Peacox – Persuasive Advisor for CO2-reducing cross-modal trip planning

Project Reference: 288466

FP7-ICT 2011: 6.6 Low carbon multi-modal mobility and freight transport)

Project Duration: 1 Oct 2011 – 30 Sept 2014

D 4.3

Implementation and evaluation of learning routines for mode and trip purpose detection

ETHZ

Author(s)

Lara Montini (ETHZ)

Nadine Rieser (ETHZ)

Date: 25/06/2014

Dissemination level: (PU, PP, RE, CO): PU

Project Co-Funded by the European Commission within the 7th Framework Program

Table of Contents

1. Introduction

This deliverable provides an overview over the current status of the work in Work Package 4 "Automated Travel Mode and Trip Purpose Detection". The deliverable is focused on Tasks 4.4 "Design, implementation and testing of the prompted-recall survey" and 4.5 "Development and implementation of learning routines". The goal of these tasks is to first collect annotated data of users and then use this knowledge to increase the accuracy of the mode as well as the trip purpose detection routines.

For field trial 1 annotations were collected using a paper-and-pen travel diary. This will be replaced in field trial 2 by a smartphone application providing a map to visualize the daily activities and allowing to correct the transport mode as well as the activity type. The prompted recall tool is described in detail in Section 2.

For trip purpose detection two aspects of personalization are shown, first, the effect of person-specific features is shown. In that respect most important are the distance to home and work, which implies that a person's home and work locations have to be either asked for or learned from the data. Second, it is shown that classification is improved if corrected data of a person is included when training the classifier. This implies that survey participants should annotate some of their data.

For transport mode identification most improvement is achieved if car, bike and public transport mode shares are calculated for each person from annotated data. These shares are then introduced as features when learning the classifier.

2. Prompted Recall Tool

Having corrected travel diaries is extremely valuable as it can be used to improve the quality of the automatically produced diaries, as shown in the next section. To get these corrected annotations from the survey participants a prompted recall tool is needed. The tool should help participants recalling their diary, that is transport modes and activity types, by presenting them GPS tracks in combination with the automatically generated diary. Most commonly prompted recall surveys are web-based as shown in Auld et al.(2009) and as implemented by us for a GPS-based survey in Zurich (Montini et al., 2013).

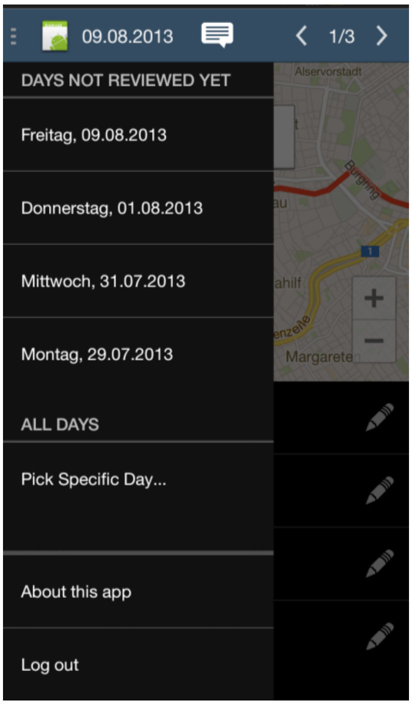
For the second PEACOX field trial, our goal is to create a clearly laid out and simple to use user interface. As PEACOX is a smartphone application, all participants have and can work with these phones. Therefore, the prompted recall tool was developed for smartphones as well. The application was jointly planned by ETHZ, CURE and Telematix, whereas Telematix developed the application.

Screenshots of the app are shown in Figure 1. First, the day which is corrected has to be selected. This is done in the menu (Figure 1 a). All days with data that have not been corrected are listed in the menu. To finish corrections the checkbox “I have reviewed this day” has to be checked in the main screen (Figure 1 b).

