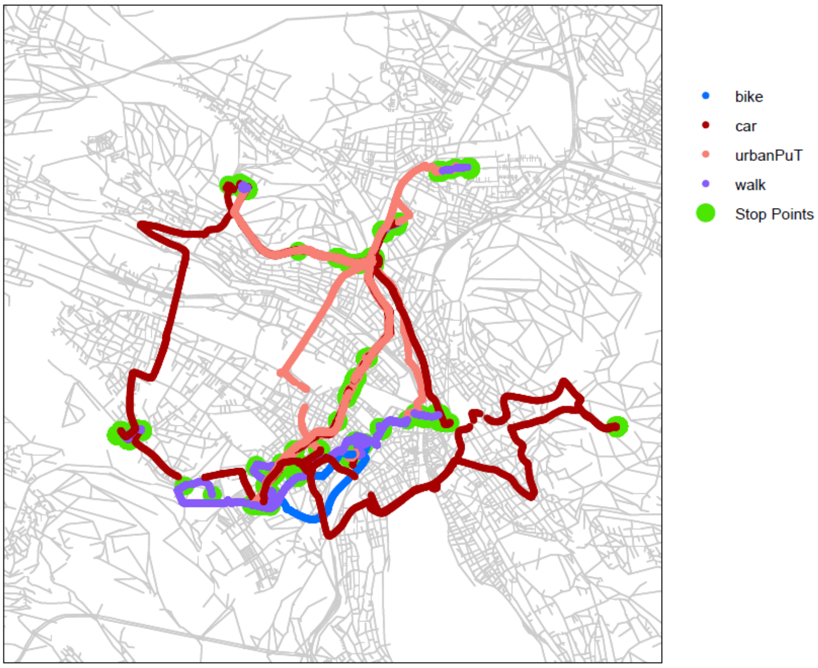


Personalisation in multi-day GPS and accelerometer data processing

Lara Montini

Nadine Rieser-Schüssler

Kay W. Axhausen

Transport and Spatial Planning

May 2014

STRC | 14th Swiss Transport Research Conference
Monte Verità / Ascona, May 14 – 16, 2014

Transport and Spatial Planning

Personalisation in multi-day GPS and accelerometer data processing

Lara Montini
IVT
ETH Zürich
CH-8093 Zürich
phone: +41-44-633 30 88
fax: +41-44-633 10 57
lara.montini@ivt.baug.ethz.ch

Nadine Rieser-Schüssler
IVT
ETH Zürich
CH-8093 Zürich
phone: +41-44-633 30 85
fax: +41-44-633 10 57
nadine.rieser@ivt.baug.ethz.ch

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

May 2014

Abstract

In this paper, we analyse how personalisation during processing of multi-day GPS and accelerometer data can improve the quality of the produced travel diaries. The main focus is on trip purpose detection using random forests. Two main approaches are followed. First, the effect of person-based input features is shown, in particular distance to home and work improve classification result (median accuracy + 3.8 %). Second, it is analysed how usage of annotated data improves prediction. Most strategies like selecting the best classifier out of many, have no effect. But, improvements are possible if the classifier is learned including some of the participant's annotated data (median accuracy + 5.5 %).

Keywords

GPS processing, trip purpose, personalisation

1 Introduction and related work

In transportation research, GPS tracks often in combination with accelerometer data are used amongst other data sources to reconstruct diaries automatically, hence, complementing or replacing traditional travel diaries. One major goal is to use these diaries to reduce the burden of respondents which would also allow to survey longer time periods. More precisely, the prepared diaries as well as a visualisations of GPS tracks support people in recollecting their daily schedule. Ideally, participants only need to confirm the diary they are presented. In practice, when conducting a GPS-based survey (Montini et al., 2013) we made the experience that it was participant-dependent how well the daily schedule was recognised. Concerning the reconstructed diary, we showed at least for trip purpose detection that the accuracy that is the share of correctly detected purposes also highly varies between participants (Montini et al., 2014). Consequently, the potential reduction in response burden varies as well.

