Master Thesis

Predicting energy use of individual mobility on survey and tracking data

---

University of Basel
Faculty of Business and Economics

Roman Stutzer

Roman Stutzer
Lettenrain 2a
4104 Oberwil, Switzerland
roman.stutzer@unibas.ch

Matriculation number: 2015-411-820

Major: Quantitative Methods
Supervised by: Professor Dr. Beat Hintermann

Submission Date: December 29, 2020

UNIVERSITÄT BASEL

STUDIENDEKANAT DER WIRTSCHAFTSWISSENSCHAFTLICHEN FAKULTÄT

Plagiatserklärung

„Ich bezeuge mit meiner Unterschrift, dass meine Angaben über die bei der Abfassung meiner Arbeit benützten Hilfsmittel sowie über die mir zuteil gewordene Hilfe in jeder Hinsicht der Wahrheit entsprechen und vollständig sind. Ich habe das Merkblatt zu Plagiat und Betrug vom 22.02.11 gelesen und bin mir der Konsequenzen eines solchen Handelns bewusst."

Name, Vorname: Stutzer, Roman

Ort und Datum: Oberwil, 29. Dezember 2020

Unterschrift: 

Contents

List of Figures

List of Tables

1 Introduction

In Switzerland, mobility accounts for around 40% of domestic energy consumption and CO_2 emissions (Bundesamt für Statistik 2016). Thereof, individual mobility accounts for three quarters of the consumed energy and the transport of goods for one quarter. Transport policy hence has potentially great leverage on energy consumption and the internalization of related externalities. Being able to make a prediction of the change in energy use as a consequence to such a policy intervention is therefore of great relevance if the costs and benefits of an intervention are to be assessed. The cost benefit analysis is the most basic economic concept brought forward when policy decisions are prepared. They are of particular relevance in the context of the preservation of public goods like a clean environment. With regard to the current climate crisis, many policy measures which are discussed today target individual behavior and thereby are expected to affect energy consumption. If the consequences of these prospective policies are to be assessed *quantitatively* (for possible inclusion in a cost-benefit analysis), this requires the prediction of this specific outcome measure.

In this thesis, I derive a prediction model that serves the end to improve the evaluation of the effects of possible prospective policy measures on individual energy use. In the empirical application, I draw on data from a mobility study in Switzerland, i.e., the MOBIS field experiment (Axhausen et al. 2021). This prediction data allows me to connect accurately measured energy use of individuals' mobility behavior with numerous personal characteristics. The observed personal characteristics include values and beliefs, which are potentially important predictors of energy use next to people's endowment (Enzler and Diekmann 2019). On this basis, I train different prediction models making use of regression forests and lasso, two popular machine learning algorithms (see Athey and Imbens 2019 for an introduction to machine learning in economics). The model showing the best cross validation performance is then applied to impute energy use into a (prospective) stated choice experiment, which is not part of this thesis.

The imputation application relies on a restricted number of predictor variables. This is the case because the predictor variables for the imputation task in the target data set have to be surveyed. The number of questions a survey sustains is, however, limited and thus they have to be wisely chosen. To minimize the adverse effect of this restriction on prediction accuracy, I develop a variable selection procedure. The selection procedure is based on the two prediction algorithms mentioned above

and identifies the most promising predictors which the field experiment makes available to predict individual energy use. This variable selection procedure could be extended to other methods and also applied to other data sets.

In addition to the imputation of the level of energy use, the model can also be applied for the prediction of differentials in energy use. This refers to the second application, when the prediction model is combined with recorded intentions of behavioral change in a stated choice experiment. The prospective stated choice experiment is designed such that it records how participants rearrange their mobility toolbox in hypothetical scenarios of the future. The mobility toolbox summarizes all available mobility tools people own such as bicycles, cars, and public transport passes (for a description of recent developments in Switzerland see Kowald et al. 2017). From a change at the extensive margin like, for example, no longer owning a car, the prediction model allows for an estimation of the impact on the intensive margin in terms of energy use.

The remainder of the present thesis is structured as follows. Section 2 provides the theoretical economic framework for the prediction analysis. In Section 3, the empirical application is embedded in its broader context by a description of the field experiment that provides the prediction data and some explanations regarding the prospective stated choice experiment. The prediction data is described in Section 4. Section 5 introduces the prediction algorithms. This section also comments on the methodological reasoning behind the variable and model selection procedure. Section 6 then presents the application of these procedures to the prediction of individual energy use from mobility. Section 7 concludes.

2 Theory on mobility behavior

2.1 Microeconomic foundation

For approaching any prediction problem one should have some theory about the emergence of the response, the dependent variable. A good theory allows to isolate promising predictors in the abundance of available data. It also allows checking for the plausibility of the results and helps interpreting them. I inform the prediction of energy use by a microeconomic theory on an individual's mobility behavior (see, e.g., Button 2010).

From a microeconomic perspective, an individual's behavior results from the interplay of preferences and restrictions. A person, for example, prefers not to get wet on the way to work and therefore looks for an alternative mode of transport to the bicycle on a rainy day. Another person likes fast sports cars, but does not possess one because her own wealth and income does not allow it.

Besides the weather, income and wealth many other factors restrict a person's behavior by affecting the relative attractiveness or prices of the options in the choice set. One group of restrictions that is often distinguished from the others are institutions like state laws. In the focus of this thesis are transport policies that co-determine energy use. Given today's political goals towards sustainability transport policies which incentivize energy efficiency and transition to renewable energy sources are of high interest. Against this background, there is a focus on policy interventions that increase the relative price of energy intensive transport modes which rely on fossil fuels. It is an attempt to price in accruing externalities. The behavioral reaction to such a restriction depends on the expected permanence of the intervention (see, e.g., Button 2010, p. 88). For a restriction which is perceived to persist the adaptation is more far reaching. This adaptation includes all substitution effects and unfolds over time.

When the transport mode car becomes relatively more expensive due to an intervention the immediate, adaptation of a car owner which commutes into town is limited. In the short-run many choices are fix and little margin is left for adjustment. In the medium-run this same person has more possibilities to adapt and may acquire a public transport pass and sell his or her car. In the long-run even more choices may be reconsidered taking into account the change in relative prices. Perhaps the person in our example no longer commutes and lives in town now, next to his or her workplace. Thereby the behavioral reactions may well depend on a person's (environmental) attitudes (see, e.g., Enzler and Diekmann 2019).

The preceding example not only underlines that effects take time to realize, but also serves as an illustration for how different parameters become choice variables over time. In the short-run, it is possible to make small changes as, for example, using the car less often for leisure activities and to walk shorter distances, when the mode car becomes relatively more expensive. In the medium-run, greater rearrangements of the mobility toolbox, which summarizes the different modes of transport that are available to a person, are possible. The rearrangement of the mobility

toolbox can consist of buying a public transport pass, change the own car for a more eco-friendly one or abandoning the car completely. Another adaptation that might arise in the medium-run is the formation of a new habit such as to buy local or to organize home-office. In the long-run, many more parameters become choice variables. New employment contracts can be negotiated, the place of work and the place where people live are no longer fix.

2.2 Necessary assumptions for the prediction of differentials in energy use

An observer interested to learn about which choices people make to adapt to a policy intervention can find him- or herself in two possible settings. In the first setting both, the state of the world with and without the policy regime of interest, are observed. In this case, a skillful comparison of the two states can attribute behavioral differences to the two policy regimes under relatively weak assumptions. In the second setting, the regime change relates to a prospective policy. In this case, many assumptions are needed to make a prediction about the effect of the regime change. The econometirc prediction study as pursued in the application of this thesis fits the second setting with a focus on the prediction of medium-run adaptation to a policy change. A more extensive motivation for this focus is derived in the next section devoted to the empirical setting.

In the medium-run, the optimization of the mobility behavior within new boundaries is mainly reflected by the rearrangement of the mobility toolbox. To evaluate the effect of a policy within this time horizon, it is therefore crucial to form a well informed expectation on how affected individuals adapt their mobility toolbox. The strategy followed here is to ask people how they intend to adapt their mobility toolbox in case of the realization of the regime change of interest. The concrete survey type used to collect this information is a stated choice experiment. In the survey of this type it is hypothetically asked: would you sell your car or extend your public transport pass? Based on the answers to questions like this and two assumptions I can make a prediction to approximate the resulting change in energy use. The first assumption is that the individual maintains his or her endowment, attitudes and behavioral patterns that are part of the prediction data apart from the stated change in the mobility toolbox. The second assumption is that the individual adapts in all the unobserved patterns to match the people with the same mobility toolbox after the stated change which are part of the (training) data set

used to fit the prediction model.

2.3 Exemplary application: The motor-vehicle registration tax

The annual motor-vehicle registration tax is a policy tool available to the Cantons in Switzerland which is increasingly used to incentivize the purchase of cars generating low emissions (Alberini and Bareit 2019). Through out the thesis, I refer to this tax as an illustration to make some of the abstract considerations more concrete.

A first application can, for example, illustrate how a policy change can lead to predictable changes in energy use. Imagine a rise in the registration tax for cars with powerful engines. This regime change brings some people to a point where heavy cars are no longer their optimal choice because of the incentive to substitute and the income effect they experience. It can therefore be expected that the tax change has an effect on the mobility tool box that manifests in the medium-run. In their respective studies Alberini and Bareit (2019) confirm this expectation for Switzerland and Yan and Eskeland (2018) for Norway. In Section 6, we will be able to go one step further and predict what a corresponding change could mean in terms of energy consumption.