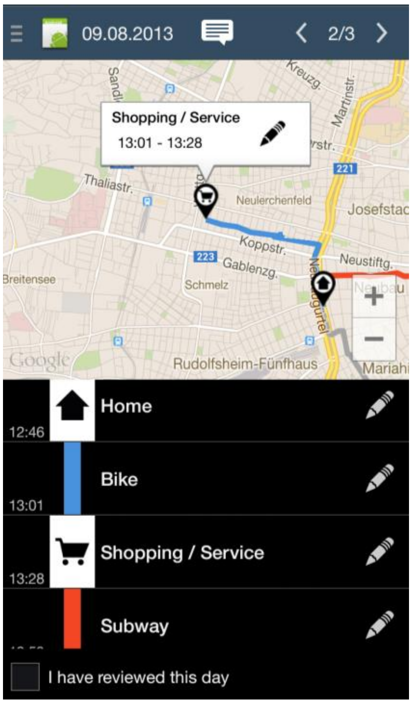
The main screen of the prompted recall tool shows the map as well as the automatically generated travel diary as list (Figure 1 b), which also includes editing possibilities for transport mode (Figure 1 c) and activity type. The different transport modes are represented by different colors and all activity types have distinct icons. This helps to quickly understand the diary when looking at the map. Colors and icons also help to create the link between the map and the diary on the bottom. On the top bar on the right, the number of activities is shown, and the possibility is given to quickly go through them by clicking on the arrows. The map is then automatically zoomed to the current activity. If no data is available for a day a checkbox “I stayed home all day” is shown instead of the diary list.

Users can also leave a comment (Figure 1d) for every day they correct. This is important because if something is unclear, users have the opportunity to write it down. The app shows very private data and is therefore login protected, the login is the same as for the PEACOX navigation client.

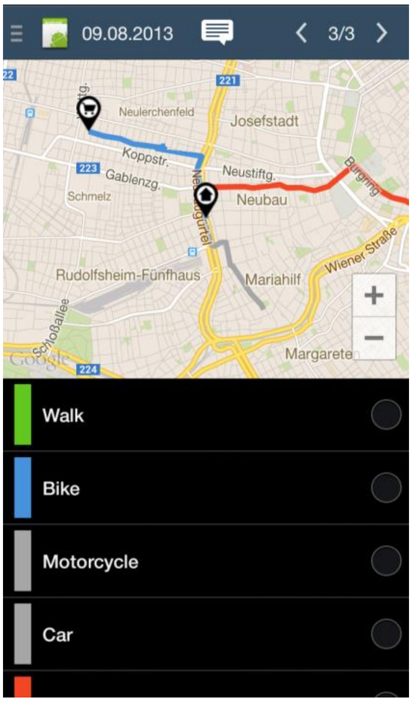
The application is currently tested by the consortium members.



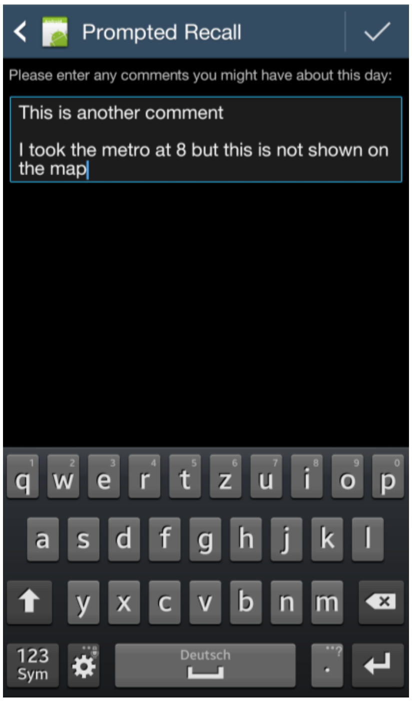
(a) Menu



(b) Main screen



(c) Main screen edit transport mode



(d) Comment screen

Figure 1 Screenshots of the prompted recall application

3. Personalization and learning in trip purpose and mode identification

This section is based on Montini et al. (2014b).

3.1 Method

The method of trip purpose detection is described in detail in Montini et al. (2014a) as well as in Deliverable 4.2. For completeness, the most important topics are repeated here. One difference is that compared to the previously mentioned paper the random forest implementation we use is FastRandomForest based on the WEKA data mining tool (Hall et al., 2009) instead of the Matlab version (TreeBagger). The Java implementation is included in our open source GPS data and accelerometer processing framework (POSDAP, 2012).

The employed method is based on multi-day GPS and accelerometer data for survey respondents living in the same region. To exploit the multi-day nature of the data, activities are clustered into locations using hierarchical clustering. Clusters are created for single persons but also for the complete set of activities as several respondents might frequent the same public locations. Classification variables, called features, are then derived for location clusters.

3.1.1 Data Set And Features Selection

The GPS data set used for evaluation was collected in and around the city of Zurich in Switzerland and is described in more detail in Montini et al. (2013). Each of the 156 respondents collected approximately one week of second-by-second GPS and accelerometer data using a dedicated GPS device. Respondents were randomly selected from an address pool and they are also reasonably representative for the study area. Respondents were asked to correct an automatically generated travel diary including transport mode and trip purpose on the survey homepage. As the quality of these corrections varies, all were double checked by the survey team. Since the participants were asked for their home and work addresses, the trip purposes being home, working and mode transfer could mostly be imputed by survey personnel. All other purposes are only available if respondents filled in the diaries.

