

**A Transfer Learning-Based Framework for
Enriching National Household Travel Survey Data with Attitudinal Variables**

Aliaksandr Malokin

(530) 902-0498, amalokin@gatech.edu

Patricia L. Mokhtarian

(404) 385-1443, patmikh@gatech.edu

and

Giovanni Circella

(530) 554-0838, giovanni.circella@ce.gatech.edu

School of Civil and Environmental Engineering
Georgia Institute of Technology
790 Atlantic Drive
Atlanta, GA 30332

January 2019

ABSTRACT

Often in practice, the problem of unavailability of specific desired knowledge within one (“target”) dataset arises. However, if this knowledge can be extracted from a different (“source”) dataset and transferred between the datasets, this could increase the value of the target dataset at relatively minimal cost. The goal of this paper is to evaluate approaches to informing one dataset with knowledge from another and to evaluate the performance of the knowledge transferred into the target dataset. We use the 2009 National Household Travel Survey as the target dataset. The missing knowledge is transportation-related attitudes, whose inclusion could greatly improve travel behavior models. Our source dataset is obtained from the 2011–12 Multitasking Survey of Northern California Commuters. To achieve the goal, the set of common variables was first augmented with a large number of built-environment attributes. Then, after applying machine-learning methods, *pro-transit*, *pro-active transportation*, and *pro-density* attitudinal factor scores were predicted with the greatest precision; correlations of the predicted and observed scores were 0.564, 0.538, and 0.571, respectively. The performance of the transferred attitudes was measured by estimating linear regression models of vehicle ownership. The results showed that in the source dataset the observed attitudes account for an 8.0% model lift (improvement in goodness of fit), while in the target dataset the predicted attitudes account for a 1.2–5.4% model lift. Although these initial results are modest, we believe they show substantial promise, and the process has identified a number of opportunities for improvement and further research.

Key words: transfer learning, statistical matching, data fusion, machine learning, attitudes, NHTS, vehicle ownership

1 INTRODUCTION

Travel demand forecasting and travel behavior modeling experience both the benefits and disadvantages associated with the increased data availability of the information age. Embracing new data acquisition techniques, such as GPS-based trajectory records of movement, smartphone geolocation, Bluetooth, and Near Field Communication sensing, has been a pioneering effort that allows gathering more travel behavior data while keeping the respondents' burden at a minimum (e.g., Chen et al., 2016). However, many important factors that influence where and how people travel lie outside of manifest travel behavior dimensions, and are still mainly collected in the form of self-reported, disaggregate survey data. Among these factors, we consider lifestyles, attitudes, motivations, intentions, and similar constructs to be especially critical.

Despite the development of internet-based surveys and smartphone-based lightning polls, a crucial problem with this type of data still exists: there is a direct relationship between the amount of useful information to be collected from respondents and their resource burden during this process, and correspondingly an inverse relationship between that burden and the likelihood of obtaining the desired information. For decades, a quest for the optimal balance, given fixed (and modest) budgets, forced investigators to target narrower topics and sacrifice breadth for depth (or, vice versa) with respect to the collected information. For example, the 2009 National Household Travel Survey (NHTS), which surveyed more than 150,000 households in all 50 US states, collected mainly socio-economic characteristics and observed travel behavior attributes. Alternatively, numerous researchers collect much smaller samples, generally within a limited geographical area, studying travel behavior phenomena and measuring numerous explanatory variables that are not captured by the NHTS.

In this study, we implement and evaluate a number of methods for using a sample containing attitudinal measures among other variables (the “source dataset”), to predict attitudes for the observations in an unrelated dataset (the “target dataset”). The choice of the NHTS as the target for the transferred information was motivated by its importance to many transportation studies in the United States and its value to the agency funding this work. The attitudinal data source is the travel-multitasking survey administered by the authors in Northern California in 2011-2012 (referred to as the Multitasking Survey of Northern California Commuters – MSNCC, in the remainder of the paper).

To inform one dataset (NHTS) with the information available in another (MSNCC) and evaluate the performance of this process, we propose two separate frameworks. The first one is the *transfer learning framework*. It is tasked to robustly evaluate the performance of predicting functions given the knowledge to be transferred (attitudes) and the pool of common variables (socio-economic and land use). The second one is the *external validation framework*. In the context of the target dataset, it assesses how valuable the transferred knowledge is for model building. These frameworks are developed to be readily transferrable beyond the context of this study and can be applied in various settings where one dataset is merged with the variables from another via statistical inferences.

The rest of this paper is organized as follows: Section 2 formally defines transfer learning and provides some background on statistical matching, data fusion, and key machine learning concepts and on the methods that are used in this study. Section 3 identifies the working substrate of this study (the NHTS and MSNCC datasets) and establishes the transfer learning and external validation frameworks, which are responsible for enriching the NHTS dataset with attitudes and evaluating their performance, respectively. Practical details of applying the transfer learning

framework and the subsequent results are laid out in Section 4. In Section 5, we describe and discuss the results of the external validation exercise, implemented as a vehicle ownership model. Section 6 summarizes the results of the study and highlights avenues for further research. Following references to the cited literature and acknowledgements, in the appendix we provide expanded detail on the literature review (offering something of a mini-tutorial on transfer learning methods, for readers who may be unfamiliar with them), our transfer learning application for categorical variables, and supporting tables.

2 BRIEF BACKGROUND AND REVIEW OF RELATED LITERATURE

2.1 An overview of approaches to combining datasets

For several decades, there has been an interest in combining independently-collected datasets and providing a “one stop shop” for a specific set of data needs. This interest only flourished as data-derived insights became more attainable and expected for decision making. The germinal attempts at data matching in the 1960s coincided with the initial spread of accessible computing power capable of handling big datasets (hundreds of thousands of records). Pioneered by governmental organizations (in the U.S., the Social Security Administration and the Internal Revenue Service) – the original “big data” powerhouses – these initial studies aimed to bridge tax, income, and demographic records for the purposes of filling in missing information, finding discrepancies in the reported data, and tracking taxpayers over time to investigate longitudinal trends. Interestingly, as Okner (1974) reported, there was a sentiment of doubt among researchers as to whether obtaining synthetic data through matching was any better than direct surveying, given the substantial amount of human and computational resources enlisted by the former.

Since these fledgling inquiries, the terminology behind the concept of informing one dataset with another has been multifarious, with several contenders coined by different scientific groups. Early into the research, *record linkage*, or *exact matching*, or *exact linking* (of records that describe the same entities; Newcombe et al., 1959) was distinguished from *synthetic (stochastic) linking* or *data synthesizing* (of records that are matched via some approximation; Okner, 1974). Later, while *record linkage* gained ground and blossomed fruitfully over the years (Winkler, 1999), the term *synthetic linking* fell out of fashion in favor of *file concatenation* (e.g., Rubin, 1986), *data fusion* (e.g., Baker et al., 1989) *statistical matching* (e.g., D’Orazio et al., 2006), *ascription* (e.g., van der Putten and Kok, 2010), *data augmentation* (e.g., Hüttenrauch, 2016), and *data triangulation* (e.g., Hand, 2018). For clarity of presentation, this work will adopt *statistical matching* to serve as a “catch-all” term for this process.

At the same time, originating in military applications, the term *data fusion*, “a process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats as well as their significance” (White, 1991, p. 5), got its footing in the fields of signal processing, statistical inference, and machine learning, usually as an abbreviation of the longer term *multisensor data fusion* (Hall and Llinas, 1997; Castanedo, 2013; Khaleghi et al., 2013). The digitization of the modern economy and the boom in consumer information and communication technology (ICT) devices expanded non-military applications of *multisensor data fusion* that now include, for example, lifestyle and medical trackers (Gravina et al., 2017); intelligent transportation systems (ITS) and traffic management (El Faouzi et al., 2011); and autonomous driving (Becker and Simon, 2000).

By using data from several sources that characterize identical entities, *multisensor data fusion* improves the confidence in and reliability of pattern detection, and is more akin to *record linkage*. The first mentions of *data fusion* as a synonym for *statistical matching* can be traced back to works of French and German market researchers in the late 1970s and 1980s (as referenced in Baker et al., 1989; and Rässler, 2002), while *multisensor data fusion* appeared on the radar at least as early as the mid-1980s (Waltz, 1986). Today it is unclear which field has a greater right to claim the definitive terminology, but the detrimental effects of their concurrent existence are apparent. The two originally distinct processes of *record linkage* and *statistical matching* have been conflated via an enveloping term *data fusion*, which hampered diffusion of knowledge within research communities and resulted in the proliferation of endemic studies, which are poorly aware of developments in the other fields. As Khaleghi et al. (2013, p. 28) put it: “Data fusion is a wide ranging subject and many terminologies have been used interchangeably. These terminologies and ad hoc methods in a variety of scientific, engineering, management, and many other publications, shows the fact that the same concept has been studied repeatedly.”

The panoply of terms describing the process of multi-source data integration could be an indirect testament to this argument. The present authors are far from the first to be perturbed by the lack of consistency in the terminology: it has previously been pointed out by Rässler (2002), D’Orazio et al. (2006), and Tsamardinos et al. (2012), for example. To avoid proliferating confusion, this work will refrain from using *data fusion* to describe exclusively the statistical matching process, due to its broader nature and conflated usage. However, readers should be aware that the practice of equating *data fusion* and *statistical matching* is still widespread, especially in European marketing research literature (e.g., Kamakura and Wedel, 1997, van der Putten, 2002; Rässler, 2004; van der Putten and Kok, 2010; Fisseler and Feher, 2010).

Not surprisingly, the problem of fusing data has been also studied within the computer science field, which led to the development of its own distinct methodology. In keeping with the general terminological theme (compare “machine learning”, “supervised learning”, “deep learning”, etc.), the computer-science-based methodology of bridging knowledge sources (i.e., different datasets) to improve task performance (i.e., predictive function accuracy), is fittingly labeled *transfer learning* (Pan and Yang, 2010). It borrows heavily from adaptive behaviors observed in the biological world, in which actors transfer their previously learned skills into new settings. Both “flavors” of *data fusion* (*multisensor data fusion* and *statistical matching*) could be encompassed by *transfer learning* (Zheng, 2015). Stemming from the computer science field, *transfer learning* is innately posed to implement machine learning methods that are capable of handling large amounts of information (i.e., *big data*) computationally efficiently. However, the compartmentalization of the fields is persistent: to our knowledge only three published works (Tsamardinos et al., 2012; Lagani et al., 2016 – explicitly; and Chen et al., 2015 – implicitly) have acknowledged the coexistence of *statistical matching* and *transfer learning* and applicability of the latter to statistical matching problems.

2.2 Transfer learning: definitions, terminology, and key concepts

In this study we aim to implement a statistical matching application by using transfer learning. *Transfer learning* is a machine-learning framework that defines the formal means of knowledge transfer between domains (datasets, or *variable spaces*) using *tasks* – combinations of predictive learning methods and target variables.

Following the transfer learning framework outlined in Pan and Yang (2010), we begin with the concept of a *variable space* \mathcal{X} , which is the set of all available input variables of interest to a study. A specific $n \times p$ input data matrix to be analyzed is denoted X , whose n rows constitute n cases or observations on p variables (following the conventions accepted in data science, this excludes any output variables of interest, i.e., dependent variables or labels, which are co-observed with the p input variables), i.e., n particular elements of a p -dimensional \mathcal{X} or of a p -dimensional subspace of a larger-dimensional \mathcal{X} . The values of n and p could change in the course of the analysis, as cases are filtered out (or, less commonly, added) and variables are added or dropped (see discussions in Sections 3.1 and 4.1). Similarly, we define the variable space \mathcal{Y} as the set of all available output variables of interest to a study. The $n \times q$ output data matrix is denoted Y , whose n rows constitute n cases or observations on q variables, and where the values of n and q could change in the course of the analysis.

A *domain* D consists of an input variable space \mathcal{X} , and probability distribution $P(X)$ over the n observations of a specific X matrix to be analyzed. The simplest case of transfer learning involves two datasets. For a source domain, $D_S = \{\mathcal{X}_S, P(X_S)\}$, let the mapping between it and the output variables of interest to be transferred, \mathcal{Y}_S , be known. Let a target or recipient domain, $D_T = \{\mathcal{X}_T, P(X_T)\}$, contain the other dataset of input variables, which will be used to transfer the information. Then, if there is an intersection between D_S and D_T (a subset of variables that are common to both source and target, with an equal probability distribution $P(X)$ in both domains), i.e., a $D'_S \subseteq D_S$ and $D'_T \subseteq D_T$ such that $D'_S = D'_T$, we can define a function $f(\cdot)$ that, given \mathcal{Y}_S associated with D_S , learns on D'_S and predicts $\hat{\mathcal{Y}}_T$ for D_T , given D'_T . As noted, we will use Y to denote specific realizations of the variable space \mathcal{Y} , i.e. a collection of q specific vectors to be predicted in the case of \hat{Y}_T , or used to train the learning function in the case of Y_S .

A combination of the to-be-transferred variables \mathcal{Y} and learning function $f(\cdot)$ constitutes a learning task, $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. In the present application of knowledge transfer, the learning task is invariant for the source and target domains. This means that the same $f(\cdot)$ is applied to the source and target domains: to calibrate function parameters on the former, and to predict \hat{Y}_T on the latter.

In a given dataset, a specific transfer variable $y \in Y = \{y_1, \dots, y_N\}$ could be either categorical or continuous. Based on this differentiation, the learning function $f(\cdot)$ would respectively involve either a *classification* or *regression* method. Note that according to the naming convention adopted in machine learning, “regression” represents a broad group of methods that go beyond the simple linear or logistic model.