3 Empirical setting

3.1 The MOBIS experiment

The data I use in my thesis are from the MOBIS experiment.[1] This is a field experiment, which was conducted between August 2019 and January 2020 in Switzerland as the timeline in Figure 1 indicates. The experiment is designed to elicit the effect mobility pricing on people's short-run mobility behavior. For this end, the sample population is randomly split into three groups: the control group and two treatment groups named information and pricing. After four weeks of the 8 week experiment the

[1] The MOBIS experiment is funded by Innosuisse and the Federal Department of the Environment, Transport, Energy and Communications (DETEC). The experiment is executed by the Swiss Federal Institute of Technology Zurich (ETH), the University of Basel and the University of Applied Sciences Zurich (zhaw). The final report of the experiment provides a detailed overview of the entire study (Axhausen et al. 2021).

treatment sets in. In the information treatment, subjects are informed about the externalities created by their mobility behavior. In the pricing treatment, subjects are additionally charged for the netto monetary value of the generated externalities. Having the two treatment groups allows the researchers to assess the relative importance of the change in price and the information provision for the observed average treatment effect of mobility pricing. A smartphone application tracks the subjects movements during the experiment and delivers the treatment.

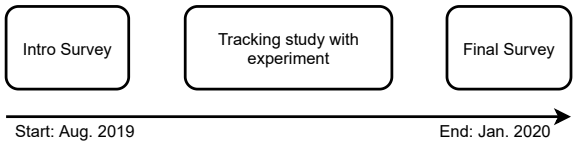

Figure 1: Elements of the MOBIS study

The observed average treatment effect of the pricing treatment is a 5% reduction in the external costs of travel (Axhausen et al. 2021). The effect is statistically significantly different from zero at a level of 0.1%. They decompose the observed treatment effect in their analysis and find, among other things that mobility pricing reduces specifically CO_2 emissions, suggesting that people's mobility becomes less energy consuming. Energy consumption creates externalities as the transport sector still relies heavily on fossil fuels.

The results of the field experiment are informative for assessing mobility pricing as a policy tool to reduce negative externalities from transportation. An important limitation to the insights that can be derived from the experiment is that from observing subjects during a one month intervention, one can make little assertions about the persistence of the observed effects in the medium to long-run. As discussed in the theory section, this is mainly due to the fact that some mobility-related investments and habits are fix in the short-run and not part of the choice set. A conceivable example to illustrate this argument is a person who keeps hold of her car during the experiment while she would sell it if the policy were actually put in place. In this specific case, it is likely that the effect in the short-run, measured in an experiment, stays behind the fully evolved effect of mobility pricing.

3.2 Stated choice experiment

The research group involved with the MOBIS experiment plans to complement their investigation of mobility behavior with a stated choice ex-

periment (SCE) to also shed some light on the medium-run effect of policy changes. For this purpose, a sample population different from the MOBIS experiment is recruited. The SCE is a one-time survey interrogation with no tracking component. In the survey, the subjects will be asked to imagine a scenario and report how they would in response rearrange their mobility toolbox. As the decisions are hypothetical, stated preferences or stated choices are observed. The hypothetical scenarios the participants are confronted with can describe conditions such as the introduction of mobility pricing, the improvement of transport infrastructure, the outbreak of a pandemic, or the increase of a motor-vehicle registration tax.

The latter condition is used for an illustrative SCE. The participant of such a SCE is, for example, asked to choose from two Volkswagen car models, a high and a low emitter. The described attribute for both cars is the associated registration tax. For the high emitter the attribute can take two levels. Either CHF 570 in the status quo or CHF 970 in the case of a tax increase for high emitters. Based on this setup, two vignettes can be constructed.

> Vignette 1:
> Imagine you have the choice between a VW Golf for which you pay a yearly registration tax of CHF 450 and a VW Tiguan for which you pay CHF 570 in taxes annually.
> Which car do you prefer?

> Vignette 2:
> Imagine you have the choice between a VW Golf for which you pay a yearly registration tax of CHF 450 and a VW Tiguan for which you pay CHF 970 in taxes annually.
> Which car do you prefer?

The random allocation of one of the two vignettes to the participants of the survey makes it an experiment, a SCE. The participants who answer the first vignette can be interpreted as the control group, while the ones answering the second vignette receive the treatment of a tax increase. If the stated choices between the two groups differ, we can tell that there is a causal effect of the registration tax on hypothetical car purchases.

Of course, a more sophisticated design of the SCE is required to elicit valuable information about the rearrangement of the mobility toolbox as a reaction to a specific condition.[2]

[2]For more conceptual information on stated choice experiments please refer to

3.3 Imputation

The SCE for the mobility toolbox provides information on the extensive margin of a person's mobility behavior. We can find out whether the person chooses to own a car and/or a public transport pass. The intensive margin, however, remains hidden. In order to nevertheless get an approximation of how much the available mobility tools are used, the imputation of such information offers a solution.

Many aspects of an individual's mobility behavior at the intensive margin can be subject to imputation. Along the lines of the motivation for my thesis, which lies in transport policies that relate to energy efficiency considerations, I envisage the imputation of energy consumption. For this purpose I propose a concrete imputation process to enrich the observations from the SCE with a prediction of the subjects' energy use for mobility. The prediction model for this imputation application is selected in Section 6.

The input variables that are to be considered in the prediction model for a subject's energy use have to satisfy two conditions. First, the variables have to be part of the data set from which we draw the information for the imputation. I call this data set the source data set. Second, the variables have to be part of the data set for which we want to increase the information value by the imputation (for example, of a measure for energy use). I call this the target data set.

In the usual imputation setting, the source and the target data set exist from the beginning. The intersecting set of variables then represents the candidate input variables for the prediction algorithm. The present thesis covers a different setting. It is peculiar in the sense that ex ante only the source data set exists.

With regard to the planed extension of the MOBIS study, the source data refers to the data gathered within the MOBIS experiment. The target data refers to the data that are yet to be collected in a SCE. Importantly, so far in the MOBIS study, each participant has filled out two surveys. This leaves a wealth of over 100 surveyed variables that can serve as potential candidate predictors for the imputation. Incorporating questions for all these variables into the SCE would not be feasible. It would make the related survey to exceed a reasonable length. To nevertheless allow for the best possible imputation within the SCE those variables from the source data set should be surveyed which are most predictive for energy

chapter 6 of Hensher et al. (2005, pp. 189). Two recent applications of the method to mobility undertake Ho et al. (2020) and Stoiber et al. (2019).

use. In the present thesis we analyze a procedure which offers a possible solution to identify the promising predictors.

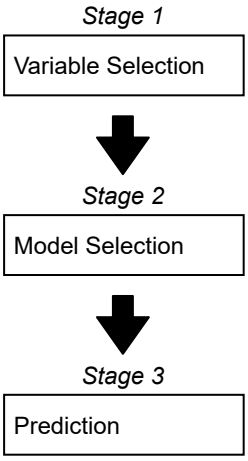

Figure 2: Imputation process

I refer to the procedure to determine the relevant predictor variables as variable selection. The variable selection is considered the first stage in the imputation process depicted in Figure 2. The second stage is then the conventional model selection to find an effective algorithm for predicting energy use based on the pre-selected variables. The third stage completes the imputation task with the application of the trained prediction model to the target data set. The third stage cannot be performed in this thesis as no SCE is undertaken. However, I present the prediction for a hypothetical behavioral reaction as an illustration.

4 Data

4.1 Sample

The data for the current analysis were compiled within the MOBIS experiment. I refer to them as the source data. Stage one and two of the imputation procedure are based on variables from this source data set. Throughout my thesis I use the original names of the variables from this data set if possible. The dependent variable, or response of interest, is the energy use of a person. It is computed from the participants' tracked movements via a smartphone application. The potential predictors offered in the source data set were collected in two surveys that

accompanied the tracking study. After the initial data cleaning[3], the total data set consists of observations from 3,373 individuals and includes 118 candidate predictors. These individuals are from different groups of the experiment as shown in Table 1. As all of them filled in the survey and were tracked, they can all be considered in this study.

Table 1— Experimental sample structure

	Number of participants
Total treatmenet group	2,232
Information treatment	1,125
Pricing treatment	1,107
Control group	1,141
Overall total	3,373

An individual eligible to participate in the MOBIS experiment lives in a metropolitan area in the German- or French-speaking part of Switzerland and travels by car at least on two days per week. Another criterion to qualify for the study is to be between 18 and 65 years old in 2018.[4]

4.2 Response: Energy use

The energy use of a person's mobility behavior is approximated by the CO_2 emissions it generates. It is thus an indicator of energy use that captures the relevant aspect for climate policy. Moreover, it is a reasonable approximation because emissions develop in direct proportion to a persons energy consumption. The CO_2 emissions were imputed based on the tracking data of the MOBIS experiment (and assess the external costs of a person's mobility behavior). The emission factors applied to the imputation are taken from the HBEFA database and are available for different vehicle types and traffic situations (Axhausen et al. 2021). The mobility tracking of the MOBIS experiment differentiates between 16 different modes of transport which are listed in Table 2. Each mode is associated with a different emission factor given the observed travel speed (Axhausen et al. 2021). The modes *walk* and *bicycle* are, for example, associated with zero emissions. For the mode *car* additional information is available regarding the participant's car which allows for a more exact determination of emissions. The surveyed characteristics of

[3] Appendix A.4 provides more information on the data cleaning procedure.

[4] A list of all seven inclusion criteria is provided in Appendix A.1.

a participant's car are its size, fuel type and year of manufacture.[5] The more precise measurement of car emissions should also be reflected in a greater standard deviation of the imputed emissions compared to other mode choices.

Table 2— Average CO_2 emissions by km and mode

Mode	Mean	Max	Std. dev.	n
Taxi, Uber	0.026969	0.1427	0.0142	842
Car	0.026092	4.9025	0.0264	380,983
Carsharing, mobility	0.023533	0.1501	0.0104	736
Motorbike, scooter	0.022471	0.4616	0.0197	4,957
Bus	0.014407	0.2307	0.0013	48,030
Boat	0.002042	0.2017	0.0156	649
Airplane	0.001262	0.0265	0.0058	986
Subway	0.000141	0.0941	0.0025	4,757
Regional train	0.000094	0.0922	0.0012	12,612
Train	0.000072	0.0356	0.0004	18,699
Tram	0.000036	0.0746	0.0011	21,592
Light rail	0.000024	0.0647	0.0009	25,164
Bicycle	0			28,541
Walk	0			50,0016
Aerialway	0			6
Ferry	0			5
Total				1,048,575

Note: n is the number of times a tracked distance is attributed to the respective mode. Aerialway and ferry are neglected because of few observations.

