Future mobility demand estimation based on socio-demographic information: A data-driven approach using machine learning algorithms

Maximilian Held

Lukas Küng

Emir Çabukoglu

Giacomo Pareschi

Gil Georges

Konstantinos Boulouchos

May 2018

STRC | 18th Swiss Transport Research Conference
Monte Verità / Ascona, May 16 – 18, 2018

Contents

List of Figures

List of Tables

Abbreviations

AIC Akaike Information Criterion
CI confidence interval
DT Decision Tree
DTD daily traveled distance
DTDC daily traveled distance by car
EV electric vehicle
HTS Household Travel Survey
ICDF Inverse Cumulative Distribution Function
mkm millions of kilometers
MLE Maximum Likelihood Estimation
MSE mean squared errors
O-D origin-destination
PCA principal component analysis
PHS Population and Households Statistics
RC Random Choice
SSE sum of squared errors
vkm vehicle kilometers

Future mobility demand estimation based on socio-demographic information: A data-driven approach using machine learning algorithms

Maximilian Held, Lukas Küng, Emir Çabukoglu, Giacomo Pareschi, Gil Georges, Konstantinos Boulouchos
Institute of Energy Technology
ETH Zurich
Sonneggstrasse 3, 8092 Zurich
phone: +41-44-632 03 53
fax: +41-44-632 11 02
held@lav.mavt.ethz.ch

May 2018

Abstract

Estimations of the future mobility demand are highly valuable for policy makers, transportation planners and the automotive industry. Knowing mobility patterns allows for targeted and optimized decarbonization of the transport sector. This work provides a model order reduction approach for clustering mobility demand according to characteristic population groups that share similar travel behavior. Using Swiss household travel survey data and machine learning algorithms, the methodology developed in this paper allows for extrapolating future mobility demand based on socio-demographic information.

Keywords

household travel survey, model order reduction, machine learning, clustering techniques, decision tree, mobility patterns, decarbonization

1 Introduction

1.1 Context of the decarbonization of transport

The transport sector was responsible for 20.5% of global CO_2-eq. emissions in 2014 (International Energy Agency, 2014). According to the Kyoto system boundary, i.e. excluding international aviation and maritime transport, the Swiss transport sector even accounted for 32.0% of the national CO_2-eq. emissions in 2015 (Federal Office for the Environment, 2017). Thus, the transport sector plays a crucial role in the transition to an ecologically sustainable society. In the Paris Climate Agreement 2015, the participating nations have expressed their commitment to substantially reduce their greenhouse gas emissions. To guide the evolution in the desired direction, strategic planning is necessary.

From an abstract perspective, a decrease of greenhouse gas emissions can be either achieved by reducing demand (vehicle kilometers (vkm)) or specific vehicle emissions (CO_2 per vkm). While the first depends on behavior, the latter is mainly driven by technology. The total emissions result as a product of the two terms; zero emissions can only be reached if one term goes to zero. As mobility is linked to productivity and crucial to sustain a functioning society and economy, it has to be technological options to cut down the emissions to zero in a long-term perspective. Changes in behavior will just gain time to postpone technological decarbonization.

Within this study, we address the technological aspect of decarbonization, in particular of Swiss passenger cars. We do not analyze mode choice or traffic flows. A rather lean description of mobility demand in form of the daily traveled distance (DTD) is sufficient. The DTD needs to be known when assessing the potential of different drivetrain technologies to decarbonize transport. It is a measure to represent energy consumption and autonomy range, i.e. the range capacity without recharging or refueling, of a vehicle. One result of this study is a lean method to derive these DTDs using Household Travel Survey (HTS) data. The investigated hypothesis is whether it is possible to reduce mobility demand to a few explaining attributes which are robust over time, ideally in a parameterized form.

1.2 Importance of daily trip distance distribution of a fleet for energy system analysis

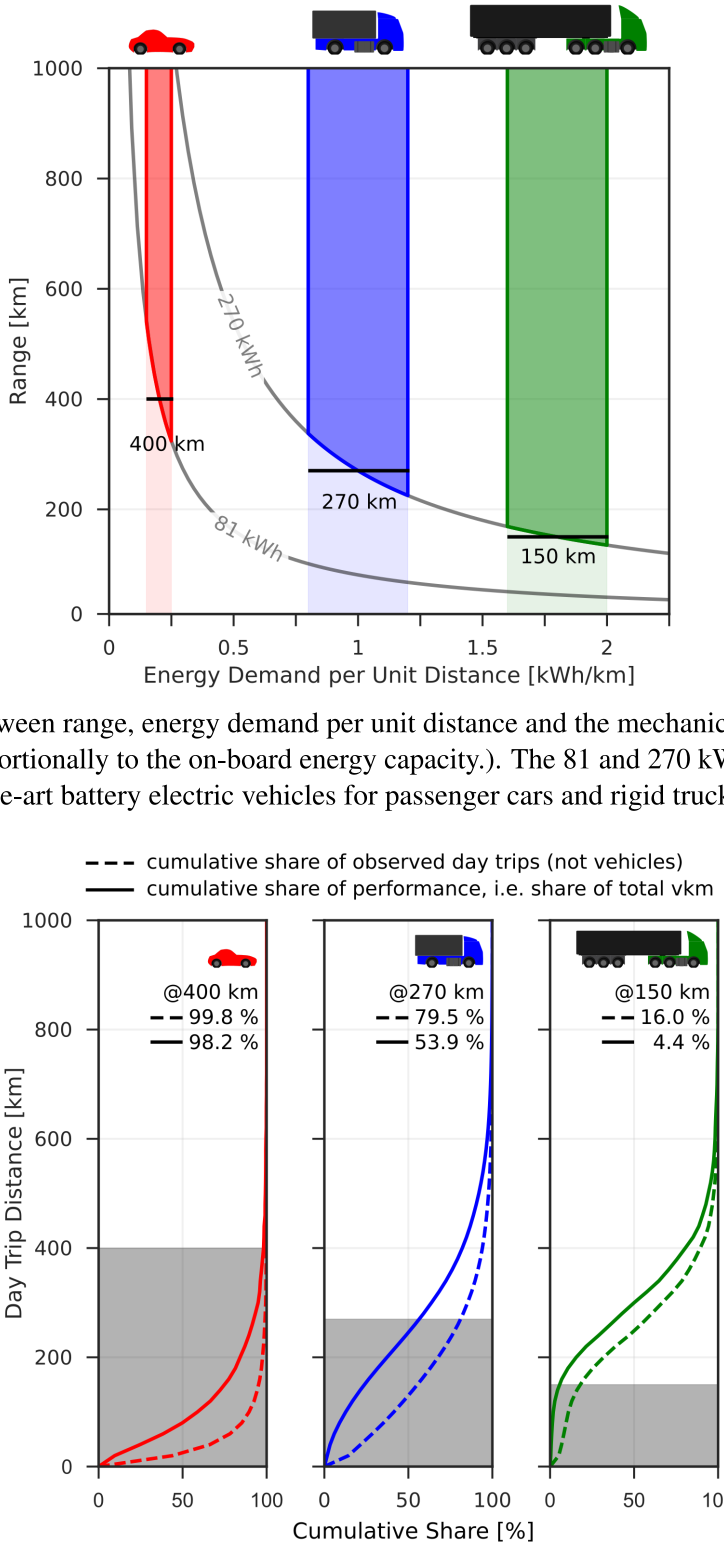
Reasonable substitutions of conventional transport technologies need to be able to provide the same service, i.e. transporting a person or good from one point to another. There is the possibility to change the mode of transport, e.g. from car to train. However, if sticking to individual mobility, technologies will not be accepted if the usage pattern cannot remain the same. Tamor *et al.* (2015) showed this for electric vehicles: They defined a threshold of inconvenience as "the number of days per year that a given electric vehicle (EV) range was insufficient for that day's driving". Below that threshold, customers accept EVs and switch to alternative technologies for the few days where EVs cannot provide the customers' mobility demand; above the threshold, they refuse to use EVs (Tamor *et al.,* 2015). An optimal substitution of conventional technologies should not force the user to change the behavior for any day of the year.