In total, 6938 valid activities including a 28.54 % share of mode transfers were reported. The other surveyed purposes are being home (22.33 %), working (11.82 %), shopping and services (9.61 %), recreational activities (9.77 %), picking up or dropping off someone (2.77 %), business, that is, work-related activities outside work place, (3.32 %) and finally other activities (2.62 %). The remaining 9.22 % of activities have no reported type. Therefore, this data was left out when training and testing the classifier. For the clustering on the other hand, this data are used, as even if the type is unknown, the location and duration provide additional information.

For the mode detection, 6990 stages with corrected transport mode are available. The reported modes are: driving by car (41.9 %), walking (30.0 %), biking (8.7 %), going by train (8.1 %), taking a tram (5.4 %) or a bus (4.3 %) and using another mode (1.6 %).

For trip purpose detection, around 40 potential features were tested and of those seventeen were selected based on a feature importance analysis (Montini et al., 2014a). The features are specific to persons, activities and clusters (per person as well as per data set) and are listed in Table 1. Home and work locations to calculate distances were asked for in the questionnaire, but these could also be learned from the data. The transport modes as well as the split in activities and stages is not imputed. The walk duration percentage is calculated using accelerometer data, which allows to detect walk with high probability.

Person-based	cat	Activity-centered	cat	Cluster-specific
Age	–	Duration	–	Mean duration
Education	✓	Start time	–	Standard deviation durations
Income	✓	Day of week	–	Occurrences per surveyed day
Marital status	✓	Walk duration percentage	–	Percentage of weekdays
		Distance to home	–	Number of persons per cluster
		Distance to work	–	
		Arrival transport mode	✓	
		Leaving transport mode	✓	

Table 1 Features used for trip purpose detection. Categorical features are marked in the cat column, the cluster-specific features are all numeric.

Table 2 shows the features used for mode detection. The mode-specific features are mostly the same as for our previously used fuzzy rule system (Rieser-Schüssler et al., 2011). Additionally, stage-centered variables such as duration and quality of the GPS signal (points per second) are used. For personalization, person-based features are needed, the same socio-demographic attributes as used for trip purpose are available. But more interesting is previously collected and corrected data from which walk and bike speeds can be calculated as well as the shares of car, bike and public transport (bus, tram, rail, metro) stages. The accelerometer measure is the moving window standard deviation of the accelerometer length, which is, according to the feature importance measure of the random forest, the most important feature.

Stage-centered	Mode-specific	Person-based
Duration	Median speed	Mean walk speed
Start time	95^{th}-percentile speed	Mean bike speed
Number of GPS points per second	Standard deviation speed	Public transport share
	Median accelerometer measure	Bike share
	95^{th}-percentile accelerometer measure	Car share
	Public transport network matching score	

Table 2 Features used for transport mode detection.

3.1.2 Classifier: Random Forests

For trip purpose as well as for transport mode imputation, decision trees, that is a set of rules learned by a machine and executed in a given order, were already used with good success (e.g. Griffin and Huang, 2005; Deng and Ji, 2010; Lu et al., 2012). Using a random forest, that is a set of decision trees also called ensemble of trees, is therefore, a natural step. Random forests were introduced by Breiman (2001). The underlying concept is comprehensible and, more importantly, it performs well in a variety of problems. They are also very popular, as they are easy to train and tune (Hastie et al., 2009). Breiman (2001) showed that random forest do not overfit even if more trees are added. A further advantage is that good results can be maintained even if data are missing, as they are estimated internally (Breiman and Cutler, 2013).

Technically, random forests work as follows. Each decision tree in the ensemble has one vote that counts for classification. The class with most votes, is the classification result. In a regular decision tree at each step the data set is split using the feature that results in the best split. Therefore, using the same data, results in the same tree. But, in a random forest different votes are needed, and correlation between trees should be reduced to obtain best classification. To achieve that, on the one hand, each tree is learned from a different subset of the training data. On the other hand, at each split in the tree a random subset of features is considered. Each tree is fully learned, that is, splits will be created until all training data are correctly classified.