The goal of this paper is to analyse if and how personalisation during processing of multi-day GPS and accelerometer data can improve the quality of the produced travel diaries, mainly focusing on trip purpose detection. For the analysis we consider applications for GPS processing where some of the available data is annotated. This is for example the case for travel surveys, where participants can be asked to at least correct some of their schedule. Having annotated data is not granted, as position data is more and more often collected as a side product, for example on smartphones as part of a navigation application.

In the literature, two main groups of imputation routines can be found both for trip purposes and mode detection. First, there are rule-based systems. For trip purposes they rely mostly on the position of the activity, its timing, and land use data (e.g. Moiseeva et al., 2010; Bohte and Maat, 2009; Stopher et al., 2008; Wolf et al., 2001) . For mode detection the rules use speed measures, stage criteria such as duration and also proximity to e.g. bus stops or roads (e.g. de Jong and Mensonides, 2003; Stopher et al., 2005; Chung and Shalaby, 2005; Bohte and Maat, 2008; Marchal et al., 2011). As a next step fuzzy logic approaches were employed that also account for the fact that the boundaries between modes are overlapping (Tsui and Shalaby, 2006; Rieser-Schüssler et al., 2011).

Second, there are machine learning approaches. For trip purpose detection the focus is more on the activity itself and less on position. Most commonly used are decision trees (e.g. Oliveira et al., 2014; Lu and Zhang, 2014; Deng and Ji, 2010; Griffin and Huang, 2005). But also other approaches are tested, Oliveira et al. (2014) compare a decision tree and a nested multinomial logit model. Liao et al. (2007) achieve good results using hierarchical conditional random fields. Lu and Zhang (2014) compare three algorithms on two datasets: decision tree, support vector machine and metalearner. And in Montini et al. (2014) we employ random forests. Machine learning is also used for mode detection e.g. Stenneth et al. (2011) use random forests with good success and Zheng et al. (2008) as well as Moiseeva et al. (2010) use Bayesian inference models.

The remainder of this paper is structured as follows. First, the underlying dataset and the classifier (random forests) are introduced as well as the evaluation method. In the results section different approaches to personalisation are tested. To conclude results are interpreted and an outlook on future work is provided.

2 Method

The method of trip purpose detection is described in detail in Montini et al. (2014), 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.

2.1 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. The very first activity of the survey period is removed for each participant, because – due to the cold start problems and unclear first handling of the device – the GPS signal can be very far off the actual location.

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 an other 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., 2014). 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.

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

person-based	cat	**activity-centered**	cat	**cluster-specific**
age	-	duration	-	mean duration
education level	✓	start time	-	standard deviation of 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 2: Features

stage-centered	mode-specific	person-based
duration	median speed	mean walk speed
start time	95th-percentile speed	mean bike speed
(distance to public transport stop)[2]	standard deviation speed	
number of GPS points per second	median accelerometer measure	
	95th-percentile accelerometer measure	

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 as well as person-based speeds are used. The accelerometer measure is the moving window standard deviation of the accelerometer length, and according to the feature importance measure of the random forest it is the most important feature.

2.2 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, and is a trademark of 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 a data set is split using the feature that results in the best split. Using the same data to learn a tree, 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. (2014) 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 $m = floor(\sqrt{nrfeatures})$ are used. For all analyses 200 trees are learned per random forest, as they provide good results in reasonable time.

2.3 Evaluation: Per Person

The main application of automated GPS travel diaries where correction or confirmation data is available 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.

