenay, B.; Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **2014**, *25*, 845–869. doi:10.1109/TNNLS.2013.2292894.
14. Zhu, X.; Wu, X. Class Noise vs. Attribute Noise: A Quantitative Study. *Artificial Intelligence Review* **2004**, *22*, 177–210. doi:10.1007/s10462-004-0751-8.
15. Brodley, C.E.; Friedl, M.A. Identifying and eliminating mislabeled training instances. Proceedings of the National Conference on Artificial Intelligence; AAAI Press: Portland, Oregon, 1996; pp. 799–805.
16. Brodley, C.E.; Friedl, M.A. Identifying mislabeled training data. *Journal of Artificial Intelligence Research* **1999**, *11*, 131–167.
17. Kotsiantis, S.B.; Kanellopoulos, D.; Pintelas, P.E. Data preprocessing for supervised learning. *International Journal of Computer Science* **2006**, *1*, 111–117.
18. Lakshminarayan, K.; Harp, S.A.; Samad, T. Imputation of missing data in industrial databases. *Applied intelligence* **1999**, *11*, 259–275.
19. Batista, G.E.; Monard, M.C. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence* **2003**, *17*, 519–533.
20. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Computers and Electrical Engineering* **2014**, *40*, 16–28. doi:10.1016/j.compeleceng.2013.11.024.
21. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
22. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **2010**, *36*, 1–13. doi:10.18637/jss.v036.i11.
23. Hanchuan Peng.; Fuhui Long.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2005**, *27*, 1226–1238. doi:10.1109/TPAMI.2005.159.
24. Ling, C.X.; Li, C. Data mining for direct marketing: Problems and solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* **1998**, *98*, 73–79.
25. Kubat, M.; Matwin, S. Addressing the curse of imbalanced training sets: one-sided selection. *Fourteenth International Conference on Machine Learning* **1997**, *97*, 179–186.
26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321–357.
27. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); IEEE: Hong Kong, China, 2008.
28. Montini, L.; Rieser-Schüssler, N.; Horni, A.; Axhausen, K. Trip purpose identification from GPS tracks. *Transportation Research Record* **2014**, *2405*, 16–23. doi:10.3141/2405-03.
29. Shen, L.; Stopher, P.R. A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies* **2013**, *36*, 261–267. doi:10.1016/j.trc.2013.09.004.
30. Oliveira, M.; Vovsha, P.; Wolf, J.; Mitchell, M. Evaluation of two methods for identifying trip purpose in GPS-based household travel surveys. *Transportation Research Record* **2014**, *2405*, 33–41. doi:10.3141/2405-05.
31. Xiao, G.; Juan, Z.; Zhang, C. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies* **2016**, *71*, 447–463. doi:10.1016/j.trc.2016.08.008.
32. Montini, L.; Rieser-Schüssler, N.; Axhausen, K.W. Personalisation in multi-day GPS and accelerometer data processing. 14th Swiss Transport Research Conference (STRC 2014), Ascona, 2014.
33. Gong, L.; Yamamoto, T.; Morikawa, T. Comparison of activity type identification from mobile phone GPS data using various machine learning methods. *Asian Transport Studies* **2016**, *4*, 114–128. doi:10.11175/eastsats.4.114.
34. Feng, T.; Timmermans, H.J. Detecting activity type from GPS traces using spatial and temporal information. *European Journal of Transport and Infrastructure Research* **2015**, *15*, 662–674. doi:10.18757/ejtir.2015.15.4.3103.
35. Liao, L.; Fox, D.; Kautz, H. Extracting places and activities from GPS traces using hierarchical conditional random fields. *International Journal of Robotics Research* **2007**, *26*, 119–134. doi:10.1177/0278364907073775.
36. Li, A.; Huang, Y.; Axhausen, K.W. An approach to imputing destination activities for inclusion in measures of bicycle accessibility. *Journal of Transport Geography* **2020**, *82*, 102566. doi:10.1016/j.jtrangeo.2019.102566.

37. Garnett, R.; Stewart, R. Comparison of GPS units and mobile Apple GPS capabilities in an urban landscape. *Cartography and Geographic Information Science* **2015**, *42*, 1–8. doi:10.1080/15230406.2014.974074.