The tuning parameters of random forests are, therefore, the number of features m that are randomly selected when deciding on the best split as well as the number of trees per forest. In Montini et al. (2014a), it is shown that for trip purpose detection the differences in accuracy are very small when varying m. In the case of 17 features used, the best results are obtained with $m = 7$. Therefore, this value is used for trip purpose. For mode detection, the default values given by $m = floor(\sqrt{nr\ features})$ are used. For all analyses, 200 trees are learned per random forest, as they provide good results in reasonable time.

3.1.3 Evaluation: Per Person

The main application of automated GPS travel diaries with correction or confirmation data are travel surveys. So when using a classifier to classify one participant's data, it will be based on previously collected data of other people. To simulate this situation the analysis presented in the following section is done per person. When evaluating classifiers it is crucial that two different subsets are used: one, the training data to learn the classifier, and two, the test data set to measure performance.

In the context of this Deliverable, if not stated otherwise, creation of the training and test data sets is based on a selection of persons and not on a selection of individual observations (that is activities for trip purpose detection and stages for mode detection). In particular, the test data set always contains data of one person only. In some cases, a subset of the person's data is incorporated in the training data and only the remaining activities or stages are used for testing. In all cases, 10 different classifiers are learned per person, to asses stability of predictions. The reported accuracy of one person's data is the mean of 10 runs.

3.2 Results Trip Purpose Identification

The main focus of the results lies on trip purpose detection. Some results for mode detection are presented in the last subsection.

First, base runs using all 17 features (Table 1) are evaluated to establish how well random forests perform per person. For the personalization strategies presented in Subsection 3.2.2 subsets of the available data have to be used to generate different classifiers. Subsets per definition consist of less data, therefore, the base scenario is run for different numbers of persons in the training data set to show the influence of its size on classification accuracy. In Figure 2, results clearly shown that classification is better the more data is used. The slope starts to flatten which indicates that the used data set is just about big enough to get a realistic estimation of how well purposes are detectable.

The base scenario results already include some personalization as socio-demographic variables are used. Their effect is shown in Subsection 3.2.1. Next, in Subsection 3.2.2, different strategies are tested how the knowledge of already annotated personal data can be included.

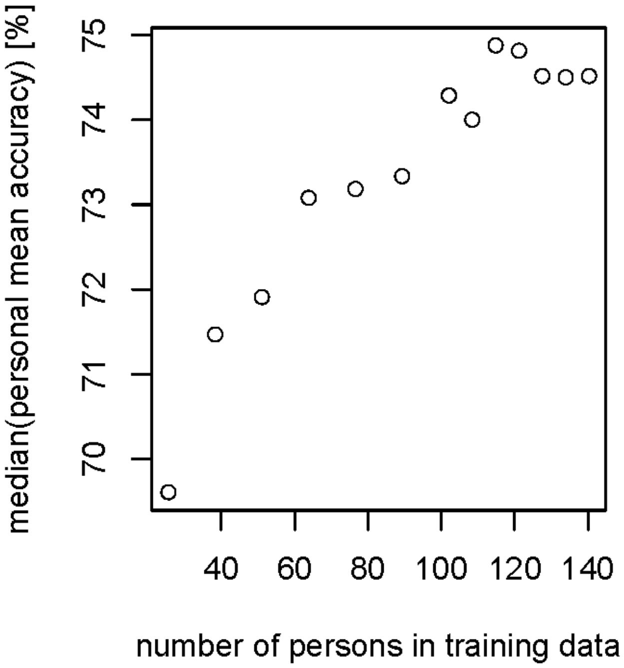


Figure 2 Median accuracy of the per person mean accuracy for different number of persons in the training set

3.2.1 Personalization Using Person-Specific Input Features

To show the effect of person-specific input features in the base scenario, a classifier is learned excluding all person-based features as well as the distances to home and work location. These distances are included in the second scenario and, finally, the socio-demographic attributes are added in the third scenario. To create 10 different classifiers for each person, all other persons are randomly split into 10 groups, for the different training sets needed in each case one of these groups is omitted. Results are illustrated in Figure 3. The mean of the accuracy of the 10 validation runs varies among participants between 25.3 % and 100 % with a median of the mean accuracies of the classifier without person-based data of 71.1 %. Including the person-specific features improves the results to 74.9 %. The per person standard deviation of accuracy are very similar for all runs. For the scenario with all features the mean is around 2.3 % but goes up to 8.7 % for the person with highest variation within the 10 runs.