From a technological perspective, not every drivetrain is a feasible technology for every mobility demand pattern or every vehicle. In order to assess the potential of different technologies to decarbonize transport on a national level, a bottom-up assessment of disaggregated mobility demand is necessary. A common approach to define such disaggregated demand is to analyze DTDs of all vehicles in a fleet. Given these DTDs and the technologies' autonomy ranges, first estimations on the decarbonization potential of different vehicle types can be drawn[1]: Figure 1(a) relates on-board energy demand for propulsion and range for different vehicles, in a simplified but reasonable manner. It shows required technology specifications for desired autonomy ranges. The two hyperboles (81 and 270 kWh) represent the state-of-the-art in battery electric vehicles for passenger cars and rigid trucks respectively, whereas there is no commercially available electrified articulated truck yet. The red/ blue/ green area indicate typical values for the specific energy demand of passenger cars/ rigid trucks/ articulated trucks. The intersection of the mean values of these areas with the hyperbolic curves result in the autonomy range of each vehicle type, i.e. the range that can be driven without recharging.

If related to the daily usage of these vehicle types (Figure 1(b)) we identify (I) the amount of day trips which are feasible for EVs given a certain day trip range, and (II) the cumulative share of performance (vkm) of day trips. The latter reflects the relative impact of trips on carbon emissions, and thus, on their decarbonization potential. Taking the rigid trucks as an example, 79.5% of all day trips can be electrified[2], given the autonomy range of 270 km (taken from Figure1(a)). However, in contrast to the found 79.5%, these trips account only for 53.9% of the overall driving performance. This is due to a large number of short trips that can be electrified but that do not contribute to decarbonization substantially. As mass and volume for energy storage is limited on a vehicle, the electrification of transport is in particular a challenge for long-distance trips. For freight transport, the challenge is even more severe since time for recharging or refueling during a trip will directly have negative impact on the profitability of the vehicle.

[1] As the results are based on an average day, we neglect intra-week and seasonal changes in mobility demand.

[2] The electrification potential of day trips sets a theoretical upper decarbonization limit. A single vehicle is executing multiple day trips over a year. In order to electrify the respective vehicle, all day trip distances needed to be below the autonomy range of 270 km.



(a) Relationship between range, energy demand per unit distance and the mechanical energy supplied to the wheels (proportionally to the on-board energy capacity.). The 81 and 270 kWh lines represent the state-of-the-art battery electric vehicles for passenger cars and rigid trucks respectively.

(b) Frequencies of observed day trip distances and the respective performance. The distribution for passenger cars is based on the HTS 2010 (Federal Statistics Office, 2010a), the distribution for trucks on data of the Federal Customs Administration (2017). The grey areas correspond to the ranges indicated in Figure 1(a) and illustrate the electrification potential.

Figure 1: Trade-off of energy demand per km and autonomy ranges of EVs for different vehicle types and distribution of day trip distances

The Swiss Federal Statistics Office collects mobility data for planning purposes, most notably the Swiss HTS. However, when looking into the decarbonization of a vehicle fleet, it is necessary to account for the change in demand over time. Future changes in population and spatial structure which can affect the distribution in daily distance and total performance have to be considered. Extrapolating perceived trends for the future mobility demand becomes necessary on a disaggregated level to allow for a bottom-up feasibility assessment of technologies.

1.3 Related work

An extensive body of literature is dedicated to both estimation of future mobility demand and decarbonizing transport. Our work differentiates from these studies by three main aspects: (1) we look at the mobility demand from a technology-centered view. (2) we use socio-demographics estimations for the future to extrapolate future mobility demand. (3) We use machine learning algorithms to cluster Swiss mobility demand.
The related literature to each aspect is provided in the following:

1. **Technology perspective:**
   There is comprehensive literature on a detailed geographical perspective on mobility demand to draw conclusions on social behavior or to model traffic flows for infrastructure planning. For that, mobility demand is usually derived from (a) agent-based models, (b) accessibility measures, or (c) origin-destination matrices.
   (a) An example of an integrated assessment of the transport and energy sector is given by Bauer *et al.* (2016). The study compares different drivetrain technologies for an average vehicle. It does not comprise a whole fleet of vehicles. The mobility demand is derived out of a status-quo agent-based modeling.
   (b) The Swiss Federal Office for Spatial Development provides mobility demand perspectives (Mathys *et al.,* 2016). Their model accounts for socio-demographic attributes and presents different scenarios resulting in aggregated activities for different regions in Switzerland until 2040. Mobility demand is modeled using origin-destination (O-D) matrices. Similarly, Saadi *et al.* (2017) use random forest classification to estimate an O-D matrix.
   (c) Loder *et al.* (2017) analyze HTS data to draw conclusions on the accessibility to reach potential destinations, in order to explain differences in the annual passenger car mileage of households.

   In contrast to the these three approaches, we provide a lean representation of daily vehicle usage profiles that is suited to assess technological feasibility of alternative drivetrains for observed vehicles. We do not focus on detailed geographic-resolved traffic flows which are not necessary for deriving decarbonization potential of alternative drivetrain technologies. In contrast to the work of Bauer *et al.* (2016), we do not only look at the ecological impact of a single vehicle type, but of the whole fleet, clustered into mobility archetypes.

2. **Socio-demographic extrapolation perspective:**
   Charleux (2018) derive mobility archetypes from HTS data using the k-means algorithm. After clustering, they analyze sociodemographic characteristics of the found clusters. In contrast to this approach, we use socio-demographic features as clustering characteristics to derive mobility archetypes. Based on demographics estimations, we intend to extrapolate the future mobility demand.
3. **Swiss perspective:**
   Most literature on clustering HTS data via machine learning algorithms to infer mobility archetypes use the HTS data set of the United States (Charleux, 2018, Diana, 2012, Mohammadian and Zhang, 2007, Pirra and Diana, 2016). Within this work, we do this analysis for Switzerland.

Table 1: Available data in both, the Swiss HTS and the PHS

Demographic data	
Nationality	*Binary variable:* Swiss, Foreigner
Age	*Continuous variable*
Gender	*Categorical variable:* Female, Male, Other
Marital Status	*Categorical variable:* Single, Married, Widowed, Divorced, Other

Geographic/ spatial data	
Municipality Population Density	*Continuous variable*
Aggregated Municipality Characteristic	*Categorical variable:* Urban agglomeration, Suburban agglomeration, Isolated, Rural
Extensive Municipality Characteristic	*Categorical variable:* Central, Suburban, Peri-Urban, Rural, Touristic, Industrial, Mixed agricultural, Agricultural, High income
Swiss Big Regions	*Categorical variable:* Zürich, Tessin, Lake Geneva Region, North-West/ West/ Central/ East Switzerland

2 Data

This work is based on two databases provided by the Swiss Federal Statistics Office (2010a,b): the Population and Households Statistics (PHS) and the Swiss Household Travel Survey (HTS). The PHS provides annual data on size and composition of the whole Swiss population. The HTS contains detailed information on the travel behavior of Swiss citizens. Since 2000, Switzerland has conducted such HTS in a 5 years-turnus.

As the PHS does only contain disaggregated numbers since 2010, this year is chosen for introducing our methodology. The HTS 2010 comprises 55'052 participants aged over 18. To get an intuition on the HTS data, we provide some statistical analysis of the data set, see Figures 7 and 8 (Appendix).

For the methodic approach introduced in the next section, only variables existing in both data sets are chosen, see Table 1. Additionally, the daily travel time and distance are extracted from HTS data.

3 Methodology

The goal of this study is to provide a method to describe mobility demand as function of socio-demographic indicators that is capable to extrapolate future mobility demand. This requires classification of the Swiss population with respect to its travel behavior. We propose a two-step approach consisting of clustering HTS data and sampling onto PHS data: We cluster the HTS participants into characteristic population groups according to similar travel behavior (I). To infer these findings onto the whole population, we sample the people of the HTS clusters onto the corresponding people of PHS that share the same socio-demographic and spatial/ geographic characteristics (II).

The methodic approach of this paper is displayed in Figure 2. The following description gives an overview on the methodology. Afterwards we provide detailed information on the single procedure steps in the subsections.