The quality of the prediction of the transferred variables depends on the quality and relevancy of the inputs, X , and the fitness of the learning function $f(\cdot)$. Intuitively it might seem that the more input variables incorporated into $f(\cdot)$, the better the predictions that are generated. However, this intuition collapses in higher dimensions, thanks to the phenomenon commonly known as the *curse of dimensionality* (Bellman, 1961). This concept refers to the exponential inflation of Euclidean hyperspace relative to the unit hypercube as the number of dimensions increases (Keogh and Mueen, 2010). This inflation causes the data to spread out sparsely across the hyperspace and to “drift” towards its edges, all of which leads to a higher variance of the fitted function $f(\cdot)$ and the prevalence of extrapolation over interpolation (Hastie et al., 2009). Possible approaches to abating the curse of dimensionality include variable selection (e.g., stepwise regression) and dimensionality reduction (e.g., principal components analysis).

The fitness of the learning function $f(\cdot)$ can be determined with *cross-validation*, a staple method in the statistical model selection toolbox. Cross-validation resamples the data at random

without replacement (unlike *bootstrapping*, which resamples with replacement) to estimate the generalization error of the model $f(\cdot)$ (Du and Swamy, 2013). In a popular variation of *leave-one-out cross-validation*, namely *k-fold cross-validation*, the dataset is partitioned randomly into k equally-sized subsets. The learning function $f(\cdot)$ is fitted over the combination of $k-1$ subsets (*training data*) while the remaining subset (*test data*) is used to evaluate the performance of the function. The process repeats k times on the same partition, with a different subset being used as the test data each time. The prediction errors of $f(\cdot)$ are averaged across the trials to get an unbiased estimate of the generalization error.

Machine learning practice offers a number of different approaches to formulating and estimating the learning function $f(\cdot)$. All primary algorithms used in this study fall into the category of *supervised learning*, namely the case in which \mathcal{Y}_S , associated with the source domain D_S , exists and is known. In contrast, *unsupervised learning* includes algorithms such as *k-means clustering*, *principal components analysis (PCA)*, and many others that do not require the prior knowledge of \mathcal{Y} (where, for the examples, \mathcal{Y} is respectively cluster membership and principal component “score”) for execution. We implemented a variety of supervised learning algorithms so as to maximize the ability to identify the best ones. Section 9.1.2 in the Appendix briefly describes the high-level mechanics of the algorithms we used.

3 METHODOLOGY

3.1 Transfer learning framework

Our target domain (the source of input variables for predicting \hat{Y}_T) is the NHTS dataset. This domain contains a wide array of disaggregate travel behavior data collected for all 50 states and different land use settings (U.S. Department of Transportation, FHWA, 2009). However, the NHTS sample lacks the attitudinal information that could be instrumental in improving our understanding of travel behavior. It is the purpose of the current study to inform the NHTS dataset with relevant attitudinal data for future use.

Given this objective, a successful source domain (the donor of transfer learning) must have attitudinal variables of interest associated with it and should be compatible with the target domain on several levels: First, the two domains should occupy a comparable spatial and temporal continuum to maximize their congruence on *unobserved* attributes. Second, the two domains should possess a pool of *observed* attributes that are equivalent (or can be made equivalent) across domains in their definition, measurement, and marginal distributions $P(X')$. We refer to this pool as *common variables*, denoted X'_S for the source domain and X'_T for the target domain.

With these requirements in mind, we selected the Multitasking Survey of Northern California Commuters (MSNCC) to be the source domain for this study. The MSNCC was administered by the authors between October 2011 and February 2012 (Neufeld and Mokhtarian, 2012). The working cleaned sample contains more than 2,000 observations of commuting adults (this number varies by variable due to scattered, residual item non-response). Attitudes are represented by general opinions (Appendix, Table 9.1), personality traits, multitasking and time use preferences, and transportation mode perceptions. They are measured on 5- and 3-point ordinal scales generally representing degrees of agreement with statements or attributes. In addition to the observed raw data, a series of factor analyses (e.g., Appendix, Table 9.2) was performed to identify the latent constructs underlying each block of interrelated statements (the technical memos describing these factor analyses are available upon request from the authors). Individuals’

estimated measurements on these latent constructs are expressed by standardized, continuous Bartlett factor scores. The sign of the factor score indicates individual agreement (+) or disagreement (−) with the latent construct while the magnitude of the score shows the extent of it. Overall, the MSNCC provides a flexible source of categorical and continuous attitudinal data available for transfer learning.

The MSNCC and NHTS data were collected within the same reasonably narrow time window, which makes the two domains temporally comparable. Yet spatially, the domains are not adequately comparable because the geographic area of the MSNCC is a small subset of that of the NHTS. So, unless only a geographically equivalent subset of the NHTS is used (which would dramatically reduce the available sample size and could limit the value of the transferred attitudes for subsequent analysis purposes), extrapolating attribute marginal distributions of the Northern California population (demonstrably not representative of the entire country) to the rest of the target domain could have tenuous validity. However, most attitudinally-rich datasets are geographically limited, and therefore for the purposes of learning more about the circumstances under which these methods are useful, it is pertinent to investigate whether information from a local/regional source can be successfully transferred to a national target. Furthermore, it is possible that although marginal distributions of variables differ between the domains, conditional relationships among multiple variables could be more stable (Babbie, 2010). Accordingly, the analysis reported here used the full nationwide scope of the NHTS dataset (a preliminary analysis showed little impact – specifically, little improvement in the effectiveness of the imputed attitudes – when choosing the California subset as the target for transfer learning).

The source and target domains were identified to have 26 common variables between them (Appendix, Table 9.3). Some of the variables have equivalent meaning and measurement in both domains (for instance, age, gender, race, and household size), while some of the variables require additional manipulation to maximize their congruence (for instance, harmonizing family income categories, determining household life cycle for the MSNCC).

The marginal distributions of the common variables are predominantly different across the two datasets, as is shown in Table 9.4 (Appendix) through visual inspection as well as Kolmogorov-Smirnov and chi-squared tests for continuous and discrete variables, respectively. There are several possible causes for this mismatch: the spatial inequality of domains, varying survey sampling rates, different sampling and data collection strategies, survey non-response, and exclusion of observations due to item non-response. In the transfer learning framework, transductive transfer learning offers specific ways to address the inequivalence of source and target domains, in general, and of the marginal distribution of variables, in particular. For example, assuming that \mathcal{Y}_T is partially known, domain adaptation (Daume and Marcu, 2006) factorizes the marginal distributions of each domain into common and specific parts and uses the three resulting distributions for model estimation and prediction. Alternatively, the iterative proportional fitting procedure could be employed for the key variables to find a set of weights that mitigates the distribution mismatch. However, given the limited time and resources allotted for the present project, the authors decided to leave for future research the process of designing, adding, and evaluating a distribution reconciliation procedure. Nonetheless, readers should keep this caveat in mind while assessing the results presented in this paper.

Additionally to the common variables that are directly available in the MSNCC and NHTS datasets, supplemental land use and “environmental” (i.e., socio-economical aggregates of the immediate surroundings) variables were obtained to aid the transfer learning exercise. As explained further in Section 4.1, data from the Decennial census 2010 (U.S. Census Bureau, 2011),

American Community Survey (ACS) 2013 (U.S. Census Bureau, 2014), and Smart Location Database 2013 (U.S. Environmental Protection Agency, 2014) were spatially matched to the residential block group of observations in the source and target domains (for the MSNCC, residential locations were reported by the respondents and were therefore available to us, whereas for the NHTS they are made available to researchers upon special request and under strict confidentiality conditions). This augmentation provides supplemental knowledge that could potentially improve the learning function goodness-of-fit. However, such a dramatic increase in the size of the variable space \mathcal{X} prompts a need to deal with the curse of dimensionality effectively to mitigate computational burden and overfitting.

The last piece of the transfer learning framework that has not been defined yet is the learning function $f(\cdot)$, which completes the learning task \mathcal{T} together with the transferred information \mathcal{Y} (attitudes), $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. In the context of this study, the learning task is invariant for both source and target domains, i.e., the learning function is estimated on the source domain and applied unmodified to the target domain to predict Y_T . Section 9.1.2 (Appendix) offers a few illustrations of how the specification of $f(\cdot)$ can differ based on inputs, outputs, their interrelationships, form, and preconceived knowledge of all of the above. Each learning function has its strengths and weaknesses with respect to the learning task at hand. The intrinsic uncertainty of which function would perform better in the current setting motivated us to develop a learning-function testing framework as a stage in the project methodology. Applied to the source domain, this framework evaluates the performance of each function by averaging the generalization errors after the 10-fold cross-validation procedure. The learner with the lowest average generalization error (mean-squared error and misclassification error for continuous and categorical dependent variables, respectively) for the test sample is considered the most effective in the current application of transfer learning.

Overall, the methodology of the transfer learning framework developed for this study involves complicated data manipulations, multiple parallel function fittings, and conditional decision-making. It can be succinctly characterized by the following sequence (see Figure 1 for a schematic representation).

0. Select and obtain data from the source and target domains: the MSNCC and NHTS (person file).
1. Identify and select common variables across the domains. Reconcile their meaning and units of measurement if necessary
2. Select and obtain supplemental land use data at the block group level from Census 2010 (summary file, all variables), ACS 2013 (summary file, all variables), and Smart Location Database 2013 (all variables). Expand variable space of the Census and ACS datasets threefold by creating interactions of all variables with the reciprocal of total population and area of a block group, respectively, to create relative, size-independent, land use measures.
3. On each expanded Census and ACS dataset, perform data reduction via principal components analysis (PCA) to extract (unrotated) orthogonal projections of the respective variable spaces.
4. Spatially match the residential locations of the observations in the source and target domains with the Smart Location Database, principal components of the Census data, and principal components of the ACS data, intelligently selecting the number of principal components used from each source. This is important for tuning the computational complexity of the subsequent analyses.

5. On the common variable data matrix X'_S (now including land use data) of the source domain, evaluate the fitness of learning functions by running the 10-fold cross-validation procedure and averaging the generalization errors across the folds.
6. Select tasks (corresponding pairs of an attitudinal dependent variable and learning function) with the lowest average generalization error.
7. For each selected task, estimate the learning function (f_S) on the entire source domain (rather than on the 90% at a time which was used at the cross-validation stage) and use that function to predict the value of the attitudinal variable using input variables from the target domain. Merge the target domain with the transferred knowledge (i.e., predicted attitudes).

Additionally to the description above, during phase #5 we evaluate the performance of the learning functions by investigating both categorical and continuous output variables. At the end of phase #7, we complete the transfer learning process by obtaining the target dataset augmented with the knowledge from the source dataset.

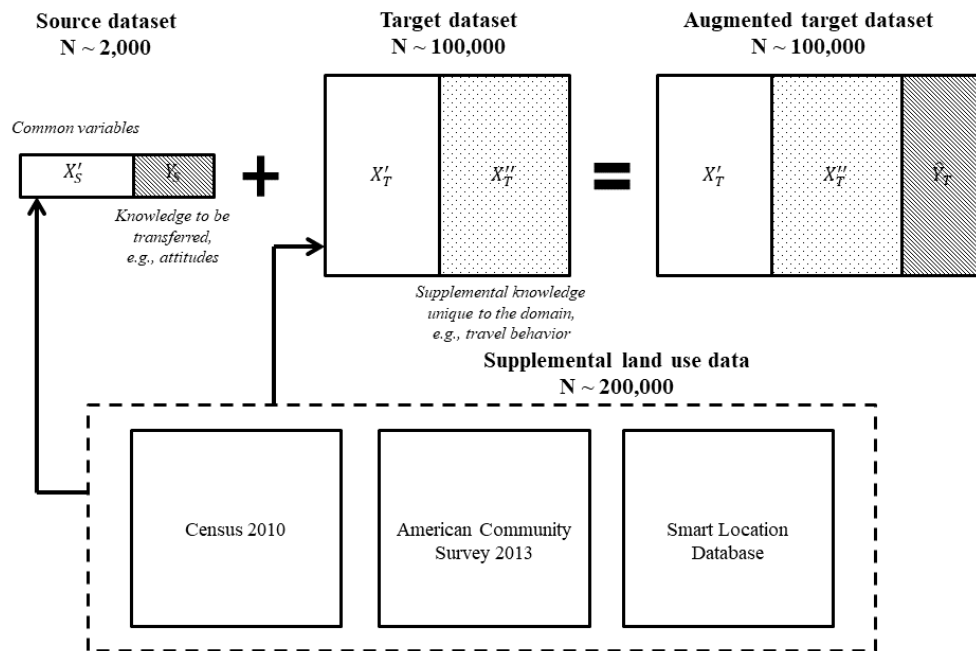


FIGURE 1 Transfer learning framework

Source: Authors' liberal modification of Fig. 1 of van der Putten et al. (2002)

3.2 External validation framework

Cross-validation is a powerful tool that evaluates how well a prediction function performs if the values of \mathcal{Y} are known. We do not have the benefit of knowing \mathcal{Y} when evaluating the transferred knowledge in the context of the target domain. Nevertheless, it is important to assess the added value of that transferred knowledge. To do that, we propose an external validation framework that is summarized in Table 1.

In this framework, external validation is conducted through the analysis of a series of (travel behavior) models implemented on the source and enriched target datasets. To begin, we select a dependent variable that is to be modeled as a function of the rest of the information,

including Y (or \hat{Y} ; note then that in the external validation stages of the analysis, Y and \hat{Y} indicate *explanatory* variables, whereas in the transfer learning stages they were *dependent* variables, or outcomes, of the learning function). Next, we develop models on the source dataset, comparing the outcomes (with respect to quality, fit, and accuracy) across models estimated respectively with Y_S , \hat{Y}_S , and neither of those. Finally, we perform a similar analysis on the target dataset, comparing the outcomes obtained across models with and without the transferred (predicted) variables \hat{Y}_T .