38. Molloy, J.; Castro Fernández, A.; Götschi, T.; Schoeman, B.; Tchervenkov, C.; Tomic, U.; Hintermann, B.; Axhausen, K.W. A national-scale mobility pricing experiment using GPS tracking and online surveys in Switzerland. Response rates and survey method results, 2020. doi:10.3929/ethz-b-000441958.
39. Molloy, J.; Tchervenkov, C.; Schatzmann, T.; Schoeman, B.; Hintermann, B.; Axhausen, K.W. MOBIS-COVID19/25. Results as of 19/10/2020 (post-lockdown), 2020. doi:10.3929/ethz-b-000447684.
40. Schlich, R.; Schönfelder, S.; Hanson, S.; Axhausen, K. Structures of Leisure Travel: Temporal and Spatial Variability. *Transport Reviews* **2004**, *24*, 219–237. doi:10.1080/0144164032000138742.
41. Stauffacher, M.; Schlich, R.; Axhausen, K.W.; Scholz, R.W. The diversity of travel behaviour: motives and social interactions in leisure time activities. *Arbeitsberichte Verkehrs-und Raumplanung* **2005**, *328*, IVT, ETH Zürich, Zürich.
42. Wu, X.; Kumar, V. *The top ten algorithms in data mining*; CRC Press, 2009.
43. Kowarik, A.; Templ, M. Imputation with the R Package VIM. *Journal of Statistical Software* **2016**, *74*, 1–16.
44. Ward Jr, J.H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **1963**, *58*, 236–244.
45. Lance, G.N.; Williams, W.T. A general theory of classificatory sorting strategies: 1. hierarchical systems. *The Computer Journal* **1967**, *9*, 373–380. doi:10.1093/comjnl/9.4.373.
46. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and regression trees*; CRC press, 1984.
47. Song, Y.Y.; Ying, L.U. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry* **2015**, *27*, 130.
48. Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123–140. doi:10.1007/bf00058655.
49. Breiman, L. Out-of-bag estimation. *Technical report, Department of Statistics: University of California, Berkeley*. **1996**.
50. Breiman, L.; Cutler, A. Manual—setting up, using, and understanding random forests V4. 0, 2003.
51. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. doi:10.1007/s11749-016-0481-7.
52. Khoshgoftaar, T.M.; Golawala, M.; Van Hulse, J. An empirical study of learning from imbalanced data using random forest. *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)* **2007**, *2*, 310–317.
53. Abellán, J.; Mantas, C.J.; Castellano, J.G.; Moral-García, S. Increasing diversity in random forest learning algorithm via imprecise probabilities. *Expert Systems with Applications* **2018**, *97*, 228–243. doi:https://doi.org/10.1016/j.eswa.2017.12.029.
54. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R news* **2002**, *2*, 18–22.
55. Pang, S.I.; Gong, J.z. C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering - Theory & Practice* **2009**, *29*, 94–104. doi:10.1016/s1874-8651(10)60092-0.
56. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
57. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C.C.; Lin, C.C.; Meyer, M.D. Package ‘e1071’. *The R Journal* **2019**.
58. Friedman, J.H. Multivariate Adaptive Regression Splines. *The Annals of Statistics* **1991**, *19*, 1–67.
59. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—an S4 package for kernel methods in R. *Journal of Statistical Software* **2004**, *11*, 1–20.
60. Branco, P.; Ribeiro, R.P.; Torgo, L. UBL: an R package for utility-based learning. *arXiv preprint arXiv:1604.08079* **2016**.
61. del Río, S.; López, V.; Benítez, J.M.; Herrera, F. On the use of MapReduce for imbalanced big data using Random Forest. *Information Sciences* **2014**, *285*, 112–137. doi:https://doi.org/10.1016/j.ins.2014.03.043.
62. Guan, D.; Yuan, W. A Survey of mislabeled training data detection techniques for pattern classification. *IETE Technical Review* **2013**, *30*, 524–530. doi:10.4103/0256-4602.125689.