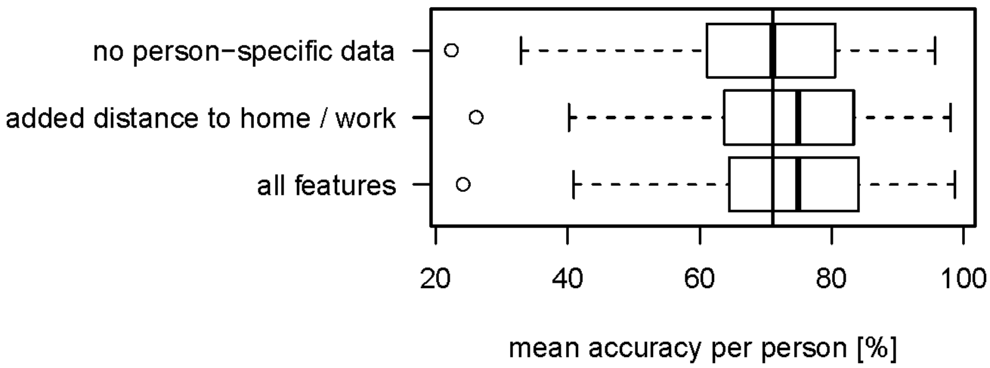


Figure 3 Distribution of mean accuracies per person for different feature sets in trip purpose detection.

It has to be considered that the number of reported days, the validity of corrections and the trip purposes vary amongst participants. The split of trip purposes has probably the biggest influence on the spread of accuracies. As reconstruction of the diary is easier if a participant is e.g., only at home and at work. This influence is also illustrated in Figure 4 which depicts the mean accuracy per person versus the per person share of the three best predicted purposes, that is: being home, working (or studying) and changing the mode.

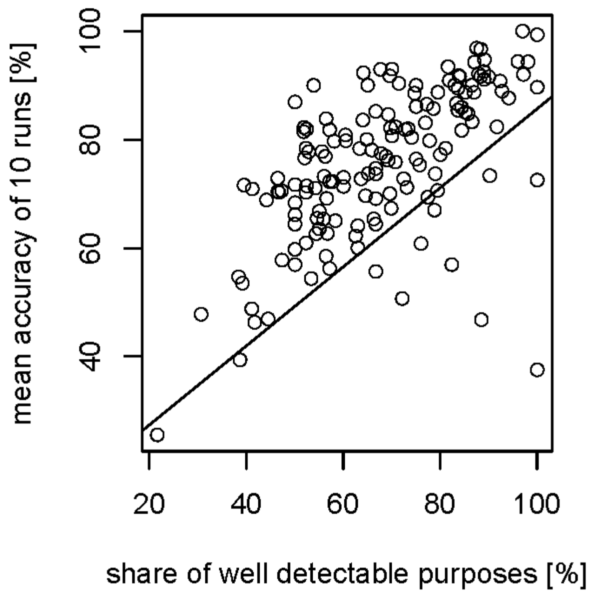


Figure 4 Mean accuracy of 10 runs for each person plotted against the share of easiest detected trip purposes

3.2.2 Personalization Based On Corrected Data

For trip purpose detection 4 strategies to personalise classification based on data corrected by participants are tested:

i. Select best: selection of one classifier out of many based on performance on a subset of a person's data
ii. Group: group participants first and learn a different classifier for each group
iii. Include person data: include some of the person's data when learning the classifier
iv. Overrule: overrule the classifier when the location is already known

All strategies are subsequently described in more detail, but it is shown in Figure 5 that only inclusion of personal data improves predictions.