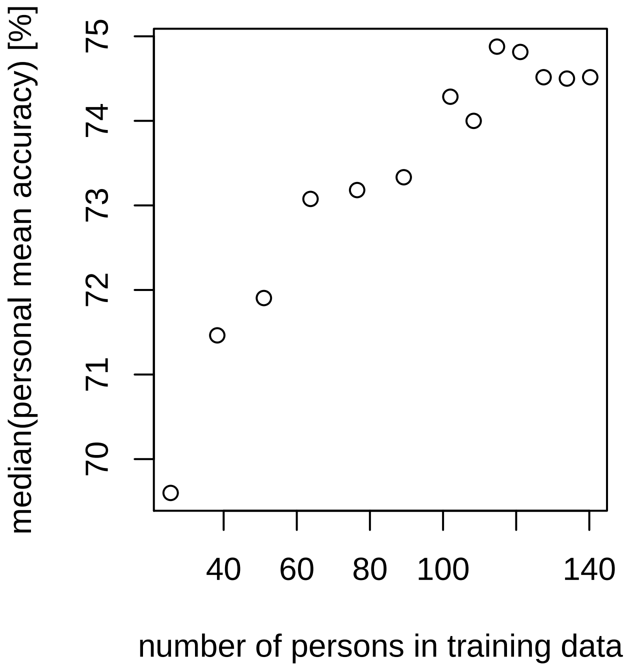
In the context of this paper creation of the training and test data sets is based on a selection of persons and not on a selection of individual observations (that is activities for trip purpose detection and stages for mode detection). In particular, the test data set always contains data of one person only, in some cases a subset of the persons'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, the reported accuracy of one person's data is therefore the mean of 10 runs.

3 Results

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 personalisation strategies, presented in the second part of this section, subsets of the available data is used to generate different classifiers. To be able to compare it to this base scenario it is run for different numbers of persons in the training data set. In Figure 1 it is clearly shown that classification is better the more data is used. The slope starts to flatten, therefore, the used data set is just about big enough to get a realistic estimation of how well purposes are detectable.

Figure 1: Median accuracy of the per person mean accuracy for different number of persons in the training set



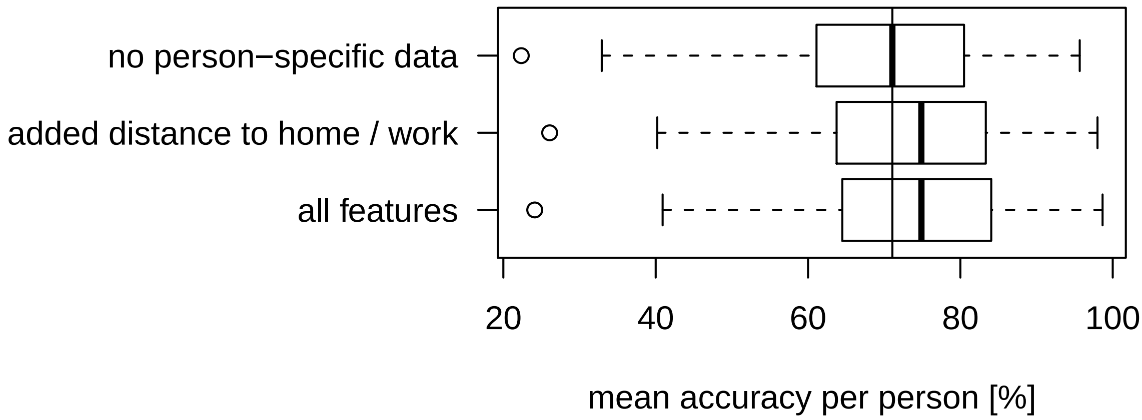
These base scenarios already include some personalisation as socio-demographic variables are included, their effect is shown in the first subsection. Next, different strategies are tested how the knowledge of already annotated personal data can be included.

3.1 Personalisation Using Person-Specific Input Features

To show the effect of person-specfic input featues a classifier is learned excluding all person-based features as well as 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, the 10 training sets then consists of 9 of these 10 groups. Results are illustrated in Figure 2. 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 %, inclusion of 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 2: Distribution of mean accuracies per person for different feature sets in trip purpose detection.