TABLE 1: External validation framework

Model specification	Explanatory variable Y (attitudes)	Specification	Rationale
<i>Source Dataset</i>			
1	Observed	Best	Benchmark
2	None	Same as 1, except w/o Y_S	Assess how much explanatory power the observed Y_S has
3	Predicted	Same as 1	Assess the loss in the goodness of fit of the benchmark model when only the predicted \hat{Y}_S is available
4	Predicted	Best new	Assess how different a model might be from the benchmark, when only the predicted \hat{Y}_S is available and the specification of the model for the true Y_S is unknown
<i>Target Dataset</i>			
5	Predicted	Same as 1 & 3	Assess how well 1 performs within the target dataset and with the predicted \hat{Y}_T
6	Predicted	Best new	Same as for 4
7	None	Same as 6, except w/o \hat{Y}_T	See how much explanatory power the estimated \hat{Y} have

The first model specification of the framework, which is estimated on the source dataset with the observed Y_S , establishes the benchmark of how well the model performs on the observed data. The second model uses the same specification of the previous model except for the exclusion of Y_S from the inputs. Comparing the fits of models 1 and 2 allows evaluating the contribution that Y_S brings to the explanatory power of the benchmark external validation model. The third model has a specification identical to the first one, only instead of the observed Y_S it uses the predicted \hat{Y}_S , that is, the output of the learning function $f(\cdot)$ trained on the common variables X'_S of the source domain. The rationale behind this step is to assess how the unavoidably incorrect prediction \hat{Y}_S influences the quality of the validation model. The fourth model seeks for the best new specification, given the *predicted* \hat{Y}_S , to assess how the model so obtained might differ from the original model (best specification given the *observed* Y_S). Those differences reflect the data's best compensation for the inaccurate prediction of the transferred knowledge (i.e., some variables may increase or decrease in importance, and other variables may enter the model, to pick up some of the

explanatory power lost by replacing the observed Y_S with an imperfect prediction). An assessment of this compensatory mechanism in the source dataset can be useful in evaluating the model’s performance in the target dataset.

These aforementioned four specifications applied to the source dataset can provide valuable initial insight into how the external validation (travel behavior) model performs with the observed and predicted data. However, the core of the framework lies in the application of the model to the target dataset, in which the observed Y_T is unknown. There, we take the changes in model quality detected in the context of the source dataset to be an indication of similar changes in the target dataset. Accordingly, the fifth model, which is estimated on the target dataset using the same specification as for the first model, has the dual role of establishing a benchmark for the *target dataset* and examining the quality change (compared to that of model 1) due to the error in the predicted \hat{Y}_T . With the search for the best new specification, the sixth model attempts to compensate for the error in the predicted \hat{Y}_T to obtain a better model. Finally, the seventh model specification allows an evaluation of the effects (on the quality of model 6) of the exclusion of the transferred knowledge.

Although the NHTS is rich in travel behavior variables, the MSNCC is not. However, vehicle ownership is one such variable common to both samples. Accordingly, in this study, we use a vehicle ownership (VO) model for external validation of the transfer learning procedure. We model VO, represented by a count of household vehicles, as a function of variables such as income, number of workers and drivers, and presence of children. Both datasets are well-equipped to allow for the specification of a reasonable baseline VO model. Moreover, attitudes have also been found to influence VO (see, e.g., Wu et al., 1999; Cao et al., 2007). Thus, VO is a suitable candidate for the external validation.

4 TRANSFER LEARNING RESULTS

4.1 Data preparation

The initial MSNCC dataset X_S consists of 1,118 attributes (p) defined for 2,849 observations (n). The person file of the NHTS supplies the dataset X_T of 113 attributes defined for 308,901 observations. Extracting common variables from the datasets shrinks the variable space to 85 attributes for both domains. Since missingness can provide additional knowledge, item non-response on the common categorical variables is coded into an extra dummy variable (=1 if the value of variable x_i is missing, =0 otherwise).

The source domain includes primarily commuters, while the person file of the NHTS dataset contains entire families. To improve the comparability of the domains, preserve commute mode variables for later analyses, and avoid arbitrary predictions for non-commuting populations, we exclude non-commuters from the target domain. After also filtering out observations with item non-response for continuous variables, the target domain shrinks to 112,026 observations.

Before spatial matching, all involved data sources (the source and target domains, ACS, Census, and Smart Location) need to be brought to a common geographic reference. The MSNCC data provides XY coordinates for residential locations. The NHTS spatial IDs are defined using the 2000 Census block-group boundaries, whereas the three supplemental land use datasets have adopted the block-group boundaries defined for the 2010 Census. We reconcile these two geographies by matching the 2000 Census block-group centroids to the 2010 Census polygons and

assigning the corresponding 2010 block-group IDs to the NHTS observations. In this way, all data sources are defined with respect to the 2010 Census geographies.

The original Census and ACS consolidated datasets contain 3,355 and 3,563 variables, respectively. In addition to the absolute numbers, block-group total population and area are used to create two sets of relative measures: share and density – expanding the variable spaces of each consolidated dataset threefold.

Inflating the common variable space of the transfer learning domains by about 20,000 attributes is computationally burdensome and potentially unjustified with respect to prediction accuracy. Moreover, the source domain, which contains just over 2,000 observations, would face the high-dimensionality problem of $p \gg n$, which requires special techniques to treat. For these reasons, we choose to employ a dimensionality reduction method, namely PCA, to decrease the number of attributes while preserving their supplemental knowledge as much as possible. PCA creates successively orthogonal linear combinations (called *principal components*, *PCs*) of the original (intercorrelated) set of variables, in such a way that the first PCs account for the largest shares of the total variance of the original variables. Census and ACS PCs are extracted separately due to the polynomial growth in runtime with the increase in p . In each case, the total number of extracted PCs is $p - 1$: 9,989 and 10,671 PCs for the Census and ACS datasets, respectively. Essentially all variance of the original set of variables is explained by the first 5,084 and 6,187 PCs for the Census and ACS, respectively. These attribute counts are far lower than in the original data, but still unmanageable. As a cutoff, we choose 75% and 50% of the cumulative variance explained, corresponding to 120 and 76 PCs, for the Census and ACS, respectively.

The Smart Location dataset contains 117 variables, which cover such attributes as demographics, employment, density, diversity, design, transit, and destination accessibility (the full data dictionary is available in Ramsey and Bell (2014)). The relatively small variable space of this supplemental land use dataset allows spatially matching the data without requiring a dimensionality reduction step. After the supplemental land use datasets are spatially matched based on the residential location, the dimensions of the domains are $2,352 \times 379$ and $91,666 \times 380$ for the source and target, respectively.

The final step of data preparation is to augment the common continuous variables in both domains by replacing them with their *natural cubic splines* (degrees of freedom = 3). This process is called *basis expansion*. Using splines is a relatively simple way to allow for non-linearity in relationships in additive models. However, a downside of expanding the basis is the inflation of the continuous variable subspace by the factor of the degrees of freedom. After replacing continuous explanatory variables (including PCs) in the transfer learning domains with their cubic splines, the common variable space of source and target datasets expanded to 968 attributes.

4.2 Best learning function search and selection

For the source domain ($2,352 \times 969$), the search for the best-performing learning function is accomplished by measuring the generalization error and averaging it over a 10-fold cross-validation (CV) routine. We explored two different approaches to predicting attitudes. In the first approach we focused on directly predicting the continuous-valued factor scores that had been previously computed from the source domain (see Table 9.2 in the appendix for examples of factor content). In the second approach, we first predicted the ordinal responses to individual attitudinal

statements (such as those in Table 9.1 in the appendix), and then factor-analyzed those predicted responses.

The search for best learning function is performed separately for the continuous (attitudinal factor scores) and categorical (attitudinal statements) dependent variables, respectively instances of the regression and classification problems described in Section 2.2¹. Recall that in the present discussion, the “dependent variable” refers to the attitudinal variable being predicted (\hat{y}_i), in contrast to the dependent variable (in our case, vehicle ownership) of the model introduced for external validation in Section 3.2, in which the observed (y_i) and predicted (\hat{y}_i) attitudes are *explanatory* variables. This subsection describes phase #6 from the transfer learning methodological sequence defined in Section 3.1.

4.2.1 Regression problem

For the regression problem, Table 2 presents selected generalization errors (the mean squared errors, MSEs) obtained for the continuous dependent variables given the corresponding learning functions (“learners”), i.e., regression tasks (the full results can be found in the appendix, Table 9.8). We tested eleven different learners: random hot deck (RHD), assigning the mean value, forward stepwise linear regression, classification and regression tree (CART), evolutionary regression tree, recursive tree, bagging, random forest, LASSO regression, support vector machine (SVM), and AdaBoost. Among these, LASSO regression (linear regression kernel) shows the best performance by having the minimum generalization error for all \mathcal{Y} variables, except for the *Time-pressure – reality* factor score. On average, the LASSO MSE is 0.894 (which, taking the square root, represents about one standard deviation off the observed value) across the nine dependent variables shown, which is an 11% and 55% improvement over assigning the mean value and RHD, respectively. The RHD learner, as expected, demonstrates the worst performance with an MSE of 1.986 (1.4 standard deviations off the observed value). Mean value assignment, the other learner that is free of conditional assumptions, outperforms only two methods: RHD and forward stepwise linear regression. The latter performed relatively poorly because of the increased prediction variance due to overfitting at the training stage. This is especially interesting since LASSO is also a linear regression method with a variable selection routine. The difference between the two is that LASSO has a built-in cross-validation procedure (done for each fold of the higher-order CV) to prevent overfitting.

The prediction performance of the tasks varies across the dependent variables. *Pro-density*, *pro-transit*, and *pro-active transportation* factor scores are predicted by LASSO regression with an MSE below 0.8. For these variables, the greatest deviation (improvement) from the mean value assignment is achieved: Δ MSE is above 0.20. *Commute benefit* has a slightly worse prediction success with a generalization error of 0.898 (Δ MSE=0.11). The other five variables show substantially less improvement over the mean value assignment method, with Δ MSEs below 0.07. Not surprisingly, the best correlations between the observed and predicted scores are obtained for these four best-predicted variables: 0.571, 0.567, 0.583, and 0.453 for *pro-density*, *pro-transit*, *pro-active transportation*, and *commute benefit*, respectively. While these correlations can be considered moderate rather than high, they compare quite favorably to a typical correlation between an instrumental variable and the endogenous explanatory variable it is replacing in a

¹ Appropriate algorithms, which are capable of handling either or both types of problems (regression and classification), are described in Section 9.1.2 (Appendix).

model. The correlations between the observed and predicted variables for the rest of the factor scores range from 0.233 to 0.343.

It stands to reason that the observed distribution of the generalization error is affected by the knowledge content (i.e., relevance) of the common variables used for prediction. The heavy prevalence of land use inputs in the source domain caused the learning functions to explain relatively well the attitudes associated with built environment attributes. Specifically, commuters who score high on *pro-density*, *pro-transit*, *pro-active transportation*, and *commute benefit* attitudes are more likely to live in denser neighborhoods with more transit, bicycling, and walking options due to residential self-selection, a phenomenon that prominently features in recent literature (Cao et al., 2009).

TABLE 2: Selected cross-validation results for the regression problem

Variable	Best learner	Lowest MSE	Mean assignment MSE	Δ MSE (mean assignment vs. best learner)
<i>Pro-transit</i>	LASSO regression	0.757	0.993	-0.236
<i>Travel is wasted time</i>	LASSO regression	0.985	1.001	-0.016
<i>Pro-technology</i>	LASSO regression	0.951	1.017	-0.066
<i>Commute benefit</i>	LASSO regression	0.898	1.008	-0.110
<i>Time pressure – reality</i>	Evolutionary regression tree	0.994	1.009	-0.015
<i>Time pressure – preference</i>	LASSO regression	0.936	0.994	-0.058
<i>Pro-active transportation</i>	LASSO regression	0.789	1.009	-0.220
<i>Satisfaction</i>	LASSO regression	0.976	1.004	-0.028
<i>Pro-density</i>	LASSO regression	0.748	1.005	-0.257

4.2.2 Classification problem

In a regression problem, trying to predict a continuous variable could produce an unsatisfactorily large generalization error if variables that strongly influence the error’s bias and variance components are unobserved and unaccounted for. In this situation, solving a classification problem, where the goal is to predict to which one of a (usually) small number of predefined categories to which the observation belongs, could mitigate the role of the unobserved inputs and decrease the influence of the error’s components. Additionally, classification problems require certain changes in the algorithm of the learning functions, or the use of completely new learners, which, potentially, might better capture the associations existing in the data. Finally, in using predicted attitudinal items as inputs to a factor analysis, we speculate that random errors associated with predicting each single item could partially counteract each other and result in predicted factors that are more accurate (closer to the “observed” factor scores previously computed from the observed attitudinal items) than those predicted directly as just described. Accordingly, we also performed the prediction of individual items with ordered categorical responses. However, the cross-validation accuracy results obtained in this way (see Section 9.2, Appendix) were apparently not superior to those obtained for the regression problem.

4.2.3 Comparison of the outcomes of the regression and classification problems

To compare the results of continuous and categorical dependent variable prediction, we investigate how the direct prediction of factor scores (regression problem) fares relative to the prediction of the raw statements (classification problem) with subsequent factor analyses of the predicted data. In all factor analyses we use the original method: principal axis factoring with oblimin rotation.