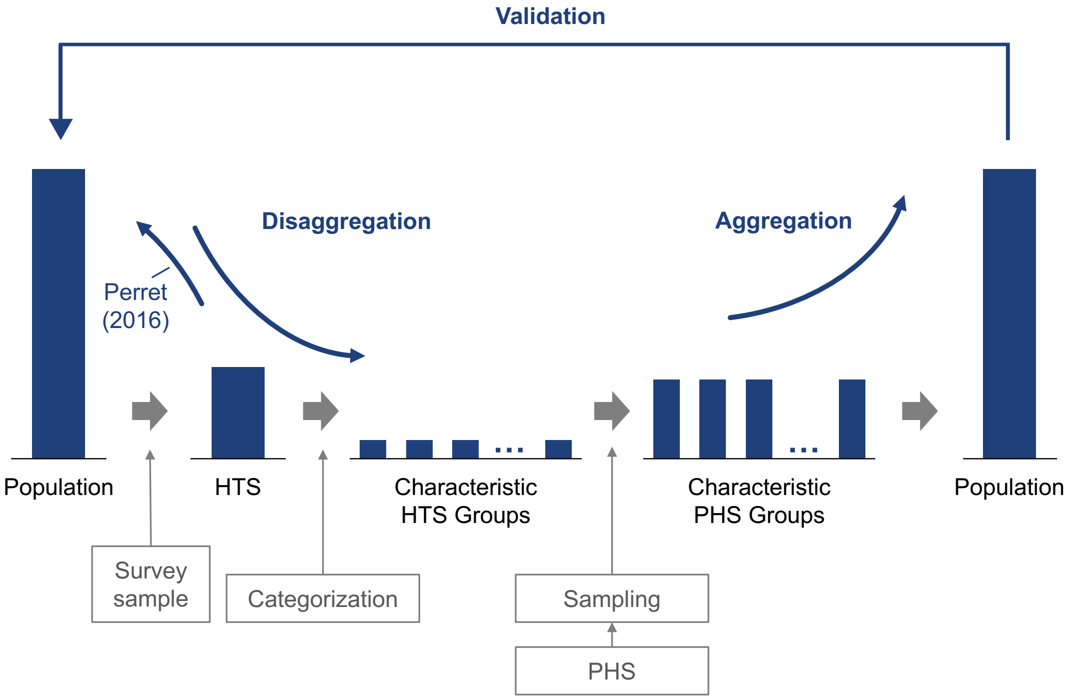


Figure 2: Methodic Approach for Data Processing

We categorize the HTS participants with respect to their daily travel distance or time using machine learning algorithms. Both include all modes of travel. For the explanation of our methodology, we use the DTD as target vector in the following.

The classification results in clusters of people sharing the same travel behavior, i.e. similar DTDs. Since we focus on car mobility, we extract the values for daily traveled distance by car (DTDC) for each cluster in a next step. The resulting DTDC for the characteristic HTS groups are then sampled onto the whole population, matching the descriptive determinants for each cluster of HTS data with the PHS data. In other words, we look up the number of people from the whole population that share the same descriptive determinants as the HTS groups found during clustering (like age, nationality, etc.). Then, we assign a DTDC value from the HTS clusters (55'052 people) to each person from PHS (6'416'153 people) using different sampling techniques. This way, every person of the whole population is assigned with a specific DTDC. Summing up these distances for all people results in the total annual car mileage for the whole population. This aggregated value can be validated by the official number that is drawn from HTS by the Federal Statistical Office (Perret, 2016).

As Figure 2 shows, this approach allows for conclusions on DTDC on a disaggregated level, e.g. from a cantonal perspective. Hence, we also had a look at spatially disaggregated cantonal DTDCs. Here, the validation assessment was done via the (arithmetic average) sum of squared errors (SSE) and the weighted average SSE. The SSE is defined as follows:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

with y_i as the reference value for canton i and $\hat{y}_i$ as the predicted value of y_i. The weighted SSE weights the SSE of each canton according to the number of samples (s) within it:

$$wSSE = \frac{1}{s} \sum_{i=1}^{n} s_i \, (y_i - \hat{y}_i)^2 \tag{2}$$

Table 2 gives an overview on methodology variations to optimize the process. We varied clustering techniques as well as their input feature vector and target variable. Furthermore, we had a look at the impact of different sampling techniques. Given the result of the clustering and sampling, we generalized the distributions via the parametrization of fitting functions.

3.1 Feature vector

Clustering is done based on features of the given data set. This feature vector can be a vector (1feature) or a n-dimensional matrix (multiple features). The feature vectors serve as descriptive determinants of travel behavior. The most important features for dividing the data set into maximal homogeneous subsets are found during clustering. However, it is beneficial to choose feature vectors in a meaningful way. Within this work, we focused on the discussed socio-demographic and geographic/ spatial variables as both seem to influence travel behavior, from a heuristic perspective. It is possible and encouraged to include other features as well. However, the constraint that the chosen features have to exist in both the HTS and the PHS limits this number of potential candidates drastically. Yet, this constraint is a prerequisite for the sampling of data from the HTS to the whole population of the PHS.

Table 2: Approaches to optimize methodology

Feature vector	• Demographic variables • Geographic/ spatial variables
Classification method	• Unsupervised Learning via PCA + k-means • Supervised Learning via decision tree classification
Target vector	• Daily Travel Time • Daily Travel Distance
Sampling method	• Random Choice • Inverse Cumulative Distribution Function
Parametrization of distance distribution	• Gamma • Lognormal • (Exponentiated) Weibull

3.2 Classification method

As clustering technique, we considered Decision Tree (DT) classification as well as k-means clustering. Both aim at finding clusters within the feature vectors from HTS regarding their travel behavior. Whereas k-means is a unsupervised clustering technique, the DT approach represents a supervised machine learning algorithm. Hence, it takes a target vector additionally to the feature vector as input.

DT classification is widely used in literature. Mohammadian and Zhang (2007) and Pirra and Diana (2016) use DT clustering for the classification of tours. Hagenauer and Helbich (2017) and Sekhar *et al.* (2016) use (Random Forest) DT clustering for mode choice analysis. Hagenauer and Helbich (2017) found this classifier as the best-performing among several machine learning algorithms. However, there is also literature using the k-means algorithm for generating mobility archetypes, e.g. by Charleux (2018).

Both methods are described in the following. Their classification performance, with respect to the validation on the annual mileage of the HTS report, is shown in Table 3.

3.2.1 Decision tree classification

Supervised machine learning algorithms like DT classification rely on training a classifier with a primary data set (train set, I), consisting of a feature vector and a target vector with corresponding labels. With that primary data set, the classifier can be fitted. The performance of the model (goodness of fit) has to be tested on a secondary data set (test set, II). However, when optimizing the hyperparameters of the classifier, there is risk of overfitting on the test set. To avoid this problem, either a third data set (validation set, III) can be set apart in the beginning to validate the performance after optimizing using the test set, or cross-validation can be used. After training, optimizing and validating, the classifier can be used for the prediction of labels of a new unknown data set comprising only the feature vector (no labels).

However, we do not follow the conventional fit-predict-path of machine learning. Instead of predicting labels on new data sets, we are rather interested in the classification structure itself since it allows for intuitive traceability of how data is clustered. The decision tree classification method suits very well to this requirement as it has an intuitive visualization of how clustering is performed in form of the decision tree. For our purposes, we rearranged the conventional procedure as follows: We keep the fitting (I) on a given primary train set (the HTS data). As categorical feature vectors are not suitable for the binary splits of the DT classifier, we converted categorical into binary features. The binary splits of the decision tree are based on the mean squared errors (MSE) cost function as classification measure. The MSE is defined as follows:

$$MSE = \frac{1}{n} SSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

Data is split such that the sub-nodes are most homogeneous, i.e. minimizing the MSE. However, we do not want to predict new labels in the next step. Hence, testing and validating can be simplified. Testing the performance of the model (II) is done by adjusting the hyperparameter settings. Larger DTs, i.e. high number of leafs[3] and low number of samples within one leaf, are prone to overfitting to noise. In contrast, small DTs may not cover all important structural information of the data. As discussed in the general methodology, validation (III) is done after sampling from the decision tree leafs onto the whole population from the PHS data set. Exemplarily, this means the following: When using the DTD as target variable, the resulting aggregated value for the annual mileage, i.e. the total distance traveled by the whole population per year, is validated against the corresponding value from the HTS report Perret (2016).

For the implementation of the illustrated classification methodology, we used the Decision Tree Classifier from the Python package *scikit-learn*.

[3] We refer to *leafs* as the end nodes of a decision *tree*.

3.2.2 k-means clustering

Unsupervised learning is used to cluster data with unknown labels. k-means is a representative of centroid-based clustering techniques where the data is initially split into k clusters and the cluster centers are computed as the mean of all data points within the clusters. Then, each point is assigned to its nearest cluster center and the cluster centers are re-calculated. This iterative procedure is an optimization task, driven by a cost function like the SSE.
Due to that procedure, k-means allows only for continuous variables as feature vectors. Therefore, we applied principal component analysis (PCA) to convert categorical to continuous variables in a preprocessing step. Depending on the initial clustering, the procedure is prone to ending up in local minima and to overfit against outliers.