The mobility of a person is understood as the trips a person undertakes between activities. Within one trip to get from one activity to another, the person can rely on one mode or a series of different transport modes. For a person who tracks his or her mobility on a given day 4.16 trips are registered, on average. Each participant in the tracking study is observed on 44.79 days on average.

For a member of the control group, the tracking information gathered over the entire experiment is considered in the prediction task. For participants who are allocated to one of the treatment groups only the first four weeks are considered. This is because after the fourth week in the

[5]Five car sizes are distinguished: luxury car or sports coupé, medium to large car, minivan or van, off-road vehicle, small car. Five fuel types are distinguished: diesel, electric, gasoline, hybrid (gasoline/diesel + electric), other. Seven time periods are distinguished for the year of manufacture.

experiment the treatment sets in which systematically influences their mobility behavior. For example, in the pricing treatment among other things, energy use becomes more expensive. Consequently the tracking information which is potentially affected by the treatment is removed from the data. This leaves us with an average of 45.23 observed days for a member of the control group and with 22.36 days for the members of the two treatment groups.

The CO_2 emitted by an individual's mobility on a given day is approximated by the sum of emissions accumulated over all trips on that day. I refrain from using the panel structure of the data because we cannot learn from the variation in individual's CO_2 emissions over time in this study. This is due to the fact that the available predictors are time invariant as they are collected in a one-time survey, the SCE.

The information which is available on an individual's emissions can be condensed in several ways. I take the average over all the days which are on record for the individual.[6] The histogram on the left in Figure 3 describes the distribution of the average CO_2 emissions per day for the 3,373 individuals in the sample. Reported are monetized emissions. For each emitted ton of CO_2, climate costs of CHF 138.9 are charged (Axhausen et al. 2021). The measurement unit of the response is therefore CHF.

The histogram on the left in Figure 3 also manifests that the distribution of the response is right skewed. The skewness of the distribution has an adverse implication on the prediction accuracy of linear models such as the lasso or the common linear regression. Predictions from these models are asymptotically normally distributed and therefore perform better to predict a response which follows a similar symmetric distribution. An attempt to bring the response in a more favorable shape is the logarithmic transformation. The transformed responses are depicted in the histogram on the right in Figure 3. It turns out that the transformed response is more left skewed than the untransformed response was right skewed (initial skewness of the response 1.734, skewness of the response in natural logarithms -3.924). I therefore refrain from transforming the response which I refer to as the approximated energy use of individuals' mobility in the task at hand.

[6] Another more sophisticated way of condensing the information is to reconstruct typical weekly emissions of an individual. This more complex way of aggregation would have the advantage that it could mitigate a potential selection bias resulting from an individual's choice of the days of the week she or he tracks her- or himself.

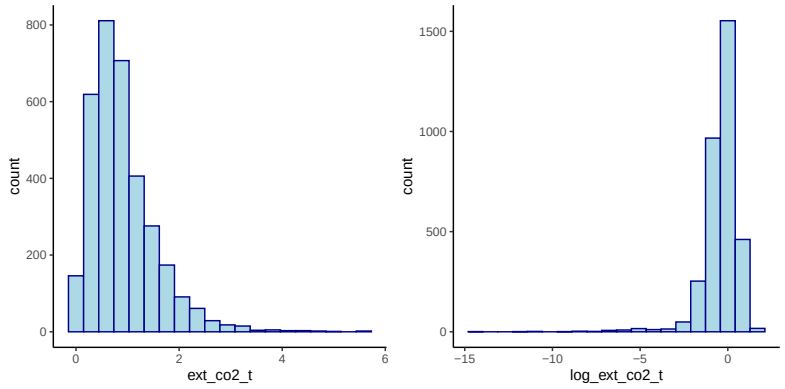

Figure 3: Distribution of the dependent variable average monetized CO_2 emissions per day

Notes: The figure on the left shows the distribution for the untransformed dependent variable. The one on the right for the dependent variable in logarithms.

4.3 Predictors

As candidates for a predictor variable qualify all variables from the source data set which are also available in the target data set. The target data set is restrained to variables which can be surveyed in the SCE. This leaves us with 118 candidates. Among these available candidates I distinguish two groups: the core variables and the supplementary variables.

The core variables are anyway collected in the SCE and hence part of the target data set. This includes general socio-economic variables, as for example, gender, education, income and age, and the employment status with the associated workload. Variables capturing the ownership of mobility tools and the use thereof belong to the core variables as well. A complete list of the 53 core predictors can be found in Table A.1 in the Appendix A.2.

The supplementary variables are the 65 remaining candidate predictors which we only want to survey if they help the imputation, i.e., they improve the prediction of energy use. The variable selection procedure for the survey which is presented in Subsection 5.6 only applies to this group of supplementary variables. The bulk of the supplementary variables covers transport and transport policy related attitudes. They contain information on whether the person thinks that public transport is too crowded or whether they consider phone use while driving a problem. Another subgroup of variables refers to comparisons of the car to public transport. Then there are four variables capturing a person's personality

in the dimensions hedonic, egoistic, altruistic and biospheric. They need to be separately mentioned because they are not directly surveyed. The variables are constructed following the Schwartz Value Survey from 26 variables capturing values, lifestyle and leisure choices as described in Bouman et al. (2018). Section A.2 of the Appendix lists all supplementary variables in Table A.2 and Section A.3 provides an example of a survey question.

4.4 Data pre-processing

The necessary data pre-processing depends on the data structure as well as the method one wants to apply to the data. Of the 118 predictor candidates, 75 are ordinal variables, 33 are nominal variables and 10 follow a numerical scale. Both the lasso and regression forests require the data in numerical form. There are several ways that offer themselves to recode the categorical variables.

Even though the regression forests have a reputation to be insensitive to the characteristics of the training data and the encoding of categorical variables (Hastie et al. 2009, Kuhn and Johnson 2013), Tibshirani et al. (n.d.) emphasizes the importance of meaningful encoding. Which data representation is meaningful depends on the empirical task and can significantly improve the quality of predictions.

For the regression forest, I therefore translate the ordinal variables such that the integer assigned to a category respects the underlying ordering. For the nominal variables, I construct as many dummy variables as there are categories and remove one dummy variable which is then the reference category. This is also the recommended solution for the ordinal variables when they serve as input for lasso.

Missing values for the predictor variables require no special treatment beforehand when applying regression forests. The decision trees in the forest incorporate the missing values meaningfully in each split (Tibshirani et al. n.d.). For lasso, missing values are taken care of by assigning an own dummy variable.

Outliers are of no concern regarding the categorical predictors which are limited to predefined levels. Table 3 supports that the numerical predictors are contained in reasonable ranges as well. The dependent variable `ext_co2_t`, on the contrary, has a few very large values. These values, however, do not contradict the self-stated mobility behavior in the survey. For example, all of the extreme emitters report that they

drive a car on 3 or more days per week. A look at the surveyed mobility behavior also indicates that the values for the six participants which allegedly emit no CO_2 are not measurement errors. The zero emitters report all to drive their car on 3 or more days per week but own an electric car with one exception. The exception is a retired woman which owns a car with combustion engine. Given that the extreme values are traceable I refrain from removing them as outliers and conclude that the distribution of the response is skewed.

Table 3— Summary statistics for the numerical variables

	Mean	Median	Min.	Max.	Std. dev.	n
ext_co2_t	0.91	0.77	0.00	5.58	0.64	3,373
age	41.05	41.00	19.00	66.00	13.62	3373
workload_jobs_main	86.81	100.00	5.00	100.00	22.12	2,673
workload_jobs_second.	21.45	20.00	1.81	80.00	13.92	120
gen_accessibility	1.15	1.23	-8.31	7.25	2.44	3,373
miv_accessibility	0.06	0.10	-0.86	0.82	0.27	3,373
oev_accessibility	0.41	0.48	-2.79	2.31	0.75	3,373
hedonic	3.89	4.00	1.00	5.00	0.70	3,373
egoistic	2.71	2.60	1.00	5.00	0.65	3,373
biospheric	4.01	4.00	1.00	5.00	0.67	3,373
altruistic	3.97	4.00	1.00	5.00	0.63	3,373

While I do not eliminate any participant from the cleaned sample, I eliminate seven predictors. The variables `homeoffice_do_days`, `homeoffice_can_yes`, `homeoffice_can_days` and `own_vehicles_motorbike` are eliminated because they are observed for less than 100 subjects. For this few observations it is improbable that they allow for meaningful model estimation especially after sub-sampling.

The seven eliminated predictors also contain the area codes of the home, the main workplace and the secondary workplace. There are too few observations per area code for these nominal variables to convey any systematic information about emissions. It is more likely that the variables contribute to the prediction as sort of an individual fixed effect. However, it would still make sense to survey the area code of the home. It allows to create the three meaningful accessibility variables listed in Table 3. Based on the area code of the home, its general accessibility as well as its access by public transport and motorized individual traffic is assigned. Another variable with many categories that count low frequencies is `citizen_1`. I do not eliminate this predictor but transform it such that only two categories remain. After transformation `citizen_1`

conveys the information whether a person is Swiss or not. I refrain from further transformation and do not construct any interaction terms between variables. I rather rely on the regression forest to detect important interactions.

5 Methodology

5.1 Prediction

For the imputation task at hand, I want to estimate a suitable model for predicting individual energy use for mobility. I refer to energy use as the response of the model. One way to think about a prediction problem is to imagine a system which receives input variables and generates the response variable as output. The system which relates the input variables to the response is unknown, a black box. Supervised learning uses a known sample of input and output variables to illuminate the dependencies between input and output. This is done without assuming causality running along the emerging input-output dependencies.