While more details are available in Malokin et al. (2017), here, we summarize them as follows. Comparison of (1) the factor scores obtained from multiplying the common factor score coefficient matrix by the various sets of predictions to (2) the scores originally computed using the observed attitudes shows consistently high correlations for the same constructs identified in the regression problem: *pro-density*, *pro-transit*, and *pro-active transportation*. However, for the most part the highest correlations obtained in this step are still worse (lower) than those obtained from the results of the regression problem. We conclude that at least in this instance, the direct prediction of factor scores is better (and more straightforward) than the two-stage process of predicting individual statements and then factor-analyzing them. Nevertheless, it is still potentially useful to have access to the predicted attitudinal statements, for situations where individual items may be of specific interest, and/or do not load heavily on any factor.

4.3 Transfer learning

For the transfer learning procedure, we apply the learning task to the entire source domain (as opposed to the CV procedure, which uses only a subset of the domain), corresponding to phase #7 of the methodological sequence defined in Section 3.1. The common variable space contains all variables described in Table 9.4 and the land use data described in Section 4.1. The learning task consists of LASSO regression as the learning function and attitudinal dependent variables (sequentially paired with the learner). However, even though the regression problem is shown to be better suited in the setting of the current study (Section 4.2), it is not computationally-burdensome to carry out the classification problem also. Accordingly, using LASSO regression with, respectively, linear regression and MNL kernels for the regression and classification tasks, we estimate the learner for each transferred variable using input variables from the source domain and apply this learner to the target domain. At the end of the transfer learning procedure, the target dataset receives 9 continuous and 39 categorical attitudinal variables defined for 91,362 observations (respondents in the NHTS person file). Table 3 presents selected distribution parameters of the transferred continuous variables, which we briefly discuss here (Tables 9.6 and 9.7 of the appendix contain similar information on the observed and predicted attitudes for the source dataset).

With respect to the continuous attitudes (factor scores), we first note that attitudes per se do not have an “absolute” zero point – they can only be measured relative to some arbitrary benchmark. Accordingly, in the source dataset, the attitudinal factor scores were standardized variables, so that each of their means were zero, and standard deviations equal to one (for the MSNCC dataset, N=2,849). This effectively makes the Northern California sample of the source dataset the benchmark against which the national sample of the target dataset is measured. A mean factor score that is close to zero in the target dataset signifies that on average, the national sample holds an attitude similar to that of Northern California. With that in mind, we can see from Table 3 that based on the nationwide predicted factor scores, respondents are considerably less pro-transit, pro-active transportation, and pro-density than those in the Northern California sample are (while national respondents are comparatively somewhat more satisfied with life and job, and view

the benefits of commuting somewhat more positively). Although this may not be surprising in terms of Northern California stereotypes, it is important to keep in mind that the MSNCC sample is deliberately enriched with non-drive-alone commuters (Neufeld and Mokhtarian, 2012), and as such, in raw form it is not even representative of Northern California.

TABLE 3: Descriptive statistics of the transferred continuous attitudes for the NHTS dataset (N=91,362)

Variable	Number missing ^a	Mean	SD	Median	Min	Max	Skew	Kurtosis
<i>Pro-transit</i>	1	-0.31	0.29	-0.34	-3.21	3.26	1.41	5.33
<i>Travel is wasted time</i>	0	0.01	0.10	0.01	-0.58	3.84	5.03	124.49
<i>Pro-technology</i>	5	0.00	0.25	0.01	-4.02	4.97	0.94	15.45
<i>Commute benefit</i>	0	0.10	0.34	0.10	-3.49	3.04	-0.66	7.96
<i>Time pressure – reality</i>	0	-0.05	0.11	-0.05	-0.77	0.76	0.01	0.02
<i>Time pressure – preference</i>	0	-0.05	0.19	-0.04	-1.86	0.99	-0.21	0.13
<i>Pro-active transportation</i>	5	-0.39	0.29	-0.42	-1.94	4.98	1.77	16.57
<i>Satisfaction</i>	1	0.12	0.21	0.14	-1.65	4.17	-0.08	12.85
<i>Pro-density</i>	3	-0.42	0.45	-0.46	-1.53	3.28	0.70	1.17

^a Predicted values beyond ± 5.0 are coded as missing. Since the learning function is trained on a smaller sample, prediction for some observations in the NHTS sample (which is larger, more heterogeneous, and with a greater chance of extreme input values) could be a result of extrapolation rather than interpolation. The former is known to be more unstable and to produce unrealistic outcomes.

It is also important to note that all the standard deviations of the predicted scores are markedly smaller than one. While in *theory* this could indicate that attitudes in the Northern California sample are considerably more variable in the aggregate than are attitudes nationwide² (which could be another consequence of the choice-based sampling strategy), it is presumably to a much greater extent a reflection of prediction error: given that most sources of variability in attitudes are unmeasured, the learning function will tend to make predictions that do not vary far from the sample mean.

This supposition is strongly supported by a comparison of Tables 9.6 (descriptive statistics for the observed scores in the source sample) and 9.7 (descriptive statistics for the predicted scores – also in the source sample): whereas standard deviations (s.d.s) of the observed scores are all close to one (by design), standard deviations of the predicted scores are never higher than 0.48. Not surprisingly, the three predicted attitudes with the largest standard deviations (where, in this case, a s.d. that is larger – therefore closer to that of the observed attitude – is better, suggesting that the learning function is better at explaining the natural variability of the factor) are pro-density

² Of course, the *range* of attitudes in Northern California will be encompassed by the range for the nation that contains Northern California, but (loosely speaking) if in the national sample extreme attitudes are a *smaller share* of the total, the standard deviation will be smaller. On the other hand, it can be argued that choosing a source sample to have greater variability than the target sample (i.e., choosing it to overrepresent more extreme opinions) is not necessarily a bad thing: a more variable source can draw on more knowledge in predicting a less variable target, than if a more homogeneous source were attempting to predict for a more variable target.

(0.48), pro-transit (0.46), and pro-active transportation (0.45) – the three best-predicted attitudes in this analysis (Section 4.2.1). Commute benefit (0.34) comes in fourth, also in keeping with its predictability.

Comparing the standard deviations of the source dataset’s predicted factor scores (Table 9.7) to those of the target dataset’s predicted scores (Table 3) offers further insight: for most of the nine factors, the s.d.s are nearly equal, whereas for two of the better-predicted factors (pro-transit and pro-active transportation), they shrink by about a third in the target dataset, indicating that the cross-sample transferred factor scores are substantially less variable than the own-sample predicted ones are. Interestingly, the *ranges* of observed and predicted factor scores display a different pattern: the ranges vary between 5 and 7 for the observed continuous attitudes (Table 9.6); they shrink at least twofold (up to ninefold in some cases) for the predicted attitudes in the source dataset (Table 9.7); and they take on more variable amplitudes – larger as well as smaller (2-9) – for the predicted attitudes in the target dataset (Table 3). While the larger ranges in the latter instance suggest a promising departure from the mainly homogeneous predictions seen in the source dataset, it might be an artificial effect created by the learning function struggling with extrapolation in the context of the greater sample heterogeneity of the NHTS.

In sum, these statistics offer a useful reminder of the relativity of attitudinal measures. It is clear that the source sample differs substantially from the target sample in its distribution of the target variables. As discussed in Section 3.1, in future work the source sample (if not initially drawn from the same population as the target sample, which would be preferable) can be weighted to be more representative of the target in terms of the common variables, which should reduce or eliminate these differences. In the meantime, future users of the scores predicted for the national target dataset may wish to re-standardize them. This would at least establish the national mean as the benchmark, although it would not resolve the lower variability in predicted values.

5 EXTERNAL VALIDATION MODEL RESULTS

For our external validation VO models, the dependent variable, number of household vehicles (*HHVEHCNT* in the NHTS data dictionary), is defined in both datasets as the count of motorized vehicles that a household owns. For the pool of potential explanatory variables, we select attributes common to both domains that have been used extensively in the literature and proven to influence VO. (Note that the same variables, albeit a superset of them, are used in the transfer learning exercise.) This pool includes race, gender, age, education, immigrant status, full/part-time work status, occupation, conditions preventing driving/taking public transit, household income, presence of children, number of children, number of drivers, number of workers, interaction between number of workers and number of drivers, distance to work, and land use variables. For greater interpretability, selected land use variables (population, employment, and network densities) are sourced from the Smart Location Database, instead of using the mechanically derived and conceptually abstract Census and ACS principal components described in Section 3.1.

In addition to this list, the pool of explanatory variables includes attitudes, represented by the three latent constructs that showed the lowest generalization error during the cross-validation step: *pro-transit*, *pro-active transportation*, and *pro-density*. We believe that these attitudes should capture effects associated with transportation mode preference and (through residential self-selection) availability, thus influencing the household’s VO.

There are several conventional ways a VO model could be specified, including using linear regression, Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, ordinal response, or multinomial discrete choice (including nested) functional forms. For this study, we choose linear regression due to the interpretability of its standard goodness-of-fit measure. Furthermore, our testing found that other formulations produced very similar results in terms of substantive interpretation of relationships.

Table 4 shows the resulting goodness-of-fit measures, together with coefficient signs and significance levels, for the seven linear regression model specifications that constitute the external validation framework. Specific coefficients are provided in Table 9.10 of the appendix. Model 1, a benchmark, is estimated on the source dataset with observed attitudes. All coefficients have the expected sign; in particular the attitude coefficients are strongly significant and negative, indicating that the more pro-transit, pro-active transportation, and/or pro-density respondents are, the fewer vehicles their households will tend to own. The adjusted R^2 of this specification is 0.45 – an indication of a reasonably well-specified model. Model 2, obtained by setting the attitudinal coefficients to zero, has an adjusted R^2 of 0.42, which signifies a 0.03 (8.0%) “model lift” (improvement in fit) attained by accounting for the three attitudinal constructs in Model 1. Using the same benchmark specification but replacing observed with predicted attitudes (Model 3) fits the data even slightly better ($\Delta R^2 = 0.0015$). In this specification, even with the *Pro-active transportation* coefficient being insignificant (yet still negative), the three predicted attitudes combined are able to explain the variance of the dependent variable better than the originally “observed” attitudes. One possible explanation for this could be the knowledge from the tens of thousands of variables used to predict the attitudes (see Section 3.1). I.e., this multitude of “hidden” variables, which are not present in Model 1, evidently contains a small amount of explanatory power above and beyond the variables that do appear in that model. Model 4 reinforces this empirical result by showing that an optimized (newly-specified “best”) model with predicted attitudes improves over the previous two ($R^2 = 0.46$).

TABLE 4: External validation framework results: linear regression VO model results

Model specification^a	1	2	3	4	5	6	7
<i>Dataset</i>	Source	Source	Source	Source	Target	Target	Target
<i>Attitudes</i>	Observed	N/A	Predicted	Predicted	Predicted	Predicted	N/A
<i>Specification</i>	Best	1 w/o atts.	1	New best	1	New best	6 w/o atts.
<i>Adjusted R-squared</i>	0.4544	0.4209	0.4559	0.4565	0.3849	0.3894	0.3848
Variable^b							
<i>Intercept</i>	+++	+++	+++	+++	+++	+++	+++
<i>Pro-transit</i>	---	0	---	---			0
<i>Pro-active transportation</i>	--	0			---	---	0
<i>Pro-density</i>	---	0	---	---	---	---	0
<i>HH_HISP</i>	0	0	0	0	0	---	--
<i>HH_RACE: Black</i>	--	--	--	---	---	---	---
<i>HH_RACE: Asian</i>	0	0	0	-	0	--	--
<i>HH_RACE: Multi</i>	--	--	---	---		0	0
<i>HH_RACE: Other</i>	-	-	-	--		0	0
<i>HHFAMINC: \$0-25k</i>	---	---	---	---	---	---	---
<i>HHFAMINC: \$25-50k</i>	---	---	---	---	---	---	---
<i>HHFAMINC: \$50-75k</i>	---	---	---	--	---	---	---
<i>HHFAMINC: \$75-100k</i>	---	--	---	0	---	---	---
<i>HHFAMINC: >\$100k</i>	0	0	0	+++	0	0	0
<i>Was born in the US?</i>	++	++	++	+	+++	+++	+++
<i>Condition preventing using public transit</i>	0	0	0	0	0	---	---
<i>EDUC: less than HS degree</i>	0	0	0	0	0	+++	+++
<i>EDUC: HS degree</i>	0	0	0	0	0	+++	+++
<i>EDUC: less than BS/BA degree</i>	0	0	0	0	0	+++	+++
<i>OCCAT: service</i>	0	0	0	0	0	+++	+++
<i>OCCAT: clerical</i>	0	0	0	0	0	+	
<i>OCCAT: manufacture</i>	+++	+++	++	++	+++	+++	+++
<i>OCCAT: professional</i>	0	0	0	0	0	+++	++

(Table 4 is continued on the next page)

TABLE 4: External validation framework results: linear regression VO model results (CONT'D)

Model specification ^a	1	2	3	4	5	6	7
<i>Dataset</i>	Source	Source	Source	Source	Target	Target	Target
<i>Attitudes</i>	Observed	N/A	Predicted	Predicted	Predicted	Predicted	N/A
<i>Specification</i>	Best	1 w/o atts.	1	New best	1	New best	6 w/o atts.
<i>Adjusted R-squared</i>	0.4544	0.4209	0.4559	0.4565	0.3849	0.3894	0.3848
Variable^b							
<i>R_SEX</i>	0	0	0	0	0	--	
<i>SELF_EMP</i>	0	0	0	0	0	+++	+++
<i>Works full time?</i>	0	0	0	0	0	---	---
<i>DRVRCNT</i>	+++	+++	+++	+++	+++	+++	+++
<i>WRKCOUNT</i>	+++	+++	+++	+++	+++	+++	+++
<i>R_AGE</i>	0	0	0	0	0	-	
<i>DISTTOWK</i>	+	+	+	+	+++	+++	+++
<i>Population density</i>	-	---		0	---	---	---
<i>Activity density</i>	0	0	0	0	0	---	---
<i>Jobs per HH</i>	-	--	-	0		0	0
<i>Road network density</i>	0	0	0	0	0	---	---
<i>Jobs within 45 mins</i>	---	---		--	---	0	0
<i>Number of children</i>	-		--	--	---	---	---
<i>Presence of children</i>	0	0	0	0	0	---	--
<i>DRVRCNT*WRKCOUNT interaction</i>	---	---	---	---		0	0

^aNumbering corresponds to Table 1.