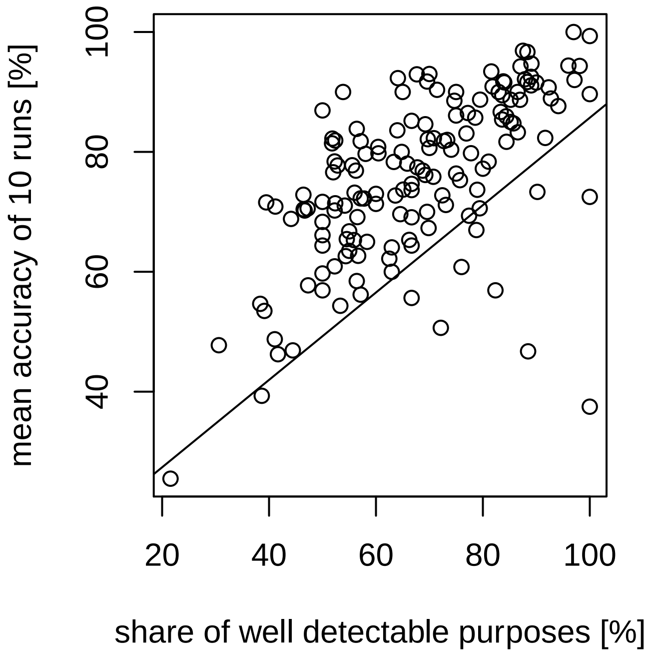
It has to be considered that the number of reported days, the validity of corrections and the trip purposes varies amongst participants. The split of trip purposes has probably the biggest influence on the spread of accuracies. As reconstruction of the diary is easier if a participant is e.g., only at home and at work. This influence is also illustrated in Figure 3 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.

3.2 Personalisation 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

Figure 3: Mean accuracy of 10 runs for each person plotted against the share of easiest detected trip purposes

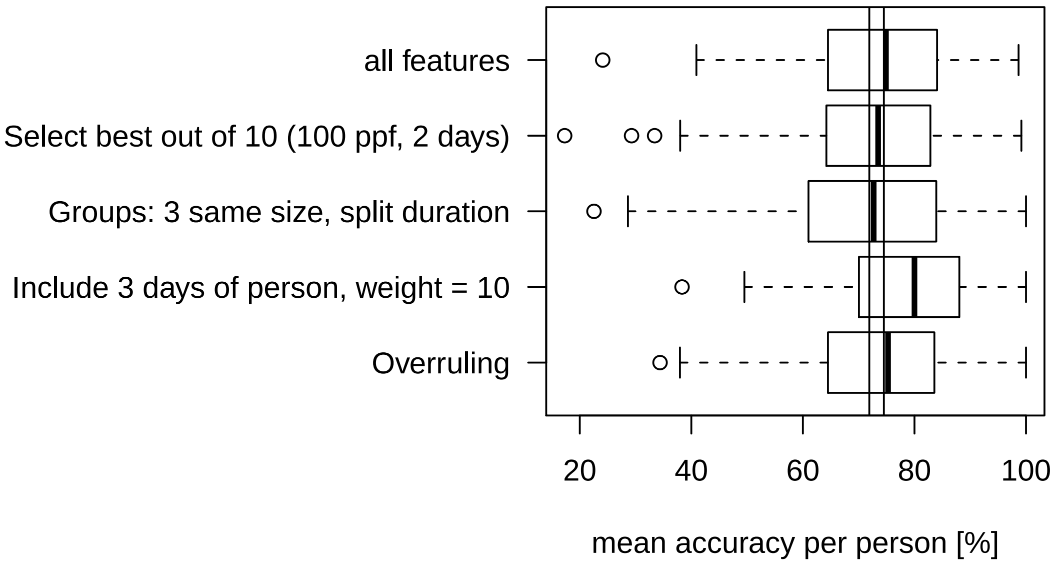


All strategies are subsequently described in more detail, but it is shown in Figure 4 that only inclusion of personal data improves predictions.