I follow an approach to statistical learning which is based on the function approximation framework (for an introduction see Hastie et al. 2009 and James et al. 2013). In this framework, learning consists of approximating the function $f(X)$ that describes the 'true' relationship between the observed inputs x and the response y up to a random error term ϵ. The error term is assumed to be independent of X and has mean zero.

$$Y = f(X) + \epsilon \tag{1}$$

For real valued responses, such as energy use, estimating $f(X)$ from a finite sample is referred to as regression problem. The methods which exist for tackling the regression problem differ in the way they restrict the set of possible functions. Only these restrictions render it feasible to fit a model to a finite sample. The data set used to fit the model is also referred to as the training set.

5.2 Model selection

Model selection implies the comparison of different models. In order to do so a measure is required which captures how well a model predicts the

response for a given set of inputs. The mean squared error (MSE), or the square root thereof, is a common measure for prediction performance in the regression setting.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2}$$

, where $\hat{f}(x_i)$ is the predicted response for the observation $\{x_i, y_i\}$.

The prediction performance of interest is how well the trained model predicts the response of observations which were not used to fit the model, they were thus not part of the training set. This prediction accuracy is also called out-of-sample performance. The focus lies on the out-of-sample performance to prevent the selection of an overfitted model for the prediction task. We speak of model overfit when the model does very well in predicting in-sample (training set) but not out-of-sample. This phenomenon arises when the estimated model is too complex for the limited information in the training set.

Many learning methods allow for calibration of model complexity along one or several so called tuning parameters. The appropriate setting of these parameters is found by minimizing the average out-of-sample error. With limited data available, the average is computed over samples which are from the same data set and are generated by a resampling algorithm. The resampling algorithm I use in this thesis for model selection is k-fold cross validation.

5.3 Model selection procedure

The selection procedure is set up to find the best model among the available models given the considered methods. I propose here a *double resampling* procedure which follows Friedman (1994, cited in Cherkassky and Mulier 2007, pp. 79). The *double resampling* consists of two steps:

Step 1: The available data is split into a training and a test set.

Step 2: The training set is then resampled by cross validation.

For each of the considered methods the cross validation samples from step 2 serve to determine optimal model complexity via the adjustment of the method specific tuning parameters. The selection of one method over the others is based on the mean cross validation error received for the optimally tuned models. The tuned model with the lowest cross

validation error is selected for prediction.

The selected model is then fit to the whole training set and applied to predict the responses of the test set. The deviance between the prediction and the actual response values is summarized in the MSE and can be interpreted as the generalization error. The generalization error indicates how well a model generalizes to unseen data, i.e., independent data. This error is not consulted for model selection because from that point onward the test set ceases to be independent and the selection is potentially biased in favor of models which predict well the responses in the test sample.

5.4 Generalized random forests

This subsection provides a general introduction to decision trees and how they are aggregated to random forests. The introduction is based on Hastie et al. (2009) and James et al. (2013) and should allow to more easily understand generalized random forests. General random forests is applied to the regression problem constituted by the imputation task in Section 6.

Decision trees is a non-parametric method, which is based on a recursive monothetic algorithm for model fitting. The algorithm starts out from what is called the initial node. At the initial node, the training sample is split into two subgroups constituting nodes on their own. The variable along which the split is performed at the node and the corresponding split value is determined in such a way that the created subgroups are as heterogeneous to each other as possible regarding the response. The binary partitioning is continued like this at each node until the halting criterion is met. A typical halting criterion is to fix a minimum number of observations that have to reach a node.

A tree which is grown by this *greedy* algorithm has a finite number of terminal nodes that are not further split. These nodes are represented by mutually exclusive regions in the predictor space framed by the determined monothetic splitting rules. In a region, the response is modeled by averaging over the residing training observations' response. It has to be noted that this averaging can only be applied to a continuous response variable. When the response is continuous, we also speak of a regression tree, and averaging the response has the property to minimize the squared error loss.

The complexity of a regression tree is determined over the tree size. Tree

size is the number of terminal nodes and can be adjusted in a model selection procedure. Nevertheless, single trees stay very sensitive to sampling. They exhibit high model variance meaning that already slight changes in training data leads to different splits and hence predictions. A possible way to reduce variance is to grow multiple trees on bootstrapped samples from the training set and estimate the final prediction by averaging over the grown multiple.[7]

The bootstrap based approach to grow a *forest* of decision trees was crucially improved by Breiman (2001). His idea is to restrict the available splitting variables at each node to a random subsample of the entire set of predictors. Thus, the optimal splitting variable might not be available at a given node for some tree. This makes trees additionally more independent from each other than they already are through the bootstrap sampling. The independence between trees is desirable as only under this condition averaging reduces model variance and hence improves on the generalization of the model. This decision tree based method is called random forests.

Athey et al. (2019) developed random forests further to general random forests (GRF). Broadly speaking, not much changes conceptually with this new development for my application of the method to the regression problem. One generalizing extension is the adaption of the quality measure for the split such that the method can accommodate different problem settings.[8] For the regression problem, the algorithm based on the new measure still maximizes heterogeneity in the response when splitting the training sample. Another change I want to mention concerns prediction. With GRF, the response is predicted based on a weighted list of neighbors which in case of the regression forest, i.e., the application of GRF to a regression problem, amounts to the same prediction as averaging over the prediction of each tree for a given observation.

A property of regression forests is that non-informative predictors do not impair the estimation because they do not have to be considered in any split. The feature that at each node the most promising variable can be chosen from the available set, implies intrinsic variable selection. The information on the number of times a variable is chosen for performing a split and by how much this split reduced the sum of the squared residual is condensed into the variable importance statistic.

Regression forests is an intuitive method to capture non-linear dependen-

[7] Bootstrap based aggregation is an ensemble method referred to as bagging.

[8] Possible applications are quantile regression, heterogeneous treatment effect estimation, instrumental variable regression or panel data analysis.

cies in data. However, the method is limited in approximating additive and linear structures due to a lack of smoothness of the prediction surface.

5.5 Lasso regression

The second method considered in the regression problem is lasso. Lasso is a parametric method which restricts the approximation $\hat{f}(X)$ to be linear in X. Tibshirani (1996) proposes to fit the linear regression model $y_i = x_i'\beta + u_i$ solving the following convex program

$$\begin{aligned} &\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i'\beta)^2 \right\} \\ &\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \ . \end{aligned} \tag{3}$$

A convenient alternative to write down the same optimization problem is the Lagrangian form (Hastie et al. 2009). In this form, the Lagrange multiplier λ presents itself as ideal parameter for tuning model complexity.

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{pmatrix} := \underset{\beta_0,\beta}{\operatorname{argmin}} \underbrace{\frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - x_i'\beta)^2}_{\text{Loss function}} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{Penalty}} \tag{4}$$

The objective function can be read as the composition of a loss function and a penalty. The impact of the penalty is adjusted via the tuning parameter λ. For $\lambda = 0$ the problem is reduced to minimizing the sum of the squared errors which implies that we receive the OLS estimates of the regression coefficients. Increasing λ above zero leads to shrinkage of the coefficients because the penalty term penalizes the model for the absolute value of its coefficients. Due to the nature of this penalty that constrains the minimization, the lasso coefficients do not only approach zero but are actually set exactly zero. This last point makes lasso a popular learning method because it performs intrinsic subset selection.

From the point of view of model selection, λ is determined in order to optimize out-of-sample prediction accuracy. However, the error resulting

from the possibility of trying to approximate a non-linear functional form with a linear model stays.

5.6 Variable selection

Variable importance

Variable subset selection is a built in feature of the two presented methods. In both cases, the selection is a byproduct of the algorithm's optimizing procedure and can inform the analyst over a variable's relative quality as a predictor. For the GRF implementation of regression forests, this information is reported as variable importance. For the lasso, I construct a variable importance measure myself. To do so, I start out from a strongly penalized model where all coefficients are set to zero and then gradually reduce the impact of the penalty by a decreasing sequence of λ. The variable which is associated with the first non-zero coefficient is considered the most important predictor. The variable which is the last one to be considered in the model is the least important predictor. According to this logic all variables can be ranked by this lasso based variable importance measure.

Decision rule for variable selection

On the basis of the two variable importance measures for the considered methods I determine the variables to survey in the SCE, which serve the imputation of energy use.[9] For illustration let us assume that the survey only allows for the inclusion of k additional variables (or questions).

In the model selection process, both methods are evaluated including the additional k most important variables in the training set according to the method specific variable importance measure. This is a necessary condition for a fair comparison between methods when they produce different variable importance rankings.

When the quality of the available candidate predictors is low, it might arise that not even k of them carry valuable information to enhance prediction. To prevent in such a case that potentially worthless variables are surveyed, the decision rule is extended around a random variable benchmark. The random variable benchmark consists of the variable importance that a constructed random variable achieves when added to

[9] For regression forests I average the variable importance over several runs before the application to variable selection because of the stochastic component of the method and its non exhaustive optimization algorithm.

the set of candidate predictors. The extended decision rule of the variable selection is as follows: an additional predictor is chosen if its importance is above the random variable benchmark and the capacity k is not yet exhausted.

Correlation-based variable pre-selection

While the correlation among predictors does not negatively affect the quality of a prediction as long as the correlation persists also outside of the training data (Harrell Jr. 2015), the respective correlation potentially has an impact on variable importance. Imagine two variables which are highly positively correlated because they measure the same underlying information. When only one of the two variables serves as predictor to fit a regression forest, the variable is considered in a certain number of splits and reaches a level of importance accordingly. If now both variables are considered in the model fitting they share the number of splits that belonged before to only one of the two. As a consequence they also score lower regarding variable importance relative to the model where only one of the two correlated variables is included. Variable selection that is based on variable importance can therefore overlook good predictors due to joint correlation. To prevent this, I propose to eliminate one of two correlated variables from the set of candidates before computing the variable importance.