3.3 Target vector

Supervised learning requires a target vector, also called label, to the feature vector. The data is clustered with respect to that target vector. We chose travel distance and travel time as the two most promising candidates.

3.4 Sampling method

The classification step results in clusters of people from the HTS sharing the same travel behavior, i.e. the same DTDC. However, the sample size is limited with 55’052 people interviewed in HTS. To infer the travel behavior of the groups of people that share similar features (i.e. similar values of the feature vectors), the number of people (p) from the PHS who show the same features for each group is extracted. For each cluster, p samples are drawn from the DTDC vector of the corresponding HTS cluster. This matching allows for extending the findings of the clustering step onto the 6’416’153 people from the PHS which are older than 18 and allowed to drive a car. To even out sampling deviations, this procedure is executed 100 times.
Sampling was done via Random Choice (RC) and via the Inverse Cumulative Distribution Function (ICDF). While the RC method allows only existing data points to be picked (e.g. existing travel distances), the ICDF represents a continuous function from which arbitrary values can be sampled.

3.5 Parametrization of distance distribution

Clustering and sampling results in a DTDC distribution curve, indicating the probability of each DTDC like shown cumulatively in Figure 1(b). Up to now, the histogram of all data points defines this curve. However, the DTDC distribution may be parametrized by fitting functions which would allow for separating the distribution from its underlying data. Hence, the parametrization aims at inferring a generalized function for the DTDC distribution that solely relies on the function parameters. The described model order reduction method enables not only a reduction of complexity in describing the travel behavior, but also faster computation.
As fitting function candidates, we chose the Lognormal, the exponentiated Weibull and the Gamma function. All of them were only defined for $x > 0$ since we only have positive distances as x values. In the following function descriptions $k, \alpha, \lambda, \vartheta$ are scaling and shape parameters, whereas μ is a location parameter:

- Lognormal function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} exp\left(-\frac{(ln(x) - \mu)^2}{2\sigma^2}\right) \tag{4}$$

- Exponentiated Weibull function:

$$f(x) = \alpha\frac{k}{\lambda}\left(\frac{x-\mu}{\lambda}\right)^{k-1}\left[1 - exp\left(-\left(\frac{x-\mu}{\lambda}\right)^k\right)\right]^{\alpha-1} exp\left(-\left(\frac{x-\mu}{\lambda}\right)^k\right) \tag{5}$$

- Gamma function:

$$f(x) = \frac{1}{\Gamma(k)\vartheta^k}\left(\frac{x-\mu}{\vartheta}\right)^{k-1} exp\left(\frac{x-\mu}{\theta}\right) \tag{6}$$

To analyze the goodness of fit of different fitting functions on a data-driven distribution, we used the Akaike Information Criterion (AIC) defined by:

$$AIC = 2k - 2ln(\mathcal{L}(\Theta_{MLE}|x)) \tag{7}$$

with k being the number of estimated parameters and $\mathcal{L}(\Theta_{MLE}|x)$ being the maximized value of the Maximum Likelihood Estimation (MLE) for the fitted function given the estimated parameters Θ_{MLE} and the sampled data x. Plötz *et al.* (2017) found that the AIC is the best measure of goodness of fit when comparing different fitting functions for DTD distributions. It has to be mentioned that the AIC only allows for comparisons within the observed data set.

4 Results & Discussion

4.1 Feature vector

For the implementation of the classification methodology, the feature vector consisted of socio-demographic and geographical/ spatial variables, as discussed in the data section. As target variable we used the daily travel time as well as the daily travel distance. The optimization of hyperparameters was done using the DTD as target vector, resulting in the following optimal values:

- maximum number of leafs: 15
- minimum number of samples per leaf: 5% of HTS data size

There is a trade-off between the maximum number of leafs and minimum number of samples per leaf. The chosen values represent the optimum with respect to the validation of the annual mileage.

Figure 3 shows the resulting decision tree when feeding the DT classifier with all demographic and geographic/ spatial features. The nodes of the DT show the following information:

- 1st line: Decision criteria that splits the data set most homogeneously.
- 2nd line: Resulting MSE for binary split.
- 3rd line: Number of samples within the node.
- 4th line: DTD value within the node.

After sampling, the found results are validated against the HTS reference value for the annual mileage of all car drivers in Switzerland which amounts to 46'225 millions of kilometers (mkm) (Perret, 2016).

Figure 4(a) validates the annual mileage of HTS (black line) against the calculated values from the DT when including only demographic values as feature vectors. The orange bins show the results of 100 iterations of re-sampling. The red lines indicate the 95% confidence interval of the sampled values. Calculation and reference number fit perfectly. The relative difference is only 0.06%. In contrast, the disaggregated cantonal DTDC values are not met, as depicted in Figure 5(a). As we have fed the DT classifier only with demographic feature vectors, this is not surprising.

To overcome this limitation, we included geographic/ spatial variables as input to the DT classifier. The results are shown in Figures 5(b). Now, the cantonal values are met much better. However, the relative difference of the annual mileage increases to 1.71%. This might indicate a

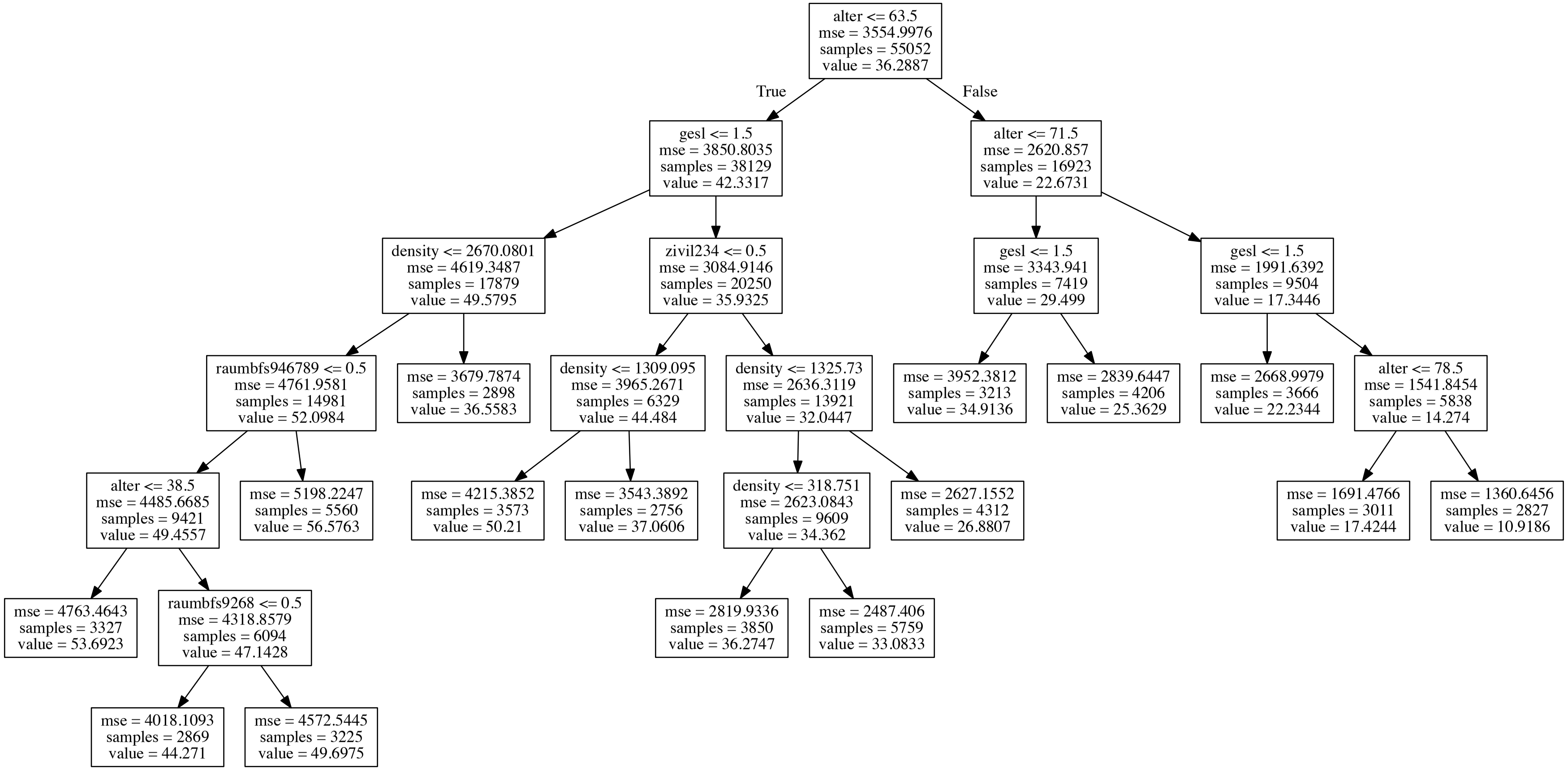


Figure 3: Decision Tree based on all demographic and geographic / spatial features and daily traveled distance as target variable