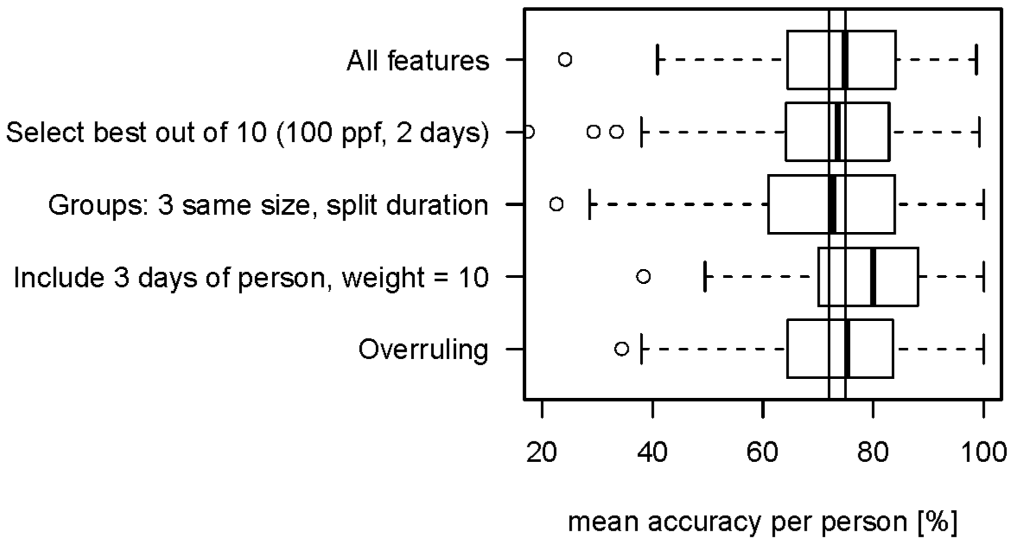


Figure 5 Distribution of mean accuracies per person for all strategies. The vertical lines are the medians of the base runs with 50 (left) and 130 (right) persons respectively.

In the *select best* approach, for every person 10 random forests are evaluated on a subset of the person's data (selector data). The classifier that performs best on the selector data is then used to produce the results on the test data. The underlying hypothesis is, that some classifiers work better for one person's data then others and that those are consistently better on this person's data. To create 10 different classifiers to select from, each classifier is learned on different person-subsets consisting of 100 persons. The selector data is fixed between the runs and consists of two randomly picked days. The mean standard deviation of the 10 test runs of the base scenario with around 100 persons is 3.2 % therefore, there is some potential to select a better classifier. But as shown in Figure 4 accuracies are not increased. Furthermore, it was tested if only selecting weekdays would yield better results, assuming they contain more information, but it did not.

The idea behind *grouping* participants is that some of them have more similar diaries than others, hence, a classifier built using similar persons should be more successful. Three different criteria for grouping were tested. First, participants are grouped based on all socio-demographic properties using hierarchical clustering. Then, for this first approach, classifiers are learned without the socio-demographic features. Second, participants are split into three groups of the same size based on the mean duration of all their activities. Finally, 86 participants were grouped as ’mostly using car’, 35 as ’mostly using public transport or bike’ and 32 as ’using both’. Neither of the groupings had any effect on the accuracies. In Figure 5, the results of the groups based on activity duration are shown. At first it looks like grouping even decreases accuracy, but it has to be considered that only 50 persons are used per group to learn the classifier. Hence, comparing it to the base classifier with 50 persons (left vertical line) shows that the grouping just does not influence results.

Including personal data when learning the classifier is straight forward. The person’s data is split into a test and a training set. The training set consists of a given number of days that are randomly selected. As one persons’s data is not enough to learn a classifier, all other person’s observations are added to the training set without special weighting. Results are compared for 1 randomly selected day and 3 randomly selected days. Whereas using 1 day does not improve classification, using 3 days increases the median accuracy to 80.0 %. Besides the number of days to be selected for training also the weights of the person’s data was varied but did not have a relevant effect as shown in the next section for the mode detection.

To implement *overruling* each person’s data is also split into test and training set. The classifier is learned on all training data (including the person’s). But when classifying the test data, it is checked whether the person’s training data includes an activity that was clustered into the same location. If this is the case the random forest is overruled and trip purpose is set to the one of the activity in the training set. If the training test set contains several activities at the same location with different purposes the purpose of the activity with the most similar duration is selected. Overall, overruling performs worse than the random forest. In total, 36627 classifications were made, of those almost 50 % (17671) are overruled and 79 % of those overrules were not necessary, 8 % were not helpful that is both the overrule and the random forest predict different but wrong purposes and 9 % of the overrules are counter-productive that is the random forest is correct. Especially home and mode transfer points are falsely corrected. Only 4 % (648) of the overrules are correct.

3.3 Results Mode Identification

To show the influence of different features mode identification is done for different feature sets. Opposed to the per person analysis in the trip purpose detection, training and test sets for the mode detection are built from random subsets of all stage observations. To be precise a 100-fold cross validation is performed, the results of which are shown in Figure 6.

The first subset includes the mode- and stage-specific features that are often used when performing mode detection. This run serves as minimum base line. The goal of adding the public transport-map-matching score is to improve distinction between car and public transport stages. In Figure 6 it can be seen that overall accuracy improves only slightly, but when looking at the confusion matrices of the two experiments it can be seen that the percentage of bus stages correctly identified is increased from 20 % to 37 %. Therefore, the general effect is as expected, but the recall of bus stages is still very low.