^bModel coefficients are represented by their sign (+ for positive, - for negative, blank for insignificant) and significance (one sign for $p < 5\%$, two signs for $p < 1\%$, three signs for $p < 0.1\%$). Zeros indicate the coefficient's absence from the model specification.

The benchmark model specification applied in the target dataset context (Model 5) shows a loss of significance of the *pro-transit* coefficient and changes its sign to a counterintuitive positive one. However, given the national coverage of the NHTS, it is not surprising that a *pro-transit* attitude plays a lesser role outside the relatively small number of transit-oriented areas. This explanation is further corroborated when the subset of urbanized regions with well-developed transit is isolated from the target dataset for model estimation purposes. For example, Model 5 estimated only on the State of New York (results not shown) produces a highly significant and negative *pro-transit* coefficient. Returning to the model estimated on the full target dataset, compared to the first specification, the goodness-of-fit measure is lower ($R^2 = 0.38$), which could be another effect of the greater heterogeneity in the nationwide sample. Model 6 is a product of the search for the best specification in the context of the target dataset. It slightly improves over Model 5 ($\Delta R^2 = 0.0045$), with the attitudinal effects demonstrating the same pattern (i.e., the *pro-transit* coefficient is not statistically significant, and is positive). Finally, Model 7 (when compared to Model 6)

answers the main question of the value of the transferred knowledge (attitudes) for future analyses. The exclusion of attitudes from the model specification results in a drop in the goodness-of-fit measure of 0.0046, or conversely, adding the three attitudinal latent constructs to the VO model specification increases the variance explained by 1.2%.

At first glance, the model lift of 1.2% is rather weak. However, it is useful to consider what variables have been used for the knowledge transfer and external validation processes. By design, the inputs of both the LASSO regression learning function and the VO model are drawn from the partly overlapping subsets of the common variables. With the same socio-economic, travel behavior, and selected land use variables being used in both of these linear-in-parameters functions, the predicted attitudes have little remaining explanatory power to offer beyond that of the other variables in the VO model. When this circularity is removed, i.e., when, for example, only land use variables are used in the transfer learning step and only socio-economic variables (together with the predicted attitudes) are used in the external validation step, the model lift rises to 5.4%, or $\Delta R^2 = 0.0197$, much closer to the difference between Models 1 and 2. Although this obtains a more reassuring performance for the transferred attitudes, it is achieved at the cost of omitting land use explanatory variables – known to be relevant to predicting VO – from the VO model.

Table 5 paints a more comprehensive picture of the competition for explanatory power, as VO model goodness-of-fit measures are cross-tabulated with respect to the groups of variables used in the transfer learning and external validation model specifications. Focusing first on the rows, the table shows that when blocks of variables are entered singly, the socio-economic block delivers the most sizable jump in R^2 (0.36) for the VO model, while attitudes and land use variables by themselves are quite modest in predicting household vehicle ownership ($R^2 \sim 0.03-0.07$). When separately combined with socio-economic variables, the attitude and land use variable blocks each enhance the goodness-of-fit measure by approximately 0.02. An even slighter further increase is demonstrated when all three groups of variables are used together to model VO, indicating the diminishing returns of including correlated explanatory variables.

Turning to the columns, it can be seen that which blocks of variables are used to predict attitudes also influences the goodness of fit of the VO models. When entering the blocks singly, using only land use variables to predict attitudes yields better-fitting VO models than using only socio-economic variables (although the latter may additionally influence the VO model via the attitudes serving as proxies for them) – which, again, is not surprising in view of the land-use-related nature of the attitudes in question. Using both socio-economic and land use variables as predictors for the attitudes further improves the VO models, but only very little beyond what having the land use variables alone delivers. Overall, as data availability increases for both the transfer learning and external validation models, the latter benefits by having a higher goodness-of-fit measure, but the incremental benefits are modest.

Table 5: Goodness-of-fit (R^2) of VO models in the target dataset (NHTS) by VO model specification and LASSO regression learning function inputs

Vehicle ownership is a function of ...	Attitudinal variables ^a are a function of ...			Not applicable
	Socio-economic variables only	Land use variables only	Socio-economic & land use variables	
Land use variables only				0.0337
Attitudinal variables only	0.0466	0.0408	0.0659	
Socio-economic vars. only				0.3572
Attitudinal & land use variables	0.0737	0.0495	0.0693	
Attitudinal & socio-economic variables	0.3655	0.3796	0.3797	
Land use & socio-economic variables				0.3764
Attitudinal & land use & socio-economic variables	0.3808	0.3844	0.3851	

^a The attitudinal variables are *Pro-transit*, *Pro-active transportation*, and *Pro-density*.

6 SUMMARY AND CONCLUSIONS

In this paper, we have developed a transfer learning-based framework for enriching one dataset with knowledge obtained from other related datasets. At the heart of this framework lies the process of identifying the set of variables common across the datasets and training a learning function that performs the knowledge transfer from the source dataset (in which the transferred variables of interest are *observed*) to the target dataset (where the transferred variables of interest are *statistically inferred*). To evaluate the performance of the transferred knowledge, we have also proposed an external validation framework. This framework employs a model, external to the transfer learning process, which is estimated using the transferred knowledge as inputs. Thus, the external validation model provides empirical insight into how valuable the transferred knowledge is to the target dataset.

The transfer learning framework of this paper is broadly applicable to many types of knowledge. The specific aim in this study was to use the framework to enrich the National Household Travel Survey data with attitudes transferred from another dataset. In our application, the *pro-transit*, *pro-active transportation*, and *pro-density* attitudinal factor scores showed the lowest generalization error (using the LASSO learner) and the greatest improvement over the benchmark (assignment of the mean value). The external validation framework was implemented by using a vehicle ownership linear regression model estimated on the source and target datasets

with observed and predicted attitudinal factor scores. The external validation revealed that in the source dataset the observed attitudes account for an 8.0% model lift, and in the target dataset the predicted attitudes account for a 1.2% model lift.

The latter modest result can be explained by the widely overlapping variable space that was used in both the transfer learning and external validation frameworks, which forced the predicted attitudes to compete with their predictors for explanatory power within the same external validation model. This effect was aggravated by the linear-in-parameters nature of the functions used in both frameworks, which created more straightforward substitution and “double counting” patterns among the same variables. If the dependency on the same variable space for both frameworks is broken (e.g., in this instance, when only land use variables are used in the transfer learning step and only socio-economic and travel behavior variables, together with the transferred attitudes, are used in the external validation step), the target dataset shows a model lift of 5.4% when the attitudinal factor scores are included. Excluding land use variables as direct predictors of VO, however, has problems of its own, as discussed in Section 4.

The benefit and cost of strict separation between inputs of the two frameworks is, perhaps, the most important finding of this study. Arguably, the transfer learning process could be viewed as a dimensionality reduction exercise that integrates a vast input variable space into a handful of attributes that gain their definition and meaning from the original knowledge (dependent variables) to be transferred. On the one hand, this suggests that for a subsequent analysis involving the transferred variables, one should avoid the knowledge recycling phenomenon identified here (i.e., including both the transferred variables and the predictors of those variables as explanatory variables in a new model), because the estimated effects on the dependent variable in such a model could distribute unpredictably across the transferred variables and their predictors. On the other hand, taking the opposite approach could *also* lead to biased effect estimates. Consider the present application: if we employ land use-related variables to impute attitudes and therefore exclude land use variables from a model of vehicle ownership, we are *also* distributing estimated effects unpredictably, in that the included attitudes will be partly accounting for the explanatory power of the excluded land use variables (in an ironic reversal of the usual residential self-selection problem, where *included* built environment variables are partly representing the explanatory power of the *excluded* attitudes; Cao et al., 2007).

This dilemma arises because of our hybrid approach to the problem: we are seeking to marry the ad hoc, correlation-based approach of machine learning to the traditional, causally-defensible econometric/statistical approach to model-building – in effect, trying to have it both ways, or put more charitably, trying to wring the best from both approaches. We are by no means the first researchers to use machine learning approaches in the service of causal models (Athey and Imbens, 2015; Dutt and Tsetlin, 2016; Sliva, et al., 2017) even if our twist is distinctive, and some scholars embrace what each approach can bring to the other (Williamson, 2004; Guyon et al., 2008; Rose et al., 2012; Dhar, 2013) – even while a spirited debate between the two “camps” lives on (e.g., Anderson, 2008; Breiman, 2001b and the comments that followed; also, the session titled “Machine learning is from Venus, econometric modeling is from Mars: Two different travel forecasting perspectives” held at the 2017 Annual Meeting of the Transportation Research Board in Washington, DC). Ardent proponents of machine learning would scoff at the desire for a conceptually-driven behavioral model from which reactions to new policies or technologies could be estimated with some confidence; they would view predictions of future vehicle ownership (for example) as just another set of missing data to be imputed by the same transfer learning methods as were used to impute attitudes. Econometric modelers assert the enduring value of understanding

cause and effect (especially in predicting reactions to a change in inputs), and of properly apportioning causality among the conceptually plausible influences on an outcome.

From the standpoint of the latter group, a *theoretical* answer to the dilemma of knowledge recycling is essentially to treat the imputed attitudes as endogenous explanatory variables (EEVs) in the subsequent behavior model – a perspective which highlights the ability to apply well-established econometric methods for dealing with EEVs. For example, an instrumental-variable-oriented approach could be used either in imputing the attitudes in the first place, or in purging them of their endogenous component after imputation (or both). We recommend the latter choice rather than the former, to maximize the amount of knowledge transferred into the imputed variables while retaining maximum flexibility with respect to future uses of those variables (e.g. the imputed attitudes may be EEVs in one context but not in another).

However, just as with the search for instrumental variables under ordinary circumstances, a *practical* answer to the knowledge recycling conundrum may not be easy to find in many instances. What is clear regardless, though, is that the value of the transferred knowledge in the context of the target dataset is determined by the relevancy of available input variables to the informational content of the variables being transferred. Strong associations between the transferred variables and the inputs to the learning functions are essential for more accurate predictions.

In light of these findings, we recommend applying the transfer learning framework to supplemental datasets (e.g., land use, marketing, socio-economic environment data, etc.) that offer reasonable ways to match them to the source and target domains, have strong associations with the transferred knowledge, and serve as valuable informational supplements to future analyses.

This study is far from conclusive. We highlight six important limitations and convenience/necessity shortcuts that warrant further investigation:

1. **Domain adaptation.** Achieving spatial and temporal equivalence between the source and target domains could be a difficult task, given the heterogeneity that exists in data acquisition. Thus, more effort should be dedicated to researching methods of assuring comparability among domains, including reconciling the marginal distributions of the common variables.
2. **Knowledge transfer functions.** The machine-learning field continues developing more advanced and sophisticated methods for more accurate and reliable predictions.
3. **Obtaining a variety of data.** Some potentially fruitful sources of additional common variables include marketing data, credit card transactions information, economic and business aggregates, social media activity, geolocation data, and so on.
4. **Evaluating performance of tasks given the available data.** The three main components of the transfer learning framework are the input (common) variables, the learning functions, and the output (transferred) variables. Options for each of these offer a large number of possible combinations. A more systematic investigation / mapping of generalization errors for various combinations of inputs, dependent variables, and learners is needed.
5. **Evaluating external validation framework.** Similar to the previous point, numerous combinations of components that come into play for the external validation framework need to be further investigated. Effects of knowledge recycling (or its absence) on model lift and different kinds of external validation models are pertinent topics for future research.

In a world where more than 2.3 million terabytes of data are generated every day (VCloudNews, 2015) – and this rate is growing rapidly – the problem of distilling data into humanly-tractable and actionable knowledge is paramount. With the current transfer learning methodology, we arrived at a dimensionality reduction technique of predicting transferred variables as a surrogate for the common variable space. We see this approach as an effective way of treating the $p \gg n$ problem with an advantage of substituting vast variable spaces with meaningful transferred variables, which are suitable for subsequent classical statistical analyses and decision-making processes. Nevertheless, there is much left to learn and improve.

7 ACKNOWLEDGEMENTS

This study was primarily funded by a grant from the Transportation Energy Evolution Modeling division of the Oak Ridge National Laboratory. Additional funding came from the Center for Teaching Old Models New Tricks (TOMNET), a University Transportation Center sponsored by the US Department of Transportation through Grant No. 69A3551747116. The authors would like to thank the Travel Monitoring Division (HPPI-30) of the Federal Highway Administration, and personally Jasmy Methipara, Adella Santos, and Tim Reuscher, for providing access to the confidential part of the National Household Travel Survey.

8 REFERENCES

- Ahfock, Daniel, Saumyadipta Pyne, Sharon X. Lee, and Geoffrey J. McLachlan (2016) Partial identification in the statistical matching problem. *Computational Statistics & Data Analysis* **104**, 79-90.
- Anderson, Chris (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired* **16(7)** (June 23). Available at <https://www.wired.com/2008/06/pb-theory/>, accessed June 30, 2017.
- Andridge, Rebecca R. and Roderick J.A. Little (2010) A review of hot deck imputation for survey non-response. *International Statistical Review* **78(1)**, 40-64.
- Antonelli, Joseph, Corwin Zigler, and Francesca Dominici (2017) Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics* **18(3)**, 553-568.
- Athey, Susan and Guido W. Imbens (2015) Machine learning methods for estimating heterogeneous causal effects. Unpublished manuscript, arXiv:1504.01132v1 [stat.ML] 5 Apr 2015, available at https://www.researchgate.net/profile/Guido_Imbens/publication/274644919_Machine_Learning_Methods_for_Estimating_Heterogeneous_Causal_Effects/links/553c02250cf2c415bb0b1720.pdf, accessed June 30, 2017.
- Babbie, Earl (2010) *The Practice of Social Research*, 12th edition. Belmont, CA: Wadsworth Publishing Company.