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 other classifiers and that this classifier is 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. First, participants are grouped based on all socio-demographic properties using a hierarchical clusterer. The classifiers are then learned without the socio-demographic features. Classification success is the same as

Figure 4: Distribution of mean accuracies per person for all strategies. The vertical are the medians of the base runs with 50 (left) and 130 (right) persons respectively.



when not grouping. Second, participants are split into three groups of the same size based on the mean duration of all their activities. And 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 4 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. All other persons are added to the training set. Results are compared for 1 randomly selected day and 3 randomly selected days where 1 day does not improve classification, 3 days on the other hand increases the median 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.

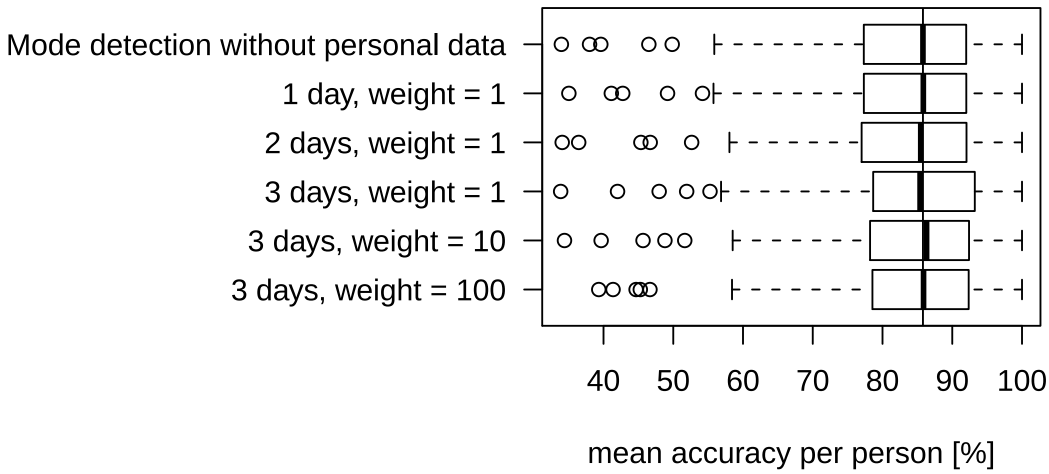
To implement *overruling* the persons's data is also split into test and training set. The classifier is learned on all training data (including the person's). But when classifying the test data, it is checked whether the person's training data includes an activity that was clustered into the same location. If this is the case the random forest is overrule and trip purpose is set to the one of the activity in the training set. If 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 of those 79 % were not necessary, 8 % are not helpful that is both the overrule and the random forest predict different but wrong purposes. 9 % of the overrules are counter-productive that is the random forest is correct, especially home and mode transfer points are falsely corrected. And only 4 % (648) of the overrules are correct.

3.3 Personalisation For Transport Mode Detection

As for trip purposes the performance of the mode detection is analysed for different feature sets. Compared to using only the mode-specific features (Table 2) adding the stage-centered features improves the median of classification from 83.9 % to 85.8 %. Adding the person-based features has no effect on accuracies.

Since including person data has a positive effect on trip purpose detection it is also tested for transport mode classification. The number of days that are included in the training data is varied between 1 and 3 and the weights given to the person's training data is also varied (1, 10, 100). Unfortunately, compared to classification without including a person's data, there is no effect at all, as can be seen in Figure 5.

Figure 5: Mean accuracies of 10 runs per person including personal training data. Varying number of training days of the person to be classified as well as different weights.



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 paper 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 personal data are included in the training set, this corresponds to an increase of 5.5 % compared to the base scenario. If 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.

All other personalisation strategies tested in this paper 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. To conclude, outsmarting the machine did not work. 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 counter productive.

To select the best classifier out of many the hope is that for each person a random grouping is found that performs better than an average classifier. First results are 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.