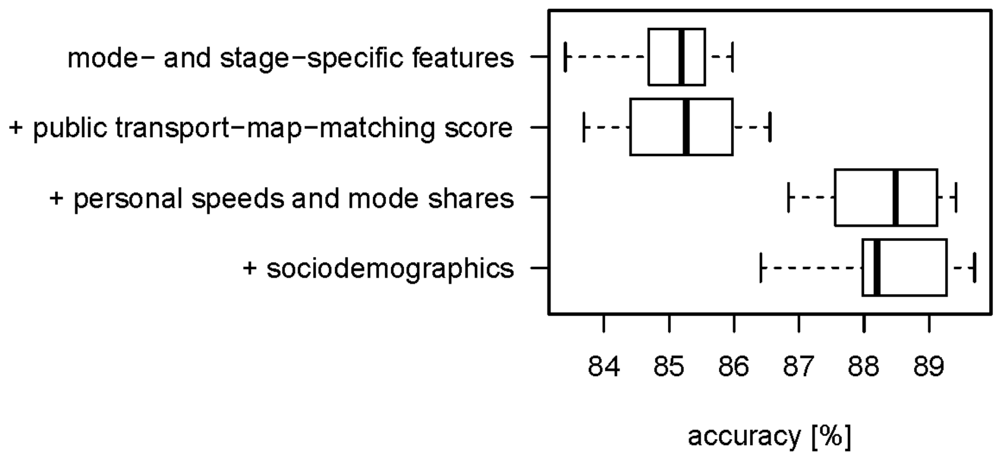


Figure 6 Accuracies for different feature sets, box plot for 100 runs each

Adding the personal mode shares computed from the available one week of data as well as the mean bike and walk speeds of a person improves the classification with a median accuracy over 88 %. Importantly, also bus stages are better detected with a recall of 55 % (see confusion matrix in Table 3). Half of the misclassified bus stages are wrongly identified as car. However, more unexpectedly most of the other errors are bus stages misclassified as walk.

		Prediction							
		Walk	Bike	Bus	Tram	Rail	Car	Other	Recall
Truth	Walk	1899	26	5	13	39	108	4	90.7%
	Bike	73	504	4	0	2	22	0	83.3%
	Bus	40	7	165	10	10	67	1	55.0%
	Tram	38	3	7	300	19	12	0	79.2%
	Rail	42	1	2	5	478	35	4	84.3%
	Car	95	8	18	7	18	2780	5	94.8%
	Other	14	0	0	5	1	37	57	50.0%
	Precision	86.3%	91.8%	82.1%	88.2%	84.3%	90.8%	80.3%	88.5%

Table 3 Confusion matrix for run number 3 including person-based features

Further adding socio-demographic information has a negative effect on classification and is therefore not useful for mode identification.

For the best performing feature set, the feature importance measure for all used features is given in Table 4. The accelerometer measure is by far the most important followed by the mode shares that are observed for a person. Personal bike and walk speeds on the other hand are not that important and can therefore be neglected.

Median accelerometer measure	19.37
Bike share	4.22
Car share	3.92
Median speed	3.63
GPS quality (number of coords per second)	2.96
95th percentile speed	2.65
Public transport share	2.37

Duration	1.79
Public transport network-matching score	0.99
95th percentile accelerometer measure	0.96
Standard deviation of walk speeds	0.41
Start time	0.40
Mean of walk speeds	0.24
Mean of bike speeds	0.19
Standard deviation of speed	0.13
Standard deviation of bike speed	0.00

Table 4 Random forest out of bag feature importance measure, blue-coloured features are person-based, mode-specific features are light purple and orange features are stage-specific

For the same feature set, a per person analysis is presented in Figure 7, where it can be seen that not only for trip purpose but also for mode identification the variation between participants is high. For one person only 30 % of the predicted modes are correct and for other persons 100 % are correct.

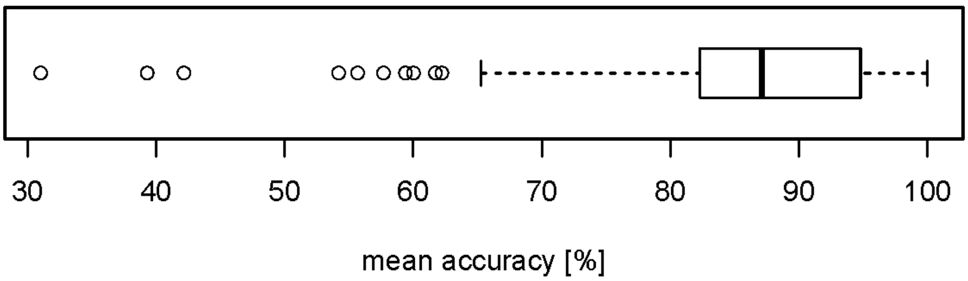


Figure 7 Per person mean accuracy of 10 runs

Since including data previously collected by a person has a positive effect on trip purpose detection of this person it was also tested for transport mode classification. Unfortunately, no effect could be shown compared to classification without including this person's data (Montini et al., 2014b).