A. Malokin, P.L. Mokhtarian, and G. Circella

- Baker, Ken, Paul Harris and John O'Brien (1989) Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society* **31(2)**, 153-212.
- Becker, Jan C. and Andreas Simon (2000) Sensor and navigation data fusion for an autonomous vehicle. In *Proceedings of the IEEE Intelligent Vehicles Symposium 2000*, 156-161. IEEE.
- Bellman, Richard E. (1961) *Adaptive Control Processes*. Princeton University Press.
- Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks.
- Breiman, Leo (1996) Bagging predictors. *Machine Learning* **24**, 123-140.
- Breiman, Leo (2001a) Random forests. *Machine Learning* **45**, 5-32.
- Breiman, Leo (2001b) Statistical modeling: Two cultures. *Statistical Science* **16(3)**, 199–231.
- Cao, Xinyu, Patricia L. Mokhtarian, and Susan L. Handy (2007) Cross-sectional and quasi-panel explorations of the connection between the built environment and auto ownership. *Environment and Planning A* **39(4)**, 830-847.
- Cao, Xinyu, Patricia L. Mokhtarian, and Susan L. Handy (2009) Examining the impacts of residential self-selection on travel behaviour: a focus on empirical findings. *Transport Reviews* **29(3)**, 359-395.
- Castanedo, Federico (2013) A review of data fusion techniques. *The Scientific World Journal* **2013**, 1-19.
- Chen, Aiyou, Art B. Owen, and Minghui Shi (2015) Data enriched linear regression. *Electronic Journal of Statistics* **9(1)**, 1078-1112.
- Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C* **68**, 285-299.
- Chen, Tianqi and Carlos Guestrin (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. ACM.
- Dhar, Vasant (2013) Data science and prediction. *Communications of the ACM* **56(12)**, 64-73.
- D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu (2006) *Statistical Matching: Theory and Practice*. Chichester, England: John Wiley & Sons.
- Du, Ke-Lin and M. N. S. Swamy (2013) *Neural Networks and Statistical Learning*. London: Springer.

- Dutt, Pushan and Ilia Tsetlin (2016) Income distribution and economic development: Insights from machine learning. INSEAD Working Paper No. 2016/62/EPS/DSC. Available at <https://ssrn.com/abstract=2701744>, accessed June 30, 2017.
- El Faouzi, Nour-Eddin, Henry Leung, and Ajeesh Kurian (2011) Data fusion in intelligent transportation systems: Progress and challenges – A survey. *Information Fusion* **12(1)**, 4-10.
- Fisseler, Jens and Imre Fehér (2009) Data fusion with probabilistic conditional logic. *Logic Journal of IGPL* **18(4)**, 488-507.
- Fosdick, Bailey K., Maria DeYoreo, and Jerome P. Reiter (2016) Categorical data fusion using auxiliary information. *The Annals of Applied Statistics* **10(4)**, 1907-1929.
- Freund, Yoav and Robert E. Schapire (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55(1)**, 119-139.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016) *Deep Learning*. Cambridge: MIT press.
- Gravina, Raffaele, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. (2017) Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion* **35**, 68-80.
- Grubinger, Thomas, Achim Zeileis, and Karl-Peter Pfeiffer (2014) evtree: Evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software* **61(1)**, 1-29.
- Guyon, Isabelle, Dominik Janzing, and Bernhard Scholkopf (2008) Causality: Objectives and assessment. *JMLR Workshop and Conference Proceedings* **6**, 1-38.
- Hall, Dave L. and James Llinas (1997) Introduction to multisensor data fusion. *Proceedings of IEEE* **85(1)**, 6-23.
- Hand, David J. (2018) Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society, Series A* **181(3)**, 1-24.
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009) *The Elements of Statistical Learning; Data Mining, Inference, and Prediction. Second Edition*. New York: Springer.
- Hornik, Kurt (1991) Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4(2)**, 251-257.
- Horton, Torsten, Kurt Hornik, and Achim Zeileis (2006) Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics* **15(3)**, 651-674.

A. Malokin, P.L. Mokhtarian, and G. Circella

- Hüttenrauch, Bettina (2016) *Targeting Using Augmented Data in Database Marketing: Decision Factors for Evaluating External Sources*. Wiesbaden, Germany: Springer.
- Kamakura, Wagner A. and Michel Wedel (1997) Statistical data fusion for cross-tabulation. *Journal of Marketing Research* **34(4)**, 485-498.
- Keogh, Eamonn and Abdullah Mueen (2010) Curse of dimensionality. In *Encyclopedia of Machine Learning*, ed. Claude Sammut and Geoffrey I. Webb, 257-258. Boston: Springer.
- Khaleghi, Bahador, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi (2013) Multisensor data fusion: A review of the state-of-the-art. *Information Fusion* **14(1)**, 28-44.
- Lagani, Vincenzo, Sofia Triantafyllou, Gordon Ball, Jesper Tegner, and Ioannis Tsamardinos (2016) Probabilistic computational causal discovery for systems biology. In *Uncertainty in Biology*, ed. Liesbet Geris and David Gomez-Cabrero, 33-73. Cham, Switzerland: Springer.
- Malokin, Aliaksandr, Patricia L. Mokhtarian, and Giovanni Circella (2017) *An Investigation of Methods for Imputing Attitudes from One Sample to Another*. School of Civil and Environmental Engineering, Georgia Institute of Technology, Research Report. Available at <http://hdl.handle.net/1853/58418>, accessed July 12, 2017.
- Neufeld, Amanda J. and Patricia L. Mokhtarian (2012) *A Survey of Multitasking by Northern California Commuters: Description of the Data Collection Process*. Institute of Transportation Studies, Research Report UCD-ITS-RR-12-32. Available at http://www.its.ucdavis.edu/?page_id=10063&pub_id=1802, accessed April 1, 2017.
- Newcombe, Howard B., James M. Kennedy, S. J. Axford and Allison P. James (1959) Automatic linkage of vital records. *Science* **130(3381)**, 954-959.
- Okner, Benjamin (1974) Data matching and merging: an overview. *Annals of Economic and Social Measurement* **3(2)**, 347-352.
- Pan, Sinno Jialin and Qiang Yang (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22(10)**, 1345-1359.
- Rässler, Susanne (2002) *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer Science & Business Media.
- Rässler, Susanne (2004) Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics* **33(1-2)**, 153-171.
- Rose, Sherri, Richard J. C. M. Starmans, and Mark J. van der Laan (2012) Targeted learning for causality and statistical analysis in medical research. UC Berkeley Division of Biostatistics Working Paper 297, available at <http://biostats.bepress.com/ucbbiostat/paper297>, accessed June 30, 2017.

A. Malokin, P.L. Mokhtarian, and G. Circella

Rubin, Donald B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* **4(1)**, 87-94.

Schifeling, Tracy, Jerome P. Reiter, and Maria DeYoreo (2016) Data Fusion for Correcting Measurement Errors. arXiv preprint arXiv:1610.00147. Available at <https://arxiv.org/pdf/1610.00147.pdf>, accessed on April 9, 2018.

Sliva, A., S. N. Reilly, D. Blumstein, S. Hookway, and J. Chamberlain (2017) Modeling causal relationships in sociocultural systems using ensemble methods. In S. Schatz and M. Hoffman, eds., *Advances in Cross-Cultural Decision Making*. Series on Advances in Intelligent Systems and Computing, Vol 480. Springer.

Tsamardinos, Ioannis, Sofia Triantafillou, and Vincenzo Lagani (2012) Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research* **13(Apr)**, 1097-1157.

U.S. Census Bureau (2011) *Decennial Census 2010*. Available at <https://www2.census.gov/>, accessed on April 1, 2017.

U.S. Census Bureau (2014) *American Community Survey 2013, 5-year estimates*. Available at <https://www2.census.gov/>, accessed on April 1, 2017.

U.S. Department of Transportation, Federal Highway Administration (2009) *National Household Travel Survey*. Available at <http://nhts.ornl.gov>, accessed on April 1, 2017.

U.S. Environmental Protection Agency (2014) *Smart Location Database, 2013*. Available at <https://www.epa.gov/smartgrowth/smart-location-mapping>, accessed on April 1, 2017.

van Buuren, Stef (2012) *Flexible Imputation of Missing Data*. Boca Raton, Florida: CRC press.

van der Putten, Peter, Joost N. Kok, and Amar Gupta (2002) Data fusion through statistical matching. MIT Sloan Working Paper No. 4342-02. Available at <http://liacs.leidenuniv.nl/~puttenpwhvander/library/2002fusionsloan.pdf>, accessed June 10, 2017.

van der Putten, Peter and Joost N. Kok (2010) Using data fusion to enrich customer databases with survey data for database marketing. In *Marketing Intelligent Systems Using Soft Computing*, ed. Jorge Casillas and Francisco J. Martínez-López, 113-130. Berlin: Springer.

VCloudNews. Every day big data statistics – 2.5 quintillion bytes of data created daily. Available at <http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/>, accessed on May 31, 2017.

Waltz, Edward (1986) Data fusion for C3I: A tutorial. In *Command, Control, Communications Intelligence (C3I) Handbook*, prepared by the editors of Defense Electronics, 217-226. Palo Alto, CA: EW Communications.

A. Malokin, P.L. Mokhtarian, and G. Circella

White, Franklin E. (1991) *Data Fusion Lexicon*. San Diego, U.S.: Joint Directors of Laboratories, Technical Panel for C3, Data Fusion Sub-Panel, Naval Ocean Systems Center. Available at <http://www.dtic.mil/dtic/tr/fulltext/u2/a529661.pdf>, accessed on May 3, 2018.

Williamson, Jon (2004) A dynamic interaction between machine learning and the philosophy of science. *Minds and Machines* **14**, 539-549.

Winkler, William E. (1999) *The State of Record Linkage and Current Research Problems*. Technical Report, Statistical Research Division, U.S. Census Bureau.

Wu, Ge, Toshiyuki Yamamoto, and Ryuichi Kitamura (1999) Vehicle ownership model that incorporates the causal structure underlying attitudes toward vehicle ownership. *Transportation Research Record* **1676**, 61-67.

Zheng, Yu (2015) Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data* **1(1)**, 16-34.

9 APPENDIX

9.1 Additional background and review of related literature

9.1.1 Statistical matching

The objective of this work could be achieved by *statistical matching* rather than *record linkage* and *multisensor data fusion*, so it is of interest to review particular methods that are implemented for statistical matching processes in the literature. To this end, D’Orazio et al. (2006) distinguish two approaches to data integration: macro and micro. In the macro approach, only the joint distribution of variables of interest (observed in one dataset and inferred in another) is transferred across the data sources. In the micro approach, the goal is to ascribe individual values to the variables of interest by inferring them via some approximation. The micro approach is most broadly used (including by the present study) due to the wider range of applications available with the resultant data. In a general case that involves two datasets, each of which consists of unique and common variables, the goal of the micro approach to statistical matching is to ascribe missing values of the unique variables in a combined stacked dataset using their partially observed relationship with the common variables.

However, there could be infinite ways of ascribing missing values to the unique variables of both datasets because they have been never observed jointly, thus, an identifiability problem exists. To overcome this problem, it is customary (and often implicit) to assume conditional independence between ascription targets, that is, given the common variables, the joint distributions of the unique variables are independent (D’Orazio et al., 2006). Viability of this assumption is extremely context-dependent; and in many real-world scenarios it is not guaranteed. In practice, there are two ways to relax the *conditional independence assumption* (CIA): (1) by expanding the common variables set, and (2) by collecting additional (small-batch) *auxiliary* data that observes all sets of unique and a set of common variables jointly (Fosdick et al., 2016; Schiefeling et al., 2016).

Typically, micro statistical matching consists of several steps. First, a model is trained on the fully observed data (“*donor*” or “*source*”), using unique and common portions as dependent and independent variables, respectively. Next, the trained model is applied to other datasets (“*recipients*” or “*targets*”), for which only the common variables are known. Finally, the predictions of this model are ascribed to the recipient datasets, synthetically supplying them with the previously unobserved variables. The result of all combinations between donors and recipients could be stacked to produce a complete synthetic dataset with unique variables defined across all observations. Different methods of micro statistical matching modify this algorithm to accommodate certain contexts and to improve validity.

Multiple factors are considered to classify methods of micro statistical matching: (1) type of function used (parametric, non-parametric, mixed, or Bayesian); (2) presence or absence of auxiliary data; (3) matching constraints used (e.g., “can an observed value be ascribed to multiple observations?” or “can a recipient observation with multiple missing values have different donor observations?”; Rässler, 2002); (4) levels of validity targeted (preserving individual values, joint distributions, correlation structures, and marginal distributions; Rässler, 2002); and (5) presence or absence of multiple outcome aggregation (in the case of multiple imputation).

Specifically, parametric methods rely on conditional mean matching (regression and log-linear for continuous and discrete ascription, respectively) to capture the observed relationship

between unique (dependent) and common (independent) variables. Stochastic noise could be used to create additional variability in the ascribed values. Non-parametric methods do not estimate parameters of the matching function explicitly; rather, they learn the marginal and joint distributions of the variables in the training data implicitly. For example, the *random hot deck (RHD)* method ascribes values in the recipient dataset with random draws from the values of unique variables observed in the donor dataset (Andridge and Little, 2010). Variations of RHD include methods such as ranked hot deck and distance hot deck (D’Orazio et al., 2006).