Table 3: Comparison of different clustering techniques and target variables

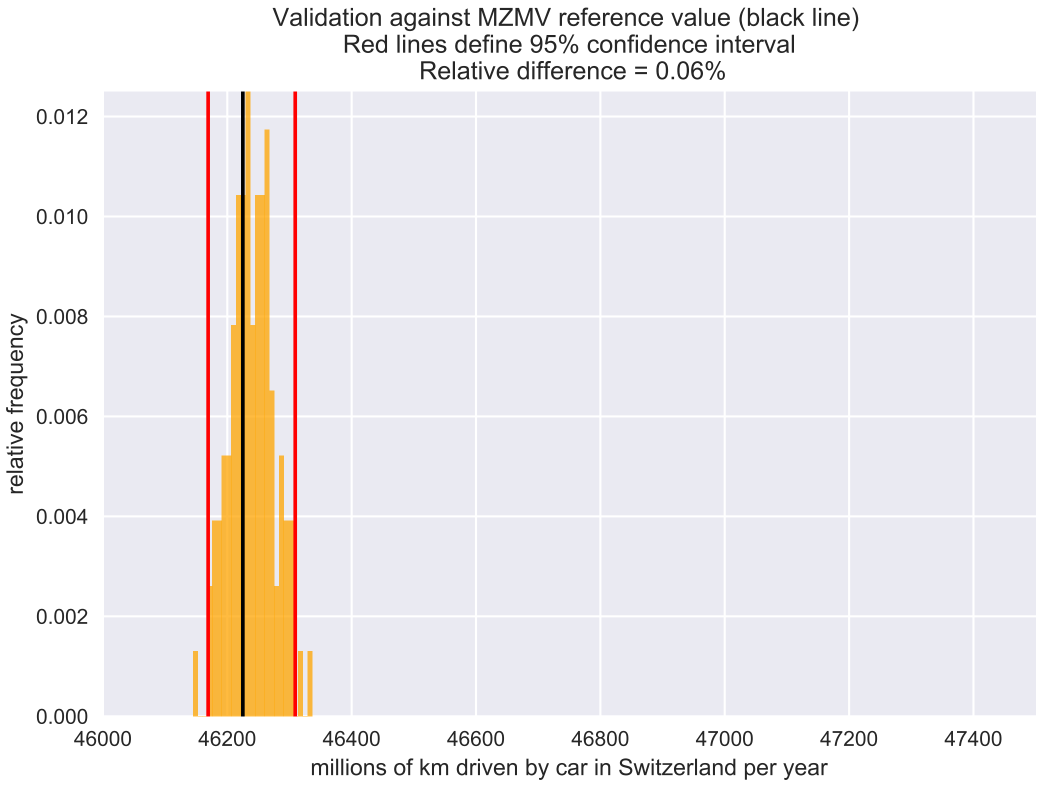
	HTS	Decision Tree (Target Variable: Distance)	Decision Tree (Target Variable: Time)	k-means Clustering
Total distance by car [mil. km]	46’225	46’252	45’410	45’363
Difference to HTS data [%]	0.00	0.06	-1.76	-1.86

trade-off on meeting disaggregated cantonal DTDC values or aggregated annual mileages, even if the found values are still on the border of the HTS confidence interval (CI). Hence, these findings have to be seen in the light of HTS uncertainty, as the HTS data relies on a sample size that represents only 0.86% of the Swiss population aged over 18. The Swiss Federal Statistics Office takes this uncertainty in form of a confidence interval into account (Perret, 2016): Here, the overall annual mileage of passenger car mobility of 46’225 mkm and the 90% CI, that is in the range between 45’508 and 46’942 mkm, is provided.

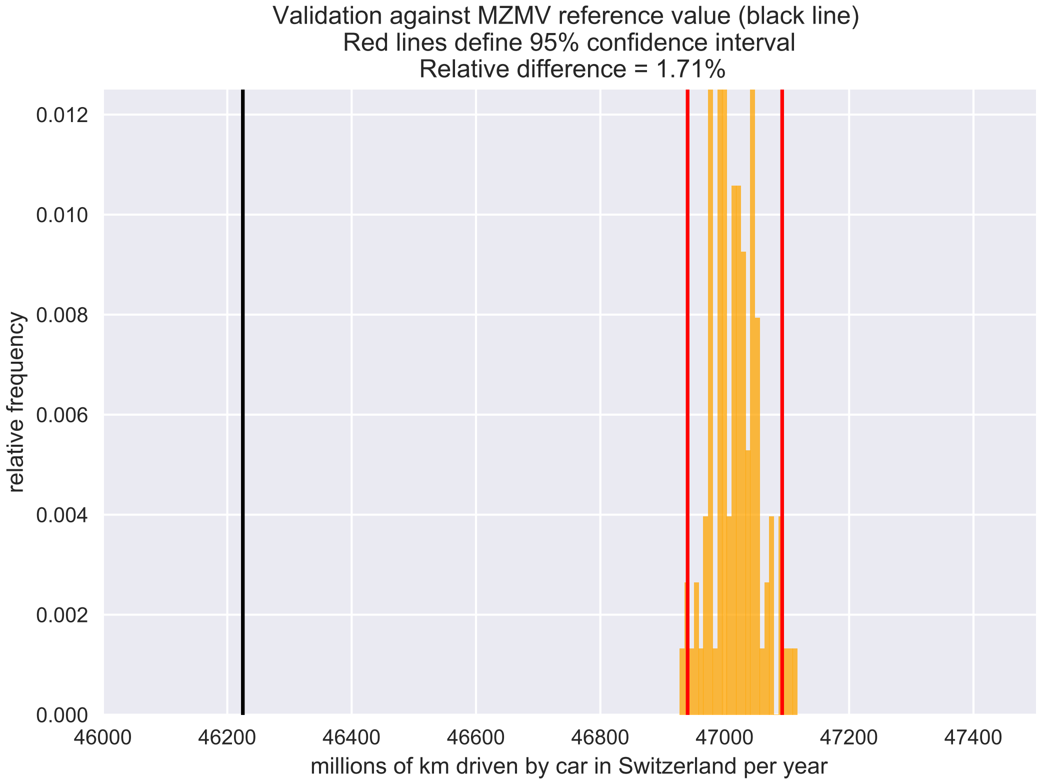
It is worth mentioning that the most descriptive features for clustering are demographic variables, even in the case when including geographical/ spatial feature vectors: The feature importances reported from sklearn are 56.9% for age, 22.2% for gender, 18.4% for marital status and 2.5% for the Extensive Municipality Characteristic.

4.2 Classification method

Table 3 shows the superiority of supervised against unsupervised learning when choosing the right target variable. The DT with distance as a target variable results in a far better value of annual mileage, compared to the DT with time as target variable or to k-means clustering. However, these findings have to be seen in the light of the CI of the HTS reference value, again. The value for the total distance by car found by k-means is not too far away from the 90% CI reported in Perret (2016).