I remove the predictors following a heuristic approach tailored to survey data. The algorithm follows in large parts the one proposed by Kuhn and Johnson (2013, p.47) and proceeds as follows:

1. Compute the correlation matrix between the predictors.
2. Determine the two predictors with the largest absolute pairwise correlation (call them predictors A and B).
3. Determine the number of non-responses for A and B.
4. If A has more non-responses than B, remove it; otherwise, remove predictor B.
5. Repeat steps 2 to 4 until no absolute correlations are above the threshold.

In step four, the number of non-responses is proposed as criterion for elimination. This proposal is based on the idea that questions are selected for the survey that subjects feel able and comfortable to answer.

This approach to predictor elimination is heuristic rather than theoretical as the choice of the threshold is somehow arbitrary. A possibility to learn about what might be a reasonable threshold is to conduct a sensitivity analysis comparing variable importance ranks for alternative choices of the threshold.

6 Application

6.1 Variable selection

A challenge everyone faces who undertakes a survey study is to keep the subjects by good humor during the questioning. With limited budget to pay participants for their answers, restricting the survey in length is important to prevent the quality of the answers to deteriorate or even an attrition of subjects. In this respect, the planned SCE is no different from any other survey, and a reduction to the essential questions is pursued to foster cooperation. For the imputation or prediction task of energy use, this motivation suggests not to blindly survey all candidate predictors from the source data set. The variable selection in this section restricts the set of candidate predictors such that only the most promising predictors are retained. For this purpose I take a closer look at the supplementary variables following the considerations about variable selection in Section 5 on methodology.

The 65 supplementary variables are subject to variable selection. A particularity of these variables we came across in Section 4 is that they all – except for two – capture different attitudes an individual holds regarding transport and transport policy.[10] Some of the attitudes may, however, be driven by common core values and beliefs the individual holds. In this case, different variables can carry the same information. Such circumstances would manifest themselves in a strong correlation between the corresponding variables. In the following variable selection, each single variable of them might then turn out to be of limited importance if jointly considered, even though that these correlated variables turn out to be of high importance if only single ones of them are considered. This consideration motivates to perform a correlation-based variable pre-selection. For this purpose I compute the correlation matrix based on Spearman's rank order correlation because most variables are ordinal. The nominal variables are transformed into dummy variables prior to computing the

[10]Please consider Appendix A.2 for a list of the supplementary variables.

correlation. The set of supplementary predictors therefore extends to 76.

Indeed, the variables correlate up to 0.69 with each other. I apply the algorithm described in Subsection 5.6 to the correlation matrix to eliminate one of two predictors which are correlated more then 0.5 in absolute value. My decision rule removes 16 variables which are listed in Table 4.[11] The non-response criteria for selecting the variable to be excluded from a pair associated with a correlation above the threshold is a reasonable choice with regard to the high variance of non-responses. Among the pairs with high correlation, the number of missing values varies between 0 and 3,270.

Table 4— Excluded supplementary variables in pre-selection process

Variable name
ext_costs_exam_r
transport_policies_extra_lanes
transport_statements_exp_capacit
attitudes_car_pt_5
transport_problems_emissions
transport_policies_red_pub_parki
transport_factors_mobility_price
worktime_No flexibility (fixed start and end time)
attitudes_car_pt_4
return_household_Lowering public transport fares
attitudes_car_pt_6
transport_policies_noise_reg_bik
attitudes_car_pt_7
satisfied_1
transport_policies_car_free_zone
transport_statements_equal_cost_

After the pre-selection, the set of predictors counts 60 supplementary and 46 core predictors. Based on this reduced data set we can compute the variable importance measure without the risk of an unforeseen bias caused by the correlation between the supplementary variables.

Figure 4 shows the top ranked candidate predictors by variable importance based on regression forests. The last predictor on the list is the variable `rand_constructed`. This variable capturing a random vector is exclusively added as benchmark. Every predictor which does not pass

[11] A more detailed table in Appendix A.5 lists additionally the correlated variables which remain in the data set and the size of the correlation.

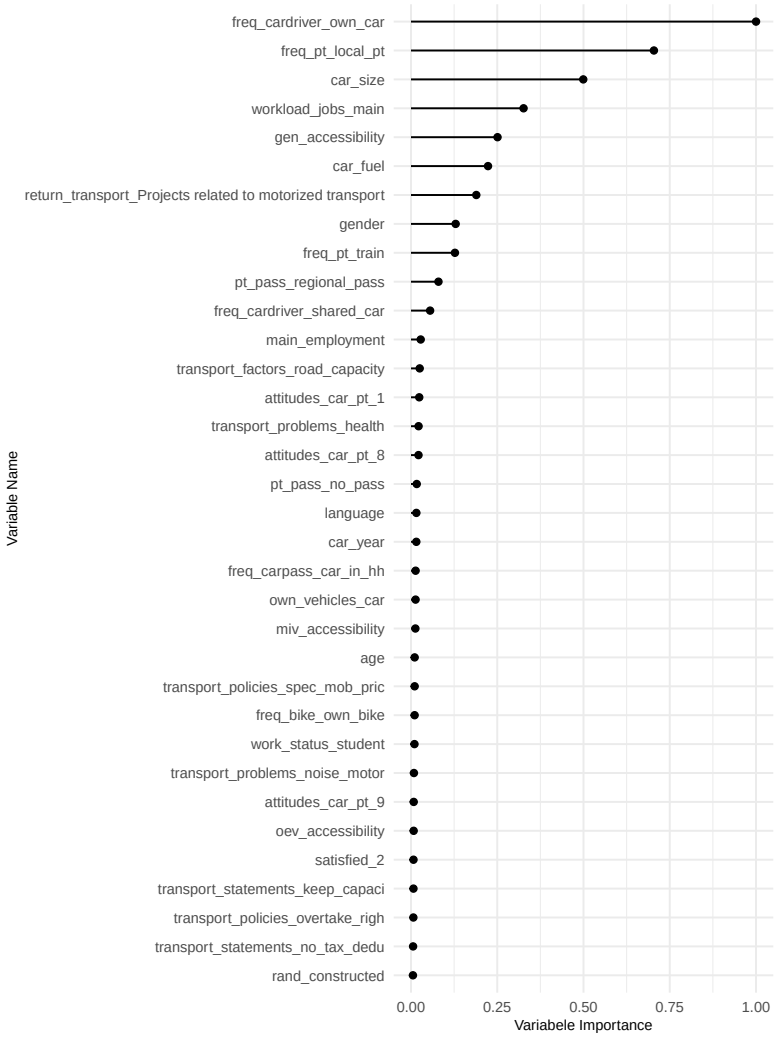

Figure 4: Regression forests variable importance

the benchmark therefore adds less information than a random variable given the source data set. The complete regression forest based ranking and a similar figure for the lasso based variable importance are provided in Appendix A.5.

Based on the rankings according to the two methods, I propose to include k of the supplementary predictors into the survey. For illustration of the remaining steps in the imputation task I assume that another ten variables can be surveyed ($k = 10$). The ten most promising supplementary predictors are listed in Table 5. This short list confirms that the variable importance of a predictor varies with respect to the applied method. When the selected model is based on lasso then four of the ten selected variables are different then when the selected model is based on regression forests.

Table 5— Supplementary variables ranked by variable importance

	Regression forest	Lasso
1	return_transport	return_transport
2	transport_factors_road_capacity	transport_problems_health
3	attitudes_car_pt_1	attitudes_car_pt_8
4	transport_problems_health	satisfied_2
5	attitudes_car_pt_8	transport_factors_road_capacity
6	transport_policies_spec_mob_pric	transport_factors_pt_price
7	transport_problems_noise_motor	revenue
8	attitudes_car_pt_9	transport_policies_overtake_righ
9	satisfied_2	transport_problems_speeding
10	transport_statements_keep_capaci	transport_problems_noise_motor

A closer look at the selected supplementary variables raises the question of whether it is possible to survey them in stand alone questions. For example, the most promising predictor `return_transport` holds the answer to the question in what kind of transport infrastructure the revenue generated by a mobility pricing scheme should be invested. In the final survey of the MOBIS experiment this question follows another one that introduces the subject to the hypothetical situation where a mobility pricing scheme is in place and asks more generally where the money should go. Transport related projects are one choice option among three.

In this regard it, is important to be aware of the original context of the questions when the selected variables are surveyed in the SCE. A change of a question's context might lead to a different understanding of it or can activate different associations which then are reflected in subjects' answers. As the variables vary in sophistication with respect

to the context they require, it might be of interest to substitute more sophisticated variables. Some suitable substitutes have potentially been eliminated from the data set in the pre-selection.

The correlation of `return_transport` with another supplementary variable did not lead to the exclusion of the latter in the pre-selection. This can be concluded from the comparison of Table 5 to Table A.3. Overall led the correlation with six of the 14 different selected supplementary variables to an exclusion in the pre-selection.

6.2 Model selection

The model selection follows the plan rolled out in Section 5.3. First, the data is split into a training and a test set. Second, the lasso and regression forests are tuned making use of resampling from the training set. Third, the calibrated models from both methods are compared by their mean cross validation error (MCVE).

Of the total observations, I put a random selection of one fourth aside as the test set. The test set thus counts 843 observations. The remaining 2,530 observations constitute the training set.

The GRF implementation of regression forests is accompanied by a tuning function. The function determines five of the method's parameters by cross validation (Tibshirani et al. n.d.). The tuning output is presented in Appendix A.7. Additional parameters of a regression forest can be subject to tuning once you allow for honest trees. I disable the option to grow honest trees because it requires a further sample split of the already small training sample.[12]

The cross validation procedure that determines optimal model complexity of the lasso can easily be traced as only one parameter is tuned. For large values of that one tuning parameter λ, the coefficient space is strongly constrained. With diminishing values of λ the coefficient space is opened up and the predictor coefficients approach the value they would take in an unconstrained linear regression. This fact is also reflected by the horizontal lines in Figure 5. On the x-axis, λ increases from left to right which leads to ever sparser models. The sparsity of the estimated models is reflected in the upper horizontal line which indicates the number of non-zero coefficients associated with the predictors. The sequence

[12]For the data set at hand, a model with honest trees and otherwise equal setting is dominated by a regression forest based on conventional trees with regard to the validation error. The corresponding results are available on request.

of λ for which a MCVE is computed based on 10-fold cross validation is selected by GLMNET, an R implementation of lasso (Friedman et al. n.d.). The proposed sequence counts 100 different values for λ and for each value are ten models fitted for cross-validation. The information of the altogether 1,000 fitted models is summarized in the mean-squared cross validation error which is plotted in Figure 5 with bands indicating plus minus one standard deviation.