3.4 Implementation for PEACOX

As it was shown above, the distance to home and work locations are important but these locations are not known for PEACOX users, these two locations are learned as soon as possible. First, if corrected data is available, the location most often annotated as home and as work are saved as that person's home and work location. If GPS data was collected but no corrections are available, a classifier that does not use distance to home and work is used to classify all activities, the locations predicted to be home and work are then used to extract an approximation of these two locations. Using these approximations, distance to home and work can be calculated and classification is run again using a classifier that takes advantage of distance to home and work.

The classifiers are learned new every day incorporating three data sources: First, the data set collected in Zurich (~7000 observations), second, data collected in field trial I (425 observations) and, third, all data that is collected and corrected during field trial II, which is the reason why the classifier has to be updated on a daily basis. As including a person's previously collected data did not improve classification for mode detection significantly, a daily update is not needed. Hence, the mode detection classifier is learned only once. In the field trials, participants are asked about their mode share, as this input proved to be important. The participants' estimation should be used as long as not enough corrected data is available.

4. Conclusion and Outlook

The base scenario shows that quite a lot of data is necessary to achieve good results for trip purpose detection in general. For a classifier learned on data of 20 persons, which are approximately 100 person days, the median accuracy is around 4 % lower than for a classifier learned on 100 persons. The base scenario also shows that including the distance to home and work is important (3.8 % increase of median accuracy).

The main conclusion of this deliverable is, that it is worth collecting annotated data from participant's in a multi-day or even multi-week survey, as a median accuracy of 80.0 % is achieved if three days of annotated personal data are included in the training set. This corresponds to an increase in accuracy of 5.5 % compared to the base scenario. If the processing of the collected data is done after the survey, this is straight forward. For continuous processing during surveys, the classifier should be updated whenever newly corrected data is available.

Interestingly, this strategy did not work for mode detection, but it was shown that using the mode shares of participants as input feature the mean classification accuracy is increased by 3 %.

All other personalization strategies tested did not have an effect on accuracies. Grouping participants seemed like a good idea, but in essence when thinking in rule-based systems, this is just adding another rule at the beginning of the decision process. This contradicts the idea of decision trees, where the best possible split at any point is found automatically. Instead of grouping people according to a new variable, probably the easiest and most successful way is to add it to the feature set and make sure that it is not counterproductive. The goal when selecting the best classifier out of many is that for each person a random grouping is found that performs better than an average classifier. However, first results were not promising. Maybe more classifiers with higher diversity would be necessary, but to achieve that, more training data is needed. A similar approach, that could be tested, is to use subsets of activities instead of creating classifiers from a subset of persons.

5. References

Auld, J., C. Williams, A. K. Mohammadian and P. Nelson (2009) An automated GPS-based prompted recall survey with learning algorithms, *Transportation Letters*, **1** (1) 59–79.

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

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.

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.

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.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten (2009) The weka data mining software: An update, *SIGKDD Explorations*, **11** (1) 10–18.

Hastie, T., R. Tibshirani and J. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. edn., Springer Series in Statistics, Springer, New York.

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.

Montini, L., N. Rieser-Schüssler and K. W. Axhausen (2013) Field Report: One-Week GPSbased Travel Survey in the Greater Zurich Area, paper presented at the *13th Swiss Transport Research Conference*, Ascona, April 2013.

Montini, L., N. Rieser-Schüssler, A. Horni and K. W. Axhausen (2014a) Trip Purpose Identification from GPS Tracks, paper presented at the *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2014.

Montini, L., N. Rieser-Schüssler and K.W. Axhausen (2014b) Personalisation in multi-day GPS and accelerometer data processing, paper presented at the *14th Swiss Transport Research Conference*, Ascona, May 2014

POSDAP (2012) Position Data Processing, webpage, sourceforge.net/projects/posdap/.

Rieser-Schüssler, N., L. Montini and C. Dobler (2011) Improving automatic post-processing routines for GPS observations using prompted-recall data, paper presented at the *9th International Conference on Survey Methods in Transport*, Termas de Puyehue, November 2011.