While RHD uses the whole training dataset to predict values of the unique variables in the recipient dataset, the prediction accuracy could be improved if the consideration pool were limited only to similar observations. The *k-nearest neighbors (kNN)* method is a non-parametric, locally-approximated algorithm that implements this “informational” homogeneity. It works well with both continuous and discrete dependent variables. In the method, observations from both datasets are mapped in the hyperspace defined by the common variables. Then, for each observation from the recipient dataset, the k closest neighbors from the donor dataset are “polled”, and the distribution of their “votes” (namely, the most-commonly-appearing class for predicting a categorical dependent variable, and an averaged value for predicting a continuous one) defines the ascribed value. The proximity of neighbors is determined by a Euclidean, weighted, or other distance function. The value of k has an inverse relation with the complexity of the model and homogeneity of neighborhoods: larger ks correspond to fewer, more heterogeneous neighborhoods.

Hot deck and *kNN* methods and their close relatives were among the first to be implemented in the early history of statistical matching due to their simplicity and low computational complexity. While they are still very popular today because of requiring fewer assumptions about the data, more elaborate non-parametric methods are being proposed: for example, the Gibbs sampler approach (Ahfock et al., 2016) performs a search in a complex multi-dimensional restricted set to fill in values.

Mixed methods employ two-stage processes that include both parametric and non-parametric methods, which first approximate some value and then ascribe an observed value based on this approximation (D’Orazio et al., 2006). Finally, alternatively to the frequentist approach of modeling, *Bayesian* methods incorporate substantial randomness into the parametric approach by allowing parameter uncertainty and outcome noise to be determined by posterior probabilities observed from the data (Rässler, 2002; van Buuren, 2012). Some interesting recent examples of the approach include the Guided Bayesian Adjustment for Confounding framework that incorporate dimension reduction and treatment for heterogeneity (Antonelli et al., 2017).

The introduction of parameter uncertainty in Bayesian methods aligns rather well with the *multiple imputation* framework. *Imputation*, or treatment of statistical matching as a nonresponse phenomenon (Rässler, 2002), is an alternative perspective on the problem. As such, donor and recipient datasets could be concatenated by (column-wise) aligning common variables, and assigning missing values to the unique variables of observations from the recipient datasets. Afterward, a desired imputation method could be applied to “recover” the missing values. Multiple imputation for statistical matching was first proposed by Rubin (1986) as a way to overcome the CIA assumption and preserve the inherent uncertainty about true values of the unobserved unique variables in a recipient dataset. By using random parameter distribution draws, multiple imputation with chained equations creates several (m) datasets that show variability in the filled-in missing values but retains respective marginal and joint distributions across the concatenated datasets. An analyst, then, needs to average distributional parameters (e.g., mean, standard deviation, regression

coefficients, etc.) across the datasets to arrive at unbiased (under missing completely at random and missing at random conditions) estimators. One apparent drawback of the method is the added analysis complexity of carrying along all imputed datasets and finding an average of m analyses. Attempting to overcome this drawback by the tempting shortcut of averaging imputed values across datasets and then proceeding with the single averaged dataset is not recommended because “imputation is not prediction” (van Buuren, 2012, p. 45). That is, faithful recreation of missing values is not the goal of imputation. In any case, several studies (e.g., Rässler, 2004) have shown promising results of multiple imputation when compared to other statistical matching methods.

Another popular imputation method is the Expectation-Maximization (EM) algorithm, which consists of two steps: expectation, which calculates the log-likelihood given some imputed values, and maximization, which maximizes the log-likelihood by adjusting the imputed values. EM is considered the best off-the-shelf method aside from multiple imputation, and has been extensively used in statistical matching applications (e.g., Kamakura and Wedel, 1997). However, Rässler (2002) showed that in the context of file concatenation, EM rarely converged to the global maximum. Further, it did not produce unique solutions, as the imputed values and model parameters were highly dependent on the algorithm’s starting conditions, unless auxiliary data were present.

Finally, machine learning methods have been gradually making inroads into statistical matching, usually by embedding into existing frameworks. D’Orazio (2011) implemented tree-based machine learning algorithms – classification and regression tree (CART; Breiman et al., 1984), random forest (RF; Breiman, 2001a), and recursive tree (Horton et al., 2006) – in a statistical matching application to find them capturing non-linearity in a synthetic dataset well. CART and RF have also been included in the popular R library for multiple imputation, MICE, where they can be used as univariate imputation methods (van Buuren, 2012). However, to date, machine learning methods are still largely absent from statistical matching applications (Putten and Kok, 2010). Transfer learning, by contrast, utilizes primarily machine learning methods, which we discuss in more detail in the next section.

No matter what framework or method for statistical matching is chosen, we should better understand the risks to validity, error propagation, and quality of inference in the fused data (Hand, 2018).

9.1.2 Overview of machine-learning methods implemented in this study

Technically not a machine-learning algorithm per se, *linear regression* provides a simple yet powerful and interpretable model of how inputs X affect outputs Y . The method may surpass more complicated, non-linear models in cases of small numbers of observations, sparse data, and low signal-to-noise ratio (Hastie et al., 2009). However, two particular challenges often arise in situations where “wide” datasets (having many variables or columns) are considered. First, by increasing the number of parameters (variable coefficients) in $f(\cdot)$, we overfit the model to the training dataset and sacrifice its transferability (the so-called *variance-bias tradeoff*, in which an overfit model reduces the bias involved in using a simpler model to reflect a more complex reality, but increases the variance between predicted and actual values when transferring the model to a new context). Second, the interpretability of such a model suffers because of the clutter created by copious parameters with associated marginal effects.

To overcome these challenges, *subset selection* and *shrinkage* methods are used in practice. Subset selection is a discrete approach in which variables are selected based on their performance

in the model. The *best-subset selection* method searches the entire combinatorial space to pick the best performing specification. However, under current computational constraints, *best-subset selection* quickly becomes infeasible as the number of input variables increases. *Forward- and backward-stepwise selection* methods test each variable and at each stage include (exclude) the variable that most improves (least reduces) the fit until convergence at the given threshold is reached. Shrinkage methods offer a continuous solution to the problem of overspecification. Instead of the discrete choice of dropping or retaining a variable coefficient, they introduce an additive penalty term into the model, e.g., $\lambda \sum_{j=1}^p \beta_j^2$ in the case of *ridge regression* (where the β_j are parameters or variable coefficients, p is the number of parameters, and λ is the shrinkage operator), which is estimated simultaneously with the model and which prevents the large coefficient magnitudes that are common in the presence of multicollinearity (i.e., it shrinks the coefficient magnitudes toward zero). The penalty term for *Least Absolute Shrinkage and Selection Operator (LASSO) regression*, $\lambda \sum_{j=1}^p |\beta_j|$, is similar to the one for ridge regression and also shrinks coefficient magnitudes; however, its non-linear nature allows *LASSO* to take the best of both discrete and continuous methods: by allowing some coefficients to shrink to zero (unlike ridge regression), it can effectively perform subset selection as well as shrinkage, which is essential for high-dimensionality problems (Hastie et al., 2009). Zou and Hastie (2005) proposed a convex combination of ridge and LASSO regression – the *elastic net*. Its penalty, $\lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|)$, is a generalization, which yields ridge or LASSO when $\alpha = 1$ or $\alpha = 0$, respectively.

If kNN attempts to create complex non-parametric boundaries between observations in hyperspace, multiple machine-learning methods, such as *logistic regression* and *linear (quadratic) discriminant analysis*, tackle this problem by imposing a functional relationship between X and Y . However, this separation could be one of infinitely many and may not be optimal, thus leading to misclassification of the new data points that map close to the class boundaries. *Support Vector Machines (SVMs)* are a method for classification and regression that solves this problem by finding optimal separating hyperplanes, that is, boundaries with the widest margins between classes. SVMs can handle inseparable problems and minimize the overlap of classes. Additionally, usage of kernel functions (e.g., n th-degree polynomials, radials, and neural networks) allows for creating non-linear boundaries in the original hyperspace.

Decision trees for classification and regression problems are a staple in the machine-learning field. The name refers to the way a model of this type is presented: it is a type of directed acyclic graph (DAG; Pelikan et al., 2001) with several nodes, each denoting a (usually binary) split based on the value of an explanatory variable. A tree starts with the first single split. Its child nodes may be iteratively split again, with the variables and split-points potentially differing by node. Tracing a branch down to a terminal node represents a set of conditions defining a group of observations that are predicted to have a certain value of the dependent variable, \hat{y} . Another way to illustrate the model is to partition the variable space into a set of volumes, each of which would correspond to a single (averaged) value of Y . These partitions are bounded by the split-planes that are equivalent to the binary split-nodes in the tree representation. The task of the model estimation algorithm, then, is to find independent variables X and their split-points that result in a more accurate prediction of Y .

Among the “tree-growing” methods used in practice, the *classification and regression tree (CART)*; Breiman et al., 1984) and *C5.0* (including its earlier versions *ID3* and *C4.5*; Quinlan, 2014) are considered to be the “classic” algorithms. Varying in small details, they both use a greedy heuristic to do iterative splitting, and implement a tree-pruning cost function to avoid

overspecification. Other variations include *recursive binary partitioning* (Horton et al., 2006), which employs non-parametric modeling, and globally-optimal trees obtained via *evolutionary learning* (Grubinger et al., 2014).

As summarized by Hastie et al. (2009), decision trees possess several desirable qualities: Decision trees can handle mixtures of continuous, ordered, and categorical independent variables naturally. They are insensitive to irrelevant inputs (e.g., case IDs, comment fields, etc.), and can treat item non-response as an explanatory variable. Scaling and other monotone transformations of the data do not affect the performance of tree models, nor does the presence of egregious outliers. Decision trees are computationally scalable and adequately interpretable. However, there are several issues common to the method:

- Trees are inherently unstable, noisy, very sensitive to the training data, and display high variance of the prediction.
- Using categorical independent variables with many levels could lead to severe overfitting because the number of possible partitions grows exponentially with the number of variable levels.
- In regression problems, lack of smoothness could degrade performance of the learning function, if it is assumed to be smooth.
- Multiway (rather than the more common binary) splits are possible but they rapidly fragment and deplete data for the next level down.
- Additive structures are increasingly difficult to capture as the number of additive effects grows.

Ensemble methods in machine-learning try to alleviate one of the most severe shortcomings of the decision trees – overfitting to the training data. One approach involves a resampling method of bootstrap aggregation, or *bagging*. Bagging averages the prediction \hat{Y} over multiple bootstrap samples, thereby reducing the variance of the fitted values and improving their accuracy at the price of losing the interpretable model structure. Moreover, bagging is beneficial only in cases when the unbagged model specification is not optimal, e.g., when parameters that cause overfitting are non-zero (Breiman, 1996). *Random forests* (Breiman, 2001a) offer a substantial improvement over bagging procedures, by building multiple “de-correlated” decision trees. By randomly picking a subset of independent variables to grow each tree in the ensemble (hence, the name), the algorithm decreases the correlation between trees and, therefore, the variance of the averaged fitted values. The random forests approach is a quite popular, ready-to-use machine-learning method that requires very little tuning and produces robust results if the ratio of relevant to all variables in the dataset is not small.

Boosting, also an ensemble method, takes a different approach by assembling the “voting committee” from separate models. *Adaptive boosting*, or *AdaBoost* (Freund and Schapire, 1997), is the most widely used boosting algorithm. The algorithm relies on producing successive “weak” learners (decision trees), whose fitted outcomes could be just slightly better than a random guess. After each fitting iteration, observations in the training dataset are reweighted based on how well the model has performed: weights for observations that were predicted correctly by the previous learner are decreased, and weights for observations with unsuccessful predictions are increased. This procedure encourages each ensuing model to pay more attention to the difficult-to-predict cases. After models are estimated, their prediction results are averaged, while weighting the *results* from the more accurate models more heavily (in contrast to the *observation* weights mentioned above) to increase the influence of the better learners.

Extreme gradient boosting (XGB; Chen and Guestrin, 2016) is another popular ensemble method that incorporates a highly-scalable, sparsity-aware gradient tree algorithm. Its advantages include computational efficiencies through imbedded parallelization, caching, and approximation search; and an ability to handle sparse data natively; all of which bolstered usage of XGB in applications in digital advertising, insurance, and particle physics.

It is widely agreed in the machine-learning community that boosting is one of the best general machine-learning treatments available off-the-shelf. In many cases, boosting, which has decision trees serving as the basic algorithm, helps to overcome the main drawback of decision trees – inaccuracy due to their large variance – at the price of lower computational speed and some reduction in interpretability.

Another large group of methods includes *artificial neural networks (ANNs)*, which are known for being universal function approximators (Hornik, 1991). Although ANNs consist of very simple elementary units, interconnected architecture with fully, partially, and recursively connected “hidden” layers could render it very complex (the infamous “black-box”). Recent advancements in network design and hardware have allowed even larger and more complex models (with millions of neurons) in the pursuit of prediction accuracy (Goodfellow et al., 2016). As such, rectified linear unit and other novel activation functions were introduced to prevent gradient decay while training deep networks. At the same time, graphics processing units (GPUs) provided massive parallelization yielding multiple fold training time decrease compared to conventional central processing units (CPUs). This enabled the sharp rise of ANN applications in computer vision, speech recognition, natural language processing, and artificial intelligence. This list, however, shows that ANNs are extremely successful in domains with homogenous input such as images, sounds, and spoken and written language.

Apparently, there is no silver bullet when it comes to selecting the best algorithm for the task: its performance is contingent on various factors, such as nature, content, and representation of the explanatory and dependent variables, and the strength of associations between them. In this study, we test all aforementioned machine-learning algorithms. For comparative performance tables, refer to Tables 9.8 and 9.9.