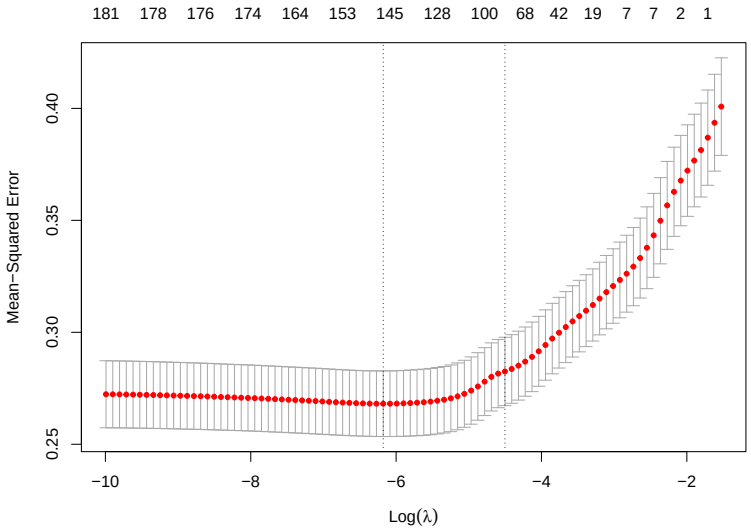

Figure 5: Optimal model complexity for lasso: Mean cross-validation error

The optimal value of λ according to this procedure is indicated by the first dashed line from the left in Figure 5. The indicated λ is around 0.002 and minimizes the validation error with a model that counts 148 non-zero coefficients.[13] The more to the left from this point, the poorer the models perform due to overfitting. The more to the right from the minimum, the poorer the models perform because they fall short to capture important dependencies due to the enforced simplicity.

Both methods are now calibrated so that they deliver accurate out-of-sample predictions. I take the tuned parameters and compute for both methods again the MCVE by 10-fold cross validation. This way I can, on the one hand, guarantee that the cross validation is based on the same ten folds from the training set. On the other hand, it additionally allows controlling the formula which is used to compute the MCVE. The com-

[13]The GLMNET summary output of the cross validation procedure is presented in Appendix A.7.

parability of model performance across methods that is thereby achieved, renders model selection between methods possible. The obtained cross validation results are listed in Table 6.

A direct comparison of MCVEs favors the lasso model over regression forests. I select on this ground the lasso model for the prediction task. However, the null hypothesis that there is no difference in the MCVE cannot be rejected according to a two sample t-test.

Table 6— Model selection between methods

	Regression forests	Lasso
MCVE	0.297	0.268
Std. dev.	0.016	0.012
CI lower bound	0.261	0.242
CI upper bound	0.333	0.295
Generalization error	0.319	0.305
In-sample error	0.244	0.239

*N*ote: The confidence intervals CI are constructed for a 5% significance level assuming that the MCVE follows a Student-distribution with 9 degrees of freedom.

After the tuning by cross validation I fit a model for both methods to the entire training data set. These models are then applied to predict the responses of the training set and the test set. The first application is used to compute the in-sample error. The second application to the test data is, in a sense, the operationalization of the models. They are used to predict responses which were not used to train the model. The prediction error over an independent test sample is for this reason also called the generalization error. It indicates how well a model generalizes to unseen data. For both methods the comparison of the generalization error to the in-sample error indicates that the models overfit.

The model selection statistics also suggest that the linear model maps the dependencies between the predictors and energy use better than the non-linear model does. This is surprising as the predictions with lasso are forced to be symmetric around the mean, while the predictions from the regression forest are more flexible to capture the skewness in the distribution of the response. However, it is conceivable that more observations would have turned the selection in favor of regression forests. The additional information would have admitted to utilize more of GRF's features that reduce overfitting such as honest trees.

If additional methods were considered in the model selection, MARS

would have offered itself.[14] This non-linear regression method is better suited to model additive structures than regression trees and thus allows for a smoother mapping of the underlying function.

6.3 Prediction model

To get familiar with the selected prediction model I look at two distinct aspects of it, which are important in the MOBIS setting. The first aspect is the imputation of the level of energy use in the SCE. The second aspect is the prediction of differences in energy use based on the answers subjects give in the SCE. The second aspect is illustrated by means of the registration tax example introduced in the theory section.

The imputation task is the prediction of energy use over an independent target data set. The most honest indicator for the prediction accuracy of this application provides the generalization error. To get a better understanding for the generalization error I take the square root of it. The transformed error then has the same unit as energy use. In our application, the square root of the generalization error amounts to $\sqrt{0.305} = 0.52$. It thus adds up to around one half of the average daily energy use of 0.91. Table 7 puts this performance in perspective. The table lists three other linear prediction models that can be adopted to impute energy use in the SCE.

Table 7— Comparison of prediction models

	LR0	LR10	Core lasso	Lasso
Generalization error	0.428	0.341	0.307	0.305
In-sample error	0.401	0.302	0.247	0.239
No. non-zero coeffs	0	10	109	148

The first model LR0 in Table 7 approximates an individual's energy use simply by the same constant for everyone, which amounts to the average daily energy use when the response is regressed on a constant. The next model LR10 in Table 7 is an OLS regression including ten predictors. The first five predictors are the top ranked according to the variable importance derived from lasso. The remaining five predictors are the dummy variables capturing the information regarding the size of an individual's car. The model *core lasso* is the optimally tuned lasso model when only the core predictors are considered. A comparison of the generalization

[14] Hastie et al. (2009) give an introduction to multivariate adaptive regression splines (MARS).

errors between the *core lasso* and the selected lasso model presented in the fourth column shows that ten additionally surveyed variables improve the prediction but not by much. It is not surprising however, that the improvement of the prediction through the inclusion of the predominately attitudinal variables is modest. The core values and beliefs behind these attitudes are in mutual dependency with the preference for the mobility tools and the intensity with which they are used. Both, the mobility tools and the frequency of their use are already considered in the set of core predictors.

Before going over to the analysis of the potential effect of a rise in the registration tax on energy use, we have a look on the existing car fleet. In the fleet five groups of cars are distinguished by size. The categories are listed in Table 8 together with the number of respondents that own a car from this category and the lasso coefficient associated with the category. The coefficients indicate for each group by how much the energy use deviates from the reference group which accounts for the energy use of an owner of a small car. The magnitudes of the coefficients match the expectation one would form based on the weight and horse power that is normally associated with each group. However, the coefficients also convey information on the mobility behavior of a typical exponent of the respective group. For example, there might exist a link that people who drive longer distances in their car are also more likely to buy a larger car that provides more comfort. It can be thought of many more such examples that suggest endogeneity of the coefficient estimates caused by omitted variables. While this fact does not limit the predictive power of the model in any way, it has to be kept in mind when interpreting the coefficients.

Table 8— Car size and predicted energy use for the existing fleet

Car size	n	Lasso coef.
Small car	843	-
Medium to large car	1,349	0.187
Minivan or van	226	0.427
Luxury car or sports coupé	85	0.451
Off-road vehicle	456	0.442
No own car	414	0.002

Starting from the existing fleet, the average energy use is at 0.91. Imagine that this is the state for a given level of the registration tax. Now, after a policy change, the world enters a state where heavier and more powerful cars are costlier due to a rise in the registration tax. For simplicity,

we assume that in the new state everyone switches to own a small car as a reaction to the tax increase. All else equal, the predicted energy use for individual mobility drops by 20%, on average, for the sample at hand. The average energy use in this new equilibrium is 0.73. The light blue bar in the center of Figure 6 depicts the predicted lower level of energy consumption and sets it in relation to the observed sample average represented by the red line.

A similar thought experiment in the same context is to assume that an otherwise average respondent owns a luxury car previous to the rise in the registration tax. The model predicts for this individual's mobility behavior an average daily energy consumption of 1.19. This level is indicated by the bar in dark blue in the center of Figure 6. Now the two bars can be compared. They reflect a scenario in which – after the increase in the registration tax – the respondent adjusts his mobility toolbox in the direction of the tax incentive, and changes to a small car. The prediction model puts the energy use accordingly down to the level indicated by the light blue bar of 0.73. The predicted change in energy use by the model amounts to $1.185 - 0.734 = 0.451$. Not surprisingly, this is exactly the magnitude of the lasso coefficient associated with the category luxury car.

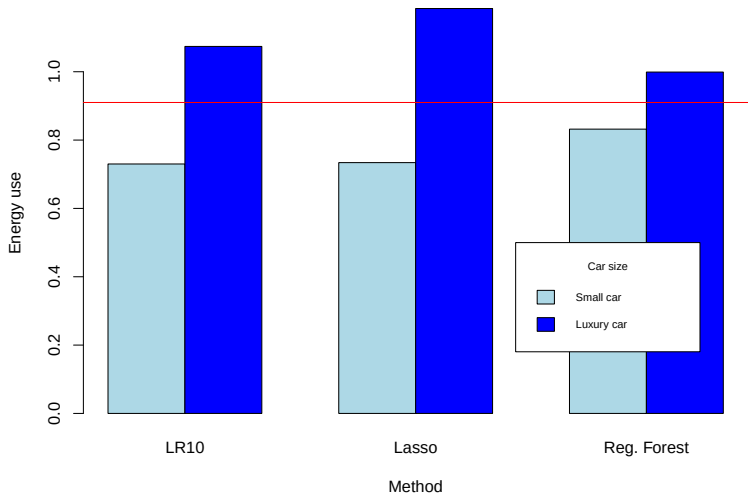

Figure 6: Prediction of energy use

The last line of the thought experiment exposes the weakness of the model application to predict differentials in energy use. With the prediction of energy differentials we enter the realm of causal inference. Unlike the

imputation of the level of energy use in an environment with given and stable restrictions, the endogeneity of the coefficient estimate limits the explanatory power of predictions when alternative regimes of restrictions are assumed. The predicted change is only valid under the two conditions derived in the theory section. The first condition is that all other attributes of the mean respondent that figure in the prediction model apart from the car size remain constant. The second condition is that the mean respondent adopts all the unobserved behavioral patterns (or omitted variables) that are typical for a small car owner and are captured by the car size variable. To what extent these conditions hold has to be carefully assessed in every application of this imputation/prediction strategy.