For mode detection, the next step is to add more person-dependent features, such as mobility tool ownership. Another step in direction of learning would be to impute the mode for a group of matched trips. Trip matching might be especially interesting for more sparse data, where gaps could be filled with knowledge from similar trips.

To conclude, one week of smartphone data is not enough to highly personalise trip purpose detection routines. Especially the variance between participants is still very high and therefore, per person analysis of automatically processed data is a bit problematic. The next step is to apply the classifiers learned here in a 6 week smartphone study that will be soon conducted in Vienna as well as Dublin. This study will allow us to analyse transferability of the classifier, that is how well it performs in a different survey context and in different countries. Further, the 6-week survey period might be sufficient to do personalisation based on weekly-rhythms.

5 Acknowledgements

This research was funded by the European Union as part of the project "Peacox - persuasive advisor for CO2-reducing cross-modal trip planning" within the Seventh Framework Programme (FP7).

6 References

Bohte, W. and K. Maat (2008) Deriving and validating trip destinations and modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands, paper presented at the *8th International Conference on Survey Methods in Transport*, Annecy, May 2008.

Bohte, W. and K. Maat (2009) Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands, *Transportation Research Part C: Emerging Technologies*, **17** (3) 285–297.

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.

Chung, E.-H. and A. Shalaby (2005) A trip bases reconstruction tool for GPS-based personal travel surveys, *Transportation Planning and Technology*, **28** (5) 381–401.

de Jong, R. and W. Mensonides (2003) Wearable GPS device as a data collection method for travel research, *Working Paper*, **ITS-WP-03-02**, Institute of Transport Studies, University of Sydney, Sydney.

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.

Liao, L., D. Fox and H. Kautz (2007) Extracting places and activities from GPS traces using hierarchical conditional random fields, *International Journal of Robotics Research*, **26** (1) 119–134.

Lu, Y. and L. Zhang (2014) Trip purpose estimation for urban travel in the u.s.: Model development, NHTS add-on data analysis, and model transferability across different states, paper presented at the *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2014.

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.

Marchal, P., J.-L. Madre and S. Yuan (2011) Post-processing procedures for person-based GPS data collected in the French National Travel Survey 2007-2008, paper presented at the *90th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2011.

Moiseeva, A., J. Jessurun and H. J. P. Timmermans (2010) Semi-automatic imputation of activity travel diaries using GPS-traces, prompted recall and context-sensitive learning algorithms, *Transportation Research Record*, **2183**, 60–68.

Montini, L., N. Rieser-Schüssler and K. W. Axhausen (2013) Field Report: One-Week GPS-based 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 (2014) Trip Purpose Identification from GPS Tracks, paper presented at the *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2014.

Oliveira, M., P. Vovsha, J. Wolf and M. Mitchell (2014) Evaluating two methods for identifying trip purpose in GPS-based household travel surveys, paper presented at the *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2014.

POSDAP (2012) Position Data Processing, webpage, `http://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.

Stenneth, L., O. Wolfson, P. S. Yu and B. Xu (2011) Transportation mode detection using mobile phones and GIS information, paper presented at the *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Chicago, November 2011.

Stopher, P. R., C. FitzGerald and J. Zhang (2008) Search for a global positioning system device to measure person travel, *Transportation Research Part C: Emerging Technologies*, **16** (3) 350–369.

Stopher, P. R., Q. Jiang and C. FitzGerald (2005) Processing GPS data from travel surveys, paper presented at the *2nd International Colloqium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications*, Toronto, June 2005.

Tsui, S. Y. A. and A. Shalaby (2006) An enhanced system for link and mode identification for GPS-based personal travel surveys, *Transportation Research Record*, **1972**, 38–45.

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.

Zheng, Y., L. Liu, L. Wang and X. Xie (2008) Learning transportation mode from raw GPS data for geographic applications on the web, paper presented at the *17th World Wide Web Conference*, Beijing, April 2008.