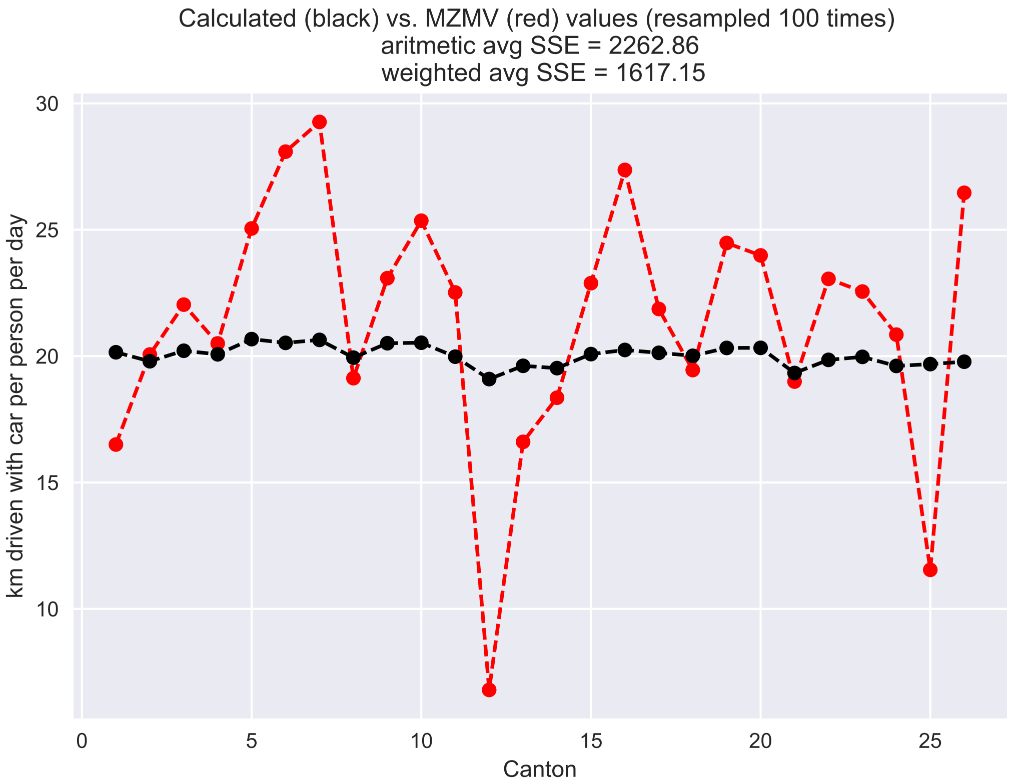


(a) Validation of aggregated annual mileage (Feature vector for DT including demographic variables)

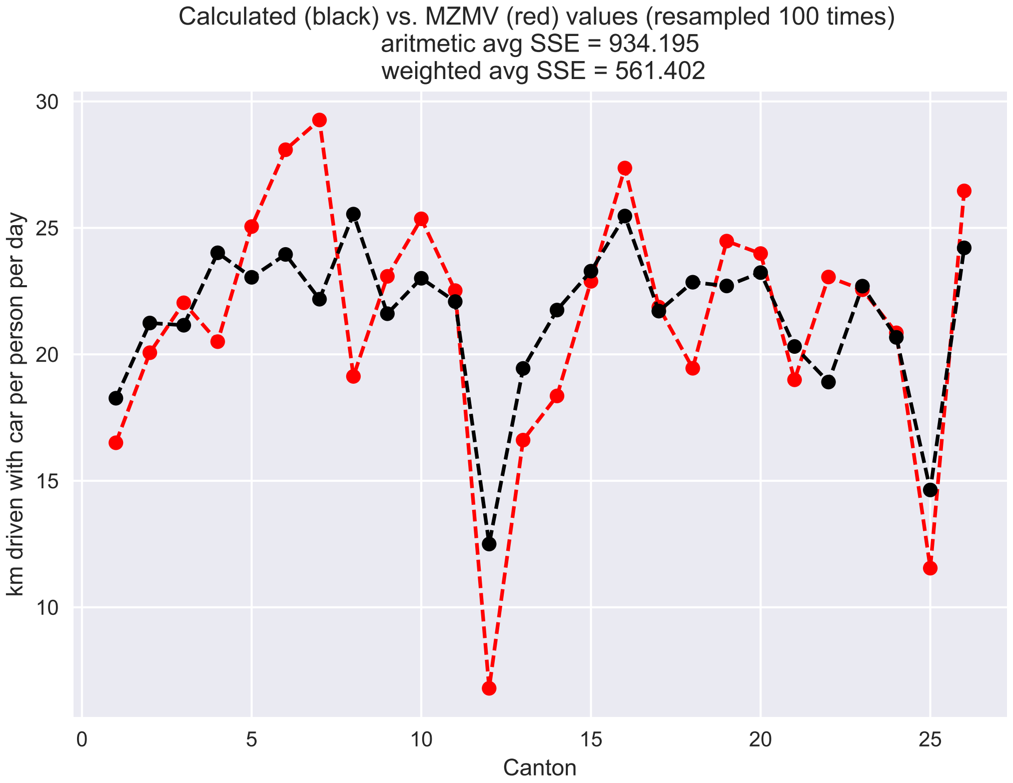


(b) Validation of aggregated annual mileage (Feature vector for DT including demographic and geographic/ spatial variables)

Figure 4: Validation of decision tree findings on aggregated annual mileage of the HTS



(a) Validation of cantonal annual mileage (Feature vector for DT including demographic variables)



(b) Validation of cantonal annual mileage (Feature vector for DT including demographic and geographic/ spatial variables)

Figure 5: Validation of decision tree findings on cantonal annual mileage of the HTS

4.3 Target vector

Table 3 shows the performance of the two different target variables using decision tree classification. It has to be mentioned that this analysis has been done as a preceding step. Hence, it is based only on demographic features. However, as demographics will show the highest feature importance even when including geographical/ spatial variables, the results are still valid.

4.4 Sampling method

Within their CIs, RC and ICDF sampling provide identical results. The deviation from the HTS reference value is in any case larger than the sampling error. Hence, the impact of the sampling method on the accuracy of the results can be neglected. However, RC results in 75 times faster sampling wherefore its usage is recommended.

4.5 Parametrization of distance distribution

We assessed the capability of the Lognormal, exponentiated Weibull and Gamma function to fit the DTDC probability distribution curve using the AIC. The differences in AIC values for the observed fitting functions show high favorization of the exponentiated Weibull distribution.
Figure 6 shows the DTDC probability distribution histogram (yellow) for distances greater than 0 and clipped at 300km. The superimposed fitting curve (red) is defined by the following exponentiated Weibull function:

$$f(x) = 1.54x^{-0.60} \cdot \left[1 - exp(-x^{0.40})\right]^{2.83} \cdot exp(-x^{0.40}) \tag{8}$$

Our findings confirm the findings of Plötz *et al.* (2017) who stated that the Weibull function performs best on several data sets of DTD.

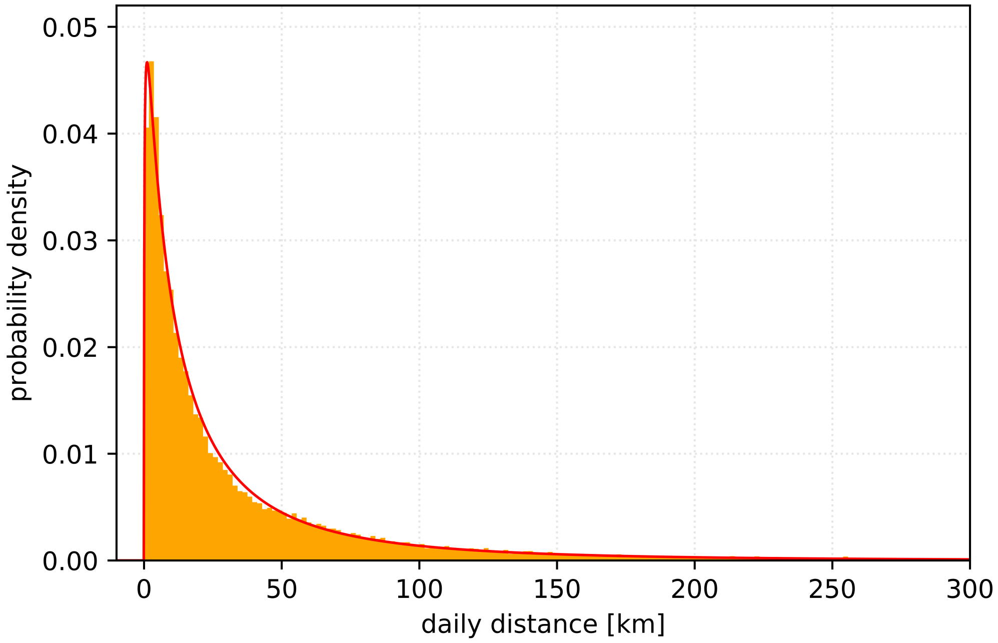


Figure 6: DTDC probability distribution with exponentiated Weibull fitting curve (red) within the range]0;300]km

5 Conclusions & Outlook

Within this work, we present a methodology for describing mobility demand based on socio-demographic and geographical/ spatial data. Following the structure of the methodology and the results section, we draw conclusions for each aspect of the proposed methodic approach:

- **Feature vector**: We used socio-demographic and geographic/ spatial variables displayed in Table 1 as feature vector for clustering. Socio-demographic features turned out to be most descriptive for clustering the population with respect to their daily driven distance. We recommend to include other features as well, considering the constraint that the chosen features have to exist in both the HTS and the PHS. This could even enhance the model results. However, enhancement does only make sense within the range of the CI of the annual mileage reported in the official HTS report (Perret, 2016).
- **Classification method**: We found that the Decision Tree is the best classification method to cluster HTS data for our purposes. Here, we do not see need for improvement.
- **Target vector**: We used travel time and distance as target vectors, with distance performing best. We encourage to enlarge the space of target vector candidates to improve the clustering results. The same constraints as described for the feature vector hold here.
- **Sampling method**: We have shown that the sampling method does not impact the results of our methodic approach significantly. No improvements are necessary.
- **Parametrization of distance distributions**: As a form of model order reduction, we analyzed approaches to parameterize the distribution of DTDC. We found that the exponentiated Weibull function has the highest goodness of fit when using the HTS from 2010. It is worth to further investigate appropriate fitting functions and their ability to generalize on different years of HTS.

In conclusion, we identified explaining features that allow for the classification of HTS samples into clusters sharing the same mobility behavior. We have proven that reconstructing aggregated mobility activity through classification and sampling is feasible for the HTS of 2010. Geographical discretization improves cantonal means, while marginally worsening the national aggregate. This can serve as a basis for future work. The future mobility demand can be extrapolated when using predictions of socio-demographics for the future. Validation of this approach will be done in the future via back-casting over all existing HTS data sets since 2000. Based on the extrapolation of mobility demand, decarbonization pathways can be modeled. This extrapolation is not a prediction including effects of new forms of mobility like autonomous driving. However, extrapolating future mobility demand based on status-quo data keeps uncertainties of assumptions low. This, in turn, improves time-resolved decarbonization

scenarios relying on the mobility demand estimations.
The beauty of our statistical approach lies in its simplicity contrary to agent-based modeling, while providing sufficient insight into mobility demand as basis for decarbonization potentials of passenger cars. As we are not interested in origin-destination- or traffic flow-based geographical information, our methodology can solely rely on data that is often open source and at least accessible for industry or policy makers.

Acknowledgement

We would like to acknowledge the work of all students that contributed to the findings of this paper. In particular we like to thank Emir Çabukoglu (2016), Iason Chontzopoulos (2017), Charles Metz (2017), Lukas Sieber (2017), Karim Ziadé (2017) and Nicola Zoppini (2017). We gratefully acknowledge the financial support of the Swiss Federal Office of Energy [SI/501311-01] and the Swiss Competence Center for Energy Research in Efficient Technologies and Systems for Mobility (SCCER Mobility), funded by Innosuisse. Furthermore, we thank the Swiss Federal Statistics Office for providing us the HTS data and the PHS data.

References

Bauer, C., B. Cox, T. Heck, S. Hirschberg, J. Hofer, W. Schenler, A. Simons, A. D. Duce, H.-J. Althaus, G. Georges, T. Krause, M. G. Vayá, F. Ciari, R. Waraich, B. Jäggi and A. Stahel (2016) Opportunities and challenges for electric mobility: an interdisciplinary assessment of passenger vehicles, *Technical Report*, ETH Zürich, Empa, PSI, Zürich.

Çabukoglu, E. (2016) Analysis of "Microcensus Mobility and Transport" and Calculation of Disaggregated Swiss Mobility Demand Using Clustering and Sampling Techniques, Master thesis, ETH Zürich.

Charleux, L. (2018) Deriving Mobility Archetypes from Household Travel Survey Data, *The Professional Geographer*, **70** (2) 186–197, apr 2018, ISSN 0033-0124.

Chontzopoulos, I. (2017) Prediction of hourly marginal CO2 emission factor for national electricity systems, Semester thesis, ETH Zürich.

Diana, M. (2012) Studying Patterns of Use of Transport Modes Through Data Mining, *Trans-*

portation Research Record: Journal of the Transportation Research Board, **2308**, 1–9, dec 2012, ISSN 0361-1981.

Federal Customs Administration (2017) Performance-related heavy vehicle charge, `https://www.ezv.admin.ch/ezv/de/home/information-firmen/transport--reisedokument--strassenabgaben/schwerverkehrsabgaben--lsva-und-psva-/lsva---allgemeines---tarife.html`.

Federal Office for the Environment (2017) Greenhouse gas inventory, `www.bafu.admin.ch/bafu/de/home/themen/klima/daten-indikatoren-karten/daten/treibhausgasinventar.html`.

Federal Statistics Office (2010a) Swiss National Household Travel Survey, `https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/personenverkehr/verkehrsverhalten.html`.

Federal Statistics Office (2010b) Swiss Population and Households Statistics.

Hagenauer, J. and M. Helbich (2017) A comparative study of machine learning classifiers for modeling travel mode choice, *Expert Systems with Applications*, **78**, 273–282, jul 2017, ISSN 0957-4174.

International Energy Agency (2014) CO2 emissions from transport, `https://data.worldbank.org/indicator/EN.CO2.TRAN.ZS`.

Loder, A., R. Tanner and K. W. Axhausen (2017) The impact of local work and residential balance on vehicle miles traveled: A new direct approach, *Journal of Transport Geography*, **64**, 139–149, oct 2017, ISSN 0966-6923.

Mathys, N., A. Justen, R. Frick, L. Ickert, M. Sieber, F. Bruns, N. Rieser, J. Uhlig, B. Dugge and J. Landmann (2016) Perspektiven des Schweizerischen Personen- und Güterverkehrs bis 2040, *Technical Report*, Swiss Federal Office for Spatial Development.

Metz, C. (2017) Data fitting with analytical functions to expand insight into mobility behaviour Bachelor Thesis, Bachelor thesis, ETH Zürich.

Mohammadian, A. and Y. Zhang (2007) Investigating Transferability of National Household Travel Survey Data, *Transportation Research Record: Journal of the Transportation Research Board*, **1993**, 67–79, jan 2007, ISSN 0361-1981.

Perret, C. (2016) Leistungen des privaten Personenverkehrs auf der Strasse, *Technical Report*, Federal Statistics Office, Neuchatel.

Pirra, M. and M. Diana (2016) Classification of Tours in the U.S. National Household Travel Survey through Clustering Techniques, *Journal of Transportation Engineering*, **142** (6) 04016021, jun 2016, ISSN 0733-947X.

Plötz, P., N. Jakobsson and F. Sprei (2017) On the distribution of individual daily driving distances, *Transportation Research Part B: Methodological*, **101**, 213–227, jul 2017, ISSN 0191-2615.

Saadi, I., A. Mustafa, J. Teller and M. Cools (2017) A bi-level Random Forest based approach for estimating O-D matrices: Preliminary results from the Belgium National Household Travel Survey, *Transportation Research Procedia*, **25**, 2566–2573, jan 2017, ISSN 2352-1465.

Schafer, A. (2000) Regularities in travel demand: An international perspective, *Journal of Transportation and Statistics*, **3**, 1–31, jan 2000.

Sekhar, C. R., Minal and E. Madhu (2016) Mode Choice Analysis Using Random Forrest Decision Trees, *Transportation Research Procedia*, **17**, 644–652, jan 2016, ISSN 2352-1465.

Sieber, L. (2017) Demand Model for multi-modal and spatially disaggregated daily Mobility using Classification and Sampling, Semester thesis, ETH Zürich.

Tamor, M. A., P. E. Moraal, B. Reprogle and M. Milači (2015) Rapid estimation of electric vehicle acceptance using a general description of driving patterns, *TRANSPORTATION RESEARCH PART C*, **51**, 136–148.

Ziadé, K. (2017) Co-evolution of Switzerland' s demography and the demand for individual mobility : trends and extrapolability Semester project, Semester thesis, ETH Zürich.

Zoppini, N. (2017) Spatially Disaggregated Analysis of Car Mobility Patterns in Switzerland Semester Project, Semester thesis, ETH Zürich.