7 Conclusion

In this thesis, an imputation setting is examined in the context of the prediction of individual energy use for mobility. It has led to the development of a variable subset selection procedure which precedes the conventional model selection. In addition to the imputation of energy use, the selected model can be applied to predict differences in energy use with regard to adaptation of individual endowments, for example, in the mobility tools.

The results indicate that the developed variable selection procedure serves its end. It allows making an informed choice on the predictors to include. For this purpose, the proposed procedure can also be generalized to other applications. However, the procedure is no substitute for a good understanding of the prediction data. It cannot be pursued in a mechanical way. Moreover, the added value of the procedure is still to be explored further. In particular, sensitivity checks would allow statements about the extent to which the procedure can enhance prediction performance.

In the current study, lasso provides the relatively best prediction model according to the conducted model selection. Of course, this statement is not conclusive for the modeling of individual energy use for mobility. The statement is made based on the prediction data of the application and with regard to the considered methods. Within the selected model, self-stated travel behavior is most predictive for energy use. This might not come as a surprise. In contrast, and more surprising, attitudinal variables improve the prediction by relatively little. This finding has to be treated with caution though as the model has a limited prediction capacity for changes in energy use. Attitudes can potentially predict more when individuals' adaptation to a regime change is observed.

Overall, the analysis suggests that data based avenue with machine learning holds a large potential for imputation tasks across different research designs.

References

Alberini, A. and Bareit, M.: 2019, The effect of registration taxes on new car sales and emissions: Evidence from Switzerland, *Resource and Energy Economics* **56**, 96–112.

Athey, S. and Imbens, G. W.: 2019, Machine learning methods that economists should know about, *Annual Review of Economics* **11**, 685–725.

Athey, S., Tibshirani, J., Wager, S. et al.: 2019, Generalized random forests, *The Annals of Statistics* **47**(2), 1148–1178.

Axhausen, K. W., Hintermann, B., Castro, A., Dubernet, T., Götschi, T., Molloy, J., Schoeman, B., Tschervenkov, C. and Tomic, U.: 2021, Empirical analysis of mobility behaviour in the context of dynamic pricing, final report of the mobis project, *ASTRA Publikationen* (in preparation).

Bouman, T., Steg, L. and Kiers, H. A.: 2018, Measuring values in environmental research: a test of an environmental portrait value questionnaire, *Frontiers in Psychology* **9**, 564.

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

Bundesamt für Statistik: 2016, *Mobilität und Verkehr*, Bundesamt für Statistik, Neuchâtel.

Button, K.: 2010, *Transport economics*, 3rd edn, Edward Elgar, Cheltenham, UK.

Cherkassky, V. and Mulier, F. M.: 2007, *Learning from data: concepts, theory, and methods*, 2nd edn, John Wiley & Sons, Hoboken.

Enzler, H. B. and Diekmann, A.: 2019, All talk and no action? An analysis of environmental concern, income and greenhouse gas emissions in Switzerland, *Energy Research & Social Science* **51**, 12–19.

Friedman, J. H.: 1994, An overview of predictive learning and function approximation, *in* V. Cherkassky, J. H. Friedman and H. Wechsler (eds), *From statistics to neural networks*, Springer, Berlin, pp. 1–61.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K. and Simon, N.: n.d., Lasso and elastic-net regularized generalized linear models, *https://glmnet.stanford.edu/index.html* (accessed December 18, 2020).

Harrell Jr., F. E.: 2015, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, 2nd edn, Springer, Cham.

Hastie, T., Tibshirani, R. and Friedman, J.: 2009, *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn, Springer, New York.

Hensher, D. A., Rose, J. M., Rose, J. M. and Greene, W. H.: 2005, *Applied choice analysis*, 2nd edn, Cambridge University Press, Cambridge.

Ho, C. Q., Mulley, C. and Hensher, D. A.: 2020, Public preferences for mobility as a service: Insights from stated preference surveys, *Transportation Research Part A: Policy and Practice* **131**, 70–90.

James, G., Witten, D., Hastie, T. and Tibshirani, R.: 2013, *An introduction to statistical learning*, Springer, New York.

Kowald, M., Kieser, B., Mathys, N. and Justen, A.: 2017, Determinants of mobility resource ownership in Switzerland: Changes between 2000 and 2010, *Transportation* **44**(5), 1043–1065.

Kuhn, M. and Johnson, K.: 2013, *Applied predictive modeling*, Springer, New York.

Stoiber, T., Schubert, I., Hoerler, R. and Burger, P.: 2019, Will consumers prefer shared and pooled-use autonomous vehicles? A stated choice experiment with Swiss households, *Transportation Research Part D: Transport and Environment* **71**, 265–282.

Tibshirani, J., Athey, S. and Wager, S.: n.d., Generalized random forests, *https://grf-labs.github.io/grf/index.html* (accessed December 18, 2020).

Tibshirani, R.: 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Yan, S. and Eskeland, G. S.: 2018, Greening the vehicle fleet: Norway's CO_2-differentiated registration tax, *Journal of Environmental Economics and Management* **91**, 247–262.

Appendix

A.1 MOBIS study inclusion criteria

The inclusion criteria for the MOBIS experiment out of the study's final report by Axhausen, Hintermann, Castro, Dubernet, Götschi, Molloy, Schoeman, Tschervenkov and Tomic (2021): "...identify subjects who qualified for the main study based on the following inclusion criteria:

- To be the recipient of the personal invitation letter (the invitation was not transferable to other persons)
- To live in a metropolitan area in the German- or French-speaking part of Switzerland (the lists of addresses included only people living in these areas but the survey doublechecked the post code)
- To be between 18 and 65 years old in 2018 (the list of addresses provided by the BFS was pre-filtered by age at this year)
- To travel by car at least two weekdays per week (including their own car, car-sharing as a driver, or with a taxi and App-based services such as Uber as passenger)
- To use of a smartphone that can install the tracking app
- To be able to walk 200 meter without assistance (to ensure that participants have free mode choice)
- To not work as a professional driver (to ensure that participants have free mode choice)"

A.2 List of predictors

Table A.1— Core predictors from the source data set

Core predictors	
Postcode based	pt_pass_other
postcode_home	pt_pass_no_pass
postcode_jobs_main	Car characteristics
postcode_jobs_secondary	car_fuel
gen_accessibility	car_year
oev_accessibility	car_size
miv_accessibility	Work related
Use of mobility tools	work_status_employed
freq_cardriver_own_car	work_status_self_employed
freq_cardriver_shared_car	main_employment
freq_carpass_car_in_hh	work_status_unemployed
freq_carpass_car_pooling	work_status_apprentice
freq_carpass_taxi	work_status_student
freq_carpass_app_based	work_status_retired
freq_pt_train	work_status_other
freq_pt_local_pt	workload_jobs_main
freq_bike_own_bike	workload_jobs_secondary
freq_bike_own_ebike	homeoffice_do_yes
freq_bike_bike_sharing	homeoffice_do_days
Ownership of mobility tools	homeoffice_can_yes
own_vehicles_motorbike	homeoffice_can_days
own_vehicles_bicycle	Socio economic
own_vehicles_car	language
bike_type_regular	gender
bike_type_ebike_45	education
bike_type_ebike_25	income
pt_pass_ga	household_size
pt_pass_half_fare	citizen_1
pt_pass_regional_pass	age
pt_pass_track_7	

Table A.2— Supplementary predictors from the source data set

Supplementary predictors	
Transport related attitudes	transport_problems_crowding_pt
transport_policies_spec_mob_pric	transport_problems_speeding
transport_policies_red_speed	transport_problems_dui
transport_policies_dyn_limits	transport_problems_distracted_dr
transport_policies_extra_lanes	transport_problems_risk_ped
transport_policies_bus_lanes	transport_problems_risk_cyc
transport_policies_exp_cyc	transport_problems_risk_dri
transport_policies_car_free_zone	Trade-off car vs. public transport
transport_policies_incr_park_cos	attitudes_car_pt_1
transport_policies_subs_etraffic	attitudes_car_pt_2
transport_policies_overtake_righ	attitudes_car_pt_3
transport_policies_noise_reg_bik	attitudes_car_pt_4
transport_policies_red_min_age	attitudes_car_pt_5
transport_policies_red_pub_parki	attitudes_car_pt_6
transport_statements_exp_capacit	attitudes_car_pt_7
transport_statements_social_cost	attitudes_car_pt_8
transport_statements_gov_pt	attitudes_car_pt_9
transport_statements_pt_dyn_pric	attitudes_car_pt_10
transport_statements_equal_cost_r	attitudes_car_pt_11
transport_statements_equal_cost_i	attitudes_car_pt_12
transport_statements_no_public_f	attitudes_car_pt_13
transport_statements_red_gov_int	Diverse
transport_statements_no_tax_dedu	satisfied_1
transport_statements_keep_capaci	satisfied_2
transport_factors_mobility_price	ext_costs_exam
transport_factors_road_capacity	revenue
transport_factors_pt_capacity	return_household
transport_factors_fuel_price	return_transport
transport_factors_pt_price	worktime
transport_problems_congestion	Value framework Bouman et al. 2018
transport_problems_emissions	hedonic
transport_problems_health	egoistic
transport_problems_too_much_trav	altruistic
transport_problems_noise_motor	biospheric
transport_problems_noise_pt	

A.3 Example of a survey question

For the variable `transport_policies_spec_mob_pric` the corresponding survey question is listed with the possible answers.

> Please indicate whether you agree or disagree with the policy.
> Time- and route-specific mobility pricing, made revenue-neutral by lowering other taxes.
>
> Possible answers:
>
> - Strongly disagree
> - Disagree
> - Neither disagree nor agree
> - Agree
> - Strongly agree

A.4 Data cleaning

The data cleaning procedure follows these five briefly described steps:

1. Merge different data files.
2. Identify the response and variables that can serve as predictor (are suitable for survey study).
3. Check variable types.
4. Bring the categories of the ordinal variables in the order that the information they hold implies.
5. Aggregate the observations by individual.

A.5 Variable pre-selection output

Table A.3— Variable pre-selection by cross-wise correlation

To Exclude	To Remain	Corr.
ext_costs_exam_r	ext_costs_exam_w	0.69
transport_policies_extra_lanes	transport_statements_keep_capaci	0.67
transport_statements_exp_capacit	transport_statements_keep_capaci	0.66
attitudes_car_pt_5	attitudes_car_pt_6	0.66
transport_problems_emissions	transport_problems_health	0.66
transport_policies_red_pub_parki	transport_policies_incr_park_cos	0.62
transport_factors_mobility_price	transport_factors_pt_price	0.60
worktime_No flexibility (fixed start and end time)	worktime_Some flexibility (flexible start and/or end time, but completing a set number of hours per day)	0.60
attitudes_car_pt_4	attitudes_car_pt_3	0.58
return_household_Lowering public transport fares	return_household_Returning the same amount to everyone (e.g., by lowering health insurance premia)	0.58
attitudes_car_pt_6	attitudes_car_pt_7	0.56
transport_policies_noise_reg_bik	transport_problems_noise_motor	0.55
attitudes_car_pt_7	attitudes_car_pt_8	0.55
satisfied_1	satisfied_2	0.52
transport_policies_car_free_zone	transport_policies_exp_cyc	0.51
transport_statements_equal_cost_	transport_statements_pt_dyn_pric	0.50

A.6 Variable importance

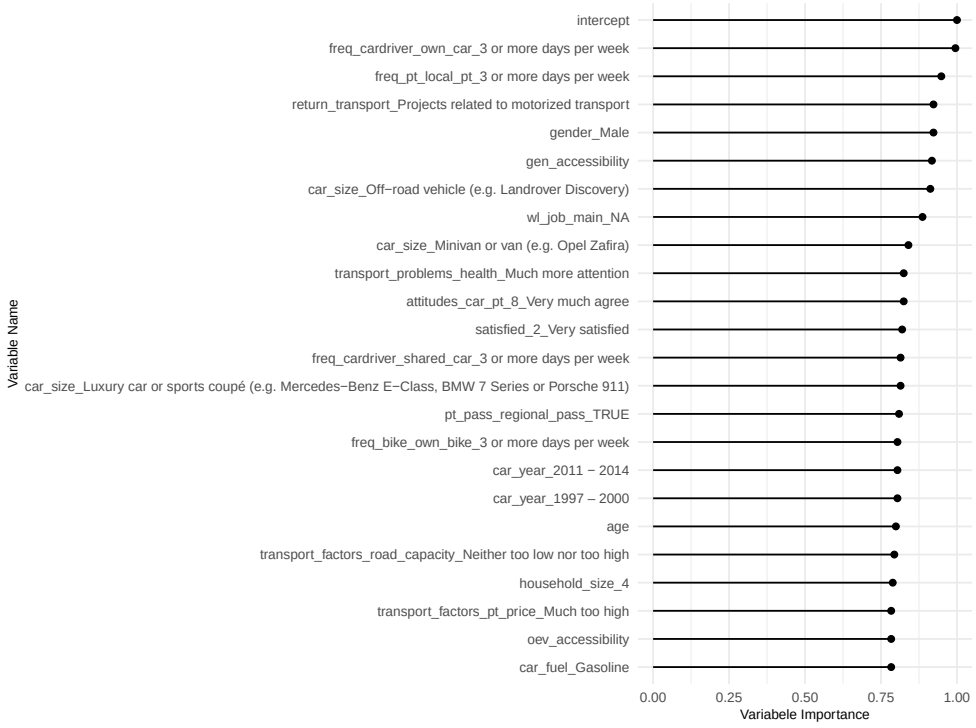

Figure A.1: Lasso variable importance

Table A.4— Predictors ranked by regression forest variable importance

Variable Part I	Importance
freq_cardriver_own_car	1.00
freq_pt_local_pt	0.70
car_size	0.50
workload_jobs_main	0.33
gen_accessibility	0.25
car_fuel	0.22
return_transport_Projects related to motorized transport	0.19
gender	0.13
freq_pt_train	0.13
pt_pass_regional_pass	0.08
freq_cardriver_shared_car	0.06
main_employment	0.03
transport_factors_road_capacity	0.03
attitudes_car_pt_1	0.02
transport_problems_health	0.02
attitudes_car_pt_8	0.02
pt_pass_no_pass	0.02
language	0.02
car_year	0.02
freq_carpass_car_in_hh	0.01
own_vehicles_car	0.01
miv_accessibility	0.01
age	0.01
transport_policies_spec_mob_pric	0.01
freq_bike_own_bike	0.01
work_status_student	0.01
transport_problems_noise_motor	0.01

Note: The importance values are normalized such that the highest ranked variable has an importance of 1.

The variable importance ranking according to lasso can be provided on request.

Variable Part II	Importance
attitudes_car_pt_9	0.01
oev_accessibility	0.01
satisfied_2	0.01
transport_statements_keep_capaci	0.01
transport_policies_overtake_righ	0.01
transport_statements_no_tax_dedu	0.01
rand_constructed	0.01
household_size	0.01
transport_statements_no_public_f	0.01
egoistic	0.01
transport_factors_pt_capacity	0.00
transport_policies_bus_lanes	0.00
freq_carpass_car_pooling	0.00
freq_carpass_app_based	0.00
biospheric	0.00
education	0.00
transport_statements_social_cost	0.00
transport_policies_subs_etraffic	0.00
work_status_employed	0.00
transport_factors_fuel_price	0.00
attitudes_car_pt_3	0.00
transport_problems_speeding	0.00
transport_policies_red_speed	0.00
altruistic	0.00
transport_statements_pt_dyn_pric	0.00
freq_bike_bike_sharing	0.00
attitudes_car_pt_10	0.00
transport_problems_dui	0.00
transport_policies_exp_cyc	0.00
transport_statements_red_gov_int	0.00
transport_problems_risk_cyc	0.00
attitudes_car_pt_2	0.00
transport_problems_congestion	0.00
return_transport_Projects related to public transport	0.00
transport_factors_pt_price	0.00
transport_problems_distracted_dr	0.00
pt_pass_ga	0.00
transport_problems_too_much_trav	0.00
transport_problems_noise_pt	0.00
transport_problems_risk_dri	0.00
transport_policies_red_min_age	0.00
transport_policies_incr_park_cos	0.00

Variable Part III	Importance
hedonic	0.00
income	0.00
revenue_The money should be returned to households	0.00
freq_carpass_taxi	0.00
transport_statements_gov_pt	0.00
attitudes_car_pt_12	0.00
v92	0.00
transport_problems_crowding_pt	0.00
transport_policies_dyn_limits	0.00
return_transport_Projects related to bicycling	0.00
transport_problems_risk_ped	0.00
freq_bike_own_ebike	0.00
attitudes_car_pt_13	0.00
work_status_self_employed	0.00
attitudes_car_pt_11	0.00
return_household_Returning the same amount to everyone (e.g., by lowering health insurance premia)	0.00
pt_pass_half_fare	0.00
pt_pass_other	0.00
own_vehicles_bicycle	0.00
workload_jobs_secondary	0.00
work_status_apprentice	0.00
work_status_other	0.00
ext_costs_exam_w	0.00
revenue_The money should be used to fund new transport-related projects	0.00
bike_type_ebike_45	0.00
bike_type_regular	0.00
homeoffice_do_yes	0.00
worktime_Some flexibility (flexible start and/or end time, but completing a set number of hours per day)	0.00
worktime_NA	0.00
bike_type_ebike_25	0.00
work_status_unemployed	0.00
revenue_Other (please specify)	0.00
return_household_Lowering existing taxes that are unrelated to transport (e.g., value added tax)	0.00
work_status_retired	0.00
return_household_Other (please specify)	0.00
citizen_1	0.00
pt_pass_track_7	0.00
return_transport_Projects related to walking	0.00

A.7 Tuning output regression methods

The lasso tuning parameter which optimizes out-of-sample performance of the model is found by cross validation which is implemented in the `cv.glmnet` function of the `glmnet` package (Friedman et al. n.d.). The function output summary below provides the λ which minimizes the mean cross validation error.

```
-----------------------------------------
Call:  cv.glmnet(x = xs, y = y, alpha = 1)

Measure: Mean-Squared Error

    Lambda   Measure SE      Nonzero
min 0.002076  0.2681 0.01464     148
1se 0.011080  0.2825 0.01527      81
-----------------------------------------
```

The function `tune_regression_forest` tunes five parameters of the regression forests model from the `grf` package (Tibshirani et al. n.d.). The five parameters are listed in the following shortened tuning output summary.

```
-----------------------------------------
Tuning status: tuned.
This indicates tuning found parameters that
are expected to perform better than default.

Predicted debiased error: 0.275325005596998

Tuned parameters:
sample.fraction: 0.416843527555466
mtry: 13
min.node.size: 2
```

```
alpha: 0.0110047467169352
imbalance.penalty: 0.923230990276616

Average error by 5-quantile:

[...]
-------------------------------------------
```