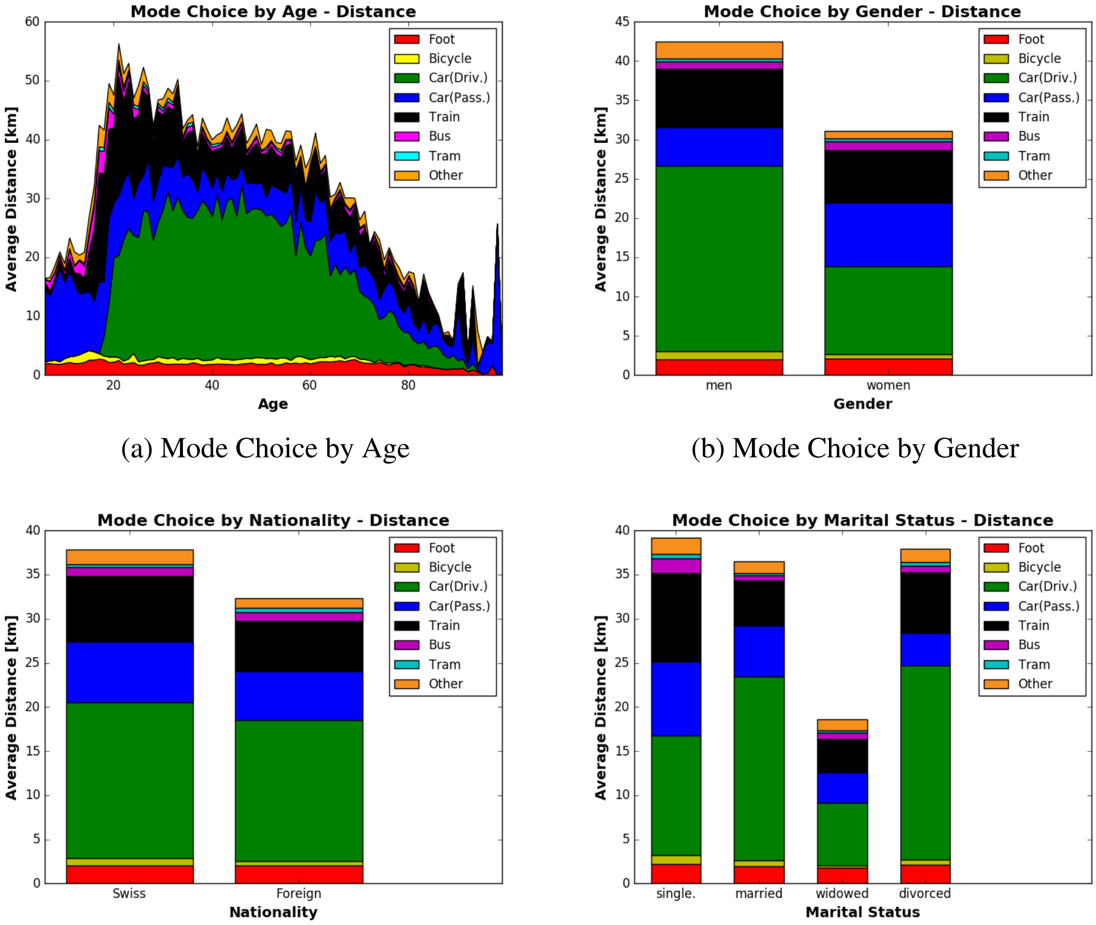
A Data Analysis

Figures 7 (a)-(d) show the average distance traveled by different modes, analyzed by a set of demographic indicators. The main findings and conclusions are the following:

- The average daily distance jumps to higher values at the age of 18 as the car is added as alternative means of transportation.
- The high share of car mobility shows the high importance of the automotive industry for decarbonization of mobility, even for a country like Switzerland where the rail and bus network is highly developed.
- The average distance by men is higher than that of women.
- The average distance of Swiss citizens is higher than that of foreigners.
- The share of car mobility increases with marriage. The increase goes hand in hand with a decrease of train mobility. Over all modes, the average distance remains approximately constant.
- Widowed persons show significantly lower average distances. This goes along with declining mobility for higher ages.

Figures 8 (a)-(d) show the average distance traveled by different modes, analyzed by a set of socio-economic indicators. The main findings and conclusions are the following:

- The average distance traveled by car increases with the number of cars of a household. If a household has no car, the share of train distance is higher. It is worth to be mentioned, that for increasing number of cars the overall distance increases as well.
- Car mobility highly depends on the availability of a car. The lack of access to a car decreases the average distance significantly.
- Having a driver license increases the average distance significantly by the distance driven by car, while the distances traveled by other modes do not decline equally.
- Higher income leads to higher mobility. This has already been proven by Schafer (2000) who found that travel distance increases with higher income as rich people are able to rely on faster means of transportation like car or train.



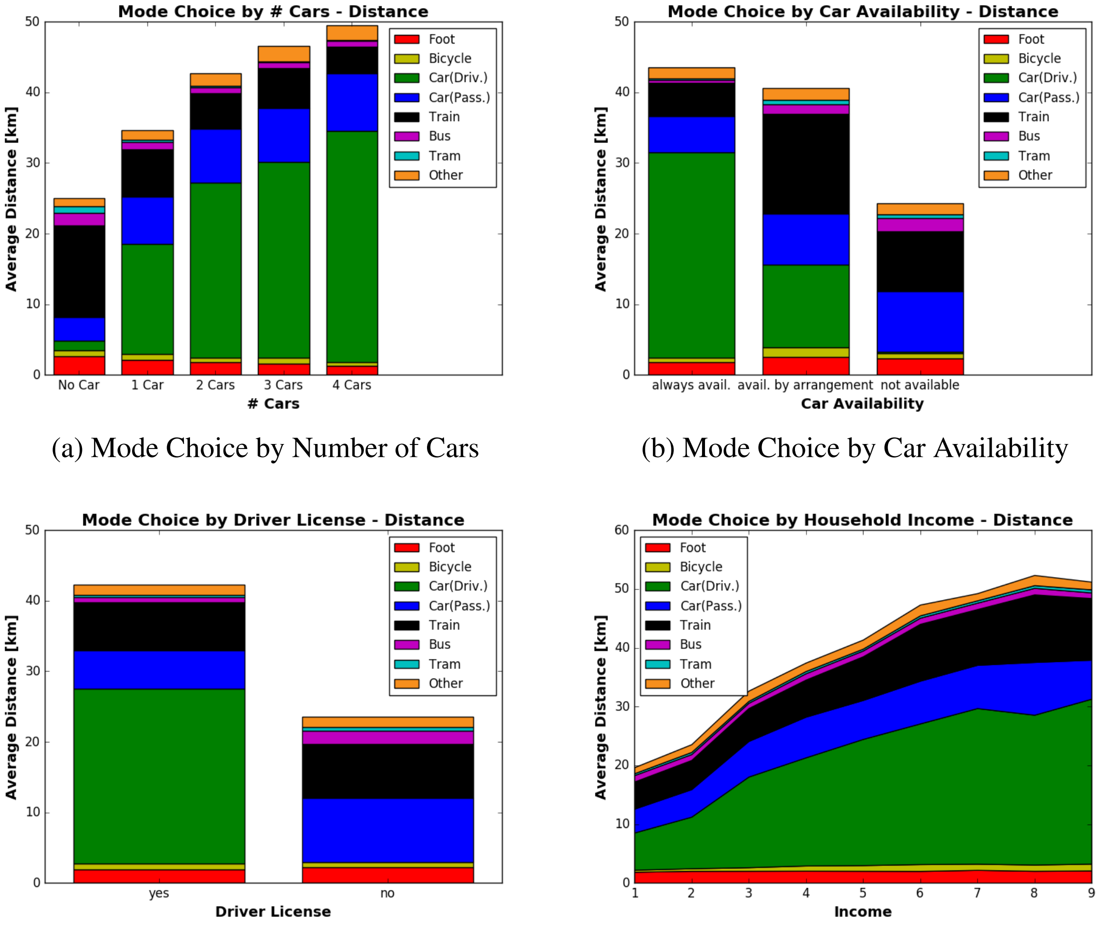
(a) Mode Choice by Age

(b) Mode Choice by Gender

(c) Mode Choice by Nationality

(d) Mode Choice by Marital Status

Figure 7: Average distance traveled by different modes, analyzed by a set of demographic indicators (Çabukoglu, 2016)



(a) Mode Choice by Number of Cars

(b) Mode Choice by Car Availability

(c) Mode Choice by Driver License

(d) Mode Choice by Income

Figure 8: Average distance traveled by different modes, analyzed by a set of socio-economic indicators (Çabukoglu, 2016)