9.2 Transfer learning results for the classification problem

Selected cross-validation results of the classification tasks are presented in Table 9.1 (the full results can be found in Table 9.9). The CV-aided search for the best learner is similar to the one presented for the regression problem, with the exception of a few details. Specifically, instead of MSE, the misclassification error (MCE) is used as the generalization error. K-nearest neighbors (kNN) and forward stepwise multinomial logit (MNL) are added to the pool of the learning functions. Assigning the mean value and CART learners are replaced by assigning the median value and C5.0, respectively. For the forward stepwise linear regression learner, predicted values are rounded to the nearest eligible integer that represents a response category. Finally, artificial neural networks (ANNs) and extreme gradient boosting (XGB) are added to the pool of the tested learning functions.

Across 15 learning functions, both kNN and XGB achieve the minimum MCE for 14 out of 39 dependent variables each, while LASSO regression (MNL kernel) delivers minimum MCE only for 11 dependent variables. However, after averaging the MCE results over all dependent variables, LASSO regression has a slight edge over kNN and XGB – 0.567, 0.569, and 0.570, respectively (indicating correct classification of 433, 431, or 430 observations out of 1,000).

Overall, the results demonstrate that even the best-performing learners do a rather inadequate job in predicting attitudes given the available inputs: on average, LASSO regression predictions are only $\Delta\text{MCE}=0.034$ more accurate (i.e., with only 34 more correct per 1,000 observations) than assigning the median.

TABLE 9.1: Cross-validation results for the classification problem

Variable	Best learner	Lowest MCE	Median assignment MCE	ΔMCE (median assignment vs. best learner)
<i>A1a_goodcommute</i>	Median/ LASSO/ SVM/ AdaBoost/ kNN	0.433	0.433	0.000
<i>A1b_jobmoney</i>	kNN	0.623	0.628	-0.005
<i>A1c_closestore</i>	AdaBoost/ kNN	0.536	0.537	-0.001
<i>A1d_prefdrive</i>	kNN/ ANN	0.673	0.751	-0.078
<i>A1e_boring</i>	Median/ kNN/ XGB/ ANN	0.577	0.577	0.000
<i>A1f_deadline</i>	XGB	0.555	0.557	-0.002
<i>A1g_yards</i>	XGB	0.590	0.619	-0.029
<i>A1h_newtech</i>	kNN	0.589	0.590	-0.001
<i>A1i_traffic</i>	LASSO	0.588	0.602	-0.014
<i>A1j_transit</i>	LASSO	0.560	0.583	-0.023
<i>A1k_trendset</i>	XGB	0.629	0.663	-0.034
<i>A1l_dayoff</i>	kNN/ XGB/ ANN	0.546	0.547	-0.001
<i>A1m_grocery</i>	LASSO	0.487	0.546	-0.059
<i>A1n_timetowork</i>	XGB	0.572	0.575	-0.003
<i>A1o_eproducts</i>	Recursive tree	0.620	0.630	-0.010
<i>A1p_travelwaste</i>	XGB	0.557	0.569	-0.012
<i>A1q_stresscommute</i>	LASSO	0.509	0.511	-0.002
<i>A1r_goodjob</i>	Median/ LASSO/ SVM/ kNN	0.407	0.407	0.000
<i>A1s_walkbike</i>	LASSO	0.627	0.786	-0.159
<i>A1t_closetransit</i>	kNN	0.614	0.764	-0.150
<i>A1u_liketravel</i>	Recursive tree/ AdaBoost/ kNN/ XGB/ ANN	0.492	0.493	-0.001
<i>A1v_useminute</i>	XGB	0.638	0.720	-0.082
<i>A1w_techproblems</i>	MNL/ LASSO	0.644	0.709	-0.065
<i>A1x_destination</i>	AdaBoost/ kNN	0.450	0.451	-0.001
<i>A1y_payquicktrip</i>	SVM	0.645	0.683	-0.038
<i>A1z_transitovercar</i>	LASSO	0.602	0.766	-0.164
<i>A1aa_noisyshops</i>	AdaBoost	0.593	0.749	-0.156
<i>A1ab_hurry</i>	Recursive tree	0.650	0.737	-0.087
<i>A1ac_carjustmove</i>	LASSO/ XGB/ ANN	0.498	0.500	-0.002
<i>A1ad_wanttravel</i>	SVM	0.639	0.823	-0.184
<i>A1ae_likeinternet</i>	kNN	0.465	0.467	-0.002
<i>A1af_driving</i>	XGB	0.543	0.546	-0.003
<i>A1ag_busy</i>	ANN	0.532	0.535	-0.003
<i>A1ah_impressivecar</i>	XGB	0.636	0.645	-0.009
<i>A1ai_fewtrips</i>	XGB	0.487	0.494	-0.007
<i>A1aj_welcomecommute</i>	Recursive tree/ kNN	0.594	0.628	-0.034
<i>A1ak_wbovercar</i>	LASSO	0.503	0.553	-0.050

<i>A1al_neverbehind</i>	kNN	0.630	0.782	-0.152
<i>A1am_goodlife</i>	XGB	0.414	0.416	-0.002

As in the regression problem, the generalization errors vary across the board for the attitudinal variables. For only nine variables out of 39 is the MCE lower than 0.5, meaning that prediction success is achieved for more than 50% of the test observations. These variables load on the factors *travel is wasted time* (3 variables), *satisfaction* (2), *commute benefit* (1), *pro-active transportation* (1), and *pro-technology* (1), in addition to the statement “*I prefer to organize my errands so that I make as few trips as possible*”, which is not included in the factor analysis. However, in all but one case the prediction rate is attained mainly because of the variable distributions (demonstrating extreme peakedness) rather than the performance of a sophisticated learning function: the MCE deviation from the median assignment learner is 0 (i.e., the median is the best or one of the best predicting functions) or very close to 0. Only for the statement, “*I like the idea of living in a neighborhood where I can walk to the grocery store*”, is the MCE 0.487 for the LASSO regression, which constitutes a 0.059 improvement over assigning the median. When the distribution of \mathcal{Y} is not very peaked, conditional learning functions could noticeably improve the generalization errors. For example, the SVM learner gains 0.184 of Δ MCE over the median for the statement “*I sometimes travel more than I have to, because I want to*”, indicating that the explanatory variables improved the prediction of this variable.

9.3 Supporting tables

TABLE 9.2: Attitudinal statements in the general opinions section of the MSNCC (N=2,849)

Variable name	Statement	Median value ^a
A1a_goodcommute	My commute is generally pleasant.	Agree
A1b_jobmoney	The main benefit of my job is that it gives me the money to pay for the things I <i>really</i> enjoy doing.	Agree
A1c_closestore	When I need to buy something, I usually prefer to get it at the closest store possible.	Agree
A1d_prefdrive	I'd rather drive than travel by any other means.	Neutral
A1e_boring	The act of traveling is boring.	Disagree
A1f_deadline	I feel more productive when I am under pressure to complete work by a deadline.	Agree
A1g_yards	I like the idea of living somewhere with large yards and lots of space between homes.	Agree
A1h_newtech	I like to track the development of new technology.	Agree
A1i_traffic	Getting stuck in traffic doesn't bother me much.	Disagree
A1j_transit	I like the idea of transit as a means of travel for me.	Agree
A1k_trendset	I often introduce new trends to my friends.	Neutral
A1l_dayoff	Occasionally, I'd be willing to give up a day's pay to get a day off work.	Agree
A1m_grocery	I like the idea of living in a neighborhood where I can walk to the grocery store.	Agree
A1n_timetowork	I do my best work when I have more than enough time to complete it.	Agree
A1o_eproducts	I like to be among the first to own new electronic products.	Disagree
A1p_travelwaste	Time spent traveling is generally wasted time.	Disagree
A1q_stresscommute	My commute is stressful.	Disagree
A1r_goodjob	I am generally satisfied with my job.	Agree
A1s_walkbike	I prefer to walk or bike rather than drive whenever possible.	Neutral
A1t_closetransit	I prefer to live close to transit, even if it means I'll have a smaller home and more people living nearby.	Neutral
A1u_liketravel	I generally enjoy the act of traveling itself.	Agree
A1v_useminute	I feel like I need to make the most of every single minute.	Neutral
A1w_techproblems	Technology brings at least as many problems as solutions.	Neutral
A1x_destination	The only good thing about traveling is arriving at your destination.	Disagree
A1y_payquicktrip	I would pay money to reduce the time I spend traveling.	Neutral
A1z_transitovercar	I prefer to take transit rather than drive whenever possible.	Neutral

(Table 9.2 is continued on the next page)

TABLE 9.2: Attitudinal statements in the general opinions section of the MSNCC (N=2,849) (CONT'D)

Variable name	Statement	Median value
A1aa_noisysshops	Mixing different types of businesses (e.g., shops, restaurants, offices) with the homes in my neighborhood causes (or would cause) too much traffic or noise.	Neutral
A1ab_hurry	I'm often in a hurry to be somewhere else.	Neutral
A1ac_carjustmove	To me, a car is mostly just a way to get from place to place.	Agree
A1ad_wantravel	I sometimes travel more than I <i>have</i> to, because I <i>want</i> to.	Neutral
A1ae_likeinternet	The internet makes life more interesting.	Agree
A1af_driving	I like the idea of driving as a means of travel for me.	Agree
A1ag_busy	I'm too busy to do many things I'd like to do.	Agree
A1ah_impressivecar	I (would) like to own a car that impresses other people.	Disagree
A1ai_fewtrips	I prefer to organize my errands so that I make as few trips as possible.	Agree
A1aj_welcomecommute	My commute serves as a welcome transition between home and work.	Agree
A1ak_wbovercar	I like the idea of walking (or biking) as a means of transportation.	Agree
A1al_neverbehind	I never get very far behind on things I'm trying to get done.	Neutral
A1am_goodlife	I am generally satisfied with my life.	Agree

^a Reporting scale has five levels: "Strongly disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree".

TABLE 9.3: General attitudinal latent constructs (factors)

Constructs ^a	Statements	Pattern matrix loadings ^b
<i>Pro-transit</i> [AVT9_protransit] ^c	I prefer to take transit rather than drive whenever possible.	0.739
	I'd rather drive than travel by any other means.	-0.588
	I like the idea of driving as a means of travel for me.	-0.536
	I like the idea of transit as a means of travel for me.	0.510
<i>Travel is wasted time</i> [AVT9_nec_oftravel]	I generally enjoy the act of traveling itself.	-0.774
	The act of traveling is boring.	0.710
	Time spent traveling is generally wasted time.	0.592
	The only good thing about traveling is arriving at your destination.	0.567
	I sometimes travel more than I have to, because I want to.	-0.389
	To me, a car is mostly just a way to get from place to place.	0.308
<i>Pro-technology</i> [AVT9_protech]	I like to be among the first to own new electronic products.	0.755
	I like to track the development of technology.	0.747
	I often introduce new trends to my friends.	0.577
	The internet makes life more interesting.	0.343
	Technology brings at least as many problems as solutions.	-0.305
<i>Commute benefit</i> [AVT9_comm_ben]	My commute is generally pleasant.	0.773
	My commute is stressful.	-0.769
	My commute serves as a welcome transition between home and work.	0.372
<i>Time pressure – reality</i> [AVT9_timepres_real]	I'm often in a hurry to be somewhere else.	0.674
	I'm too busy to do many things I'd like to do.	0.476
	I feel like I need to make the most of every single minute.	0.433
<i>Time pressure – preference</i> [AVT9_timepres_pref]	I do my best work when I have more than enough time to complete it.	-0.709
	I feel more productive when I am under pressure to complete work by a deadline.	0.532
<i>Pro-active transportation</i> [AVT9_pro_activetrans]	I like the idea of walking (or biking) as a means of transportation.	0.895
	I prefer to walk or bike rather than drive whenever possible.	0.767
	I like the idea of living in a neighborhood where I can walk to the grocery store.	0.420
<i>Satisfaction</i> [AVT9_satisfaction]	I am generally satisfied with my life.	0.806
	I am generally satisfied with my job.	0.550
<i>Pro-density</i> [AVT9_prodensity]	I like the idea of living somewhere with large yards and lots of space between homes.	-0.635
	I prefer to live close to transit, even if it means I'll have a smaller home and more people living nearby.	0.625
	Mixing different types of businesses (e.g., shops, restaurants, offices) with the homes in my neighborhood causes (or would cause) too much traffic or noise.	-0.549

^a Principal axis factor extraction with oblimin rotation was used.

^b Represents the degree of association between the statement and the construct. Only loadings greater than 0.3 in magnitude are reported.

^c Variable name in the input/output datasets.

TABLE 9.4: Socio-economic variables common to the MSNCC and NHTS datasets

NHTS variable name	Variable content (following the NHTS)	Variable description (following the NHTS)
<i>HH_HISP</i>	Hispanic	Binary
<i>HH_RACE</i>	Race	Categorical: 1=White, 2=Black, 3=Asian, 4=Native American, 5=Native Hawaiian, 6=Multi ethnic
<i>DRVRCNT</i>	Number of drivers in HH	Count
<i>HHFAMINC</i>	Derived total annual HH income	Ordinal: 1 to 16=\$0k to \$80k w/ \$5k increments, 17=\$80-100k, 18= >\$100k
<i>HHSIZE</i>	Count of HH members (HHMs)	Count
<i>HHVEHCNT</i>	Count of HH vehicles	Count
<i>NUMADLT</i>	Count of adult HHMs at least 18 years old	Count
<i>WRKCOUNT</i>	Number of workers in HH	Count
<i>LIF_CYC</i>	Life cycle classification for the HH	Categorical: 1 to 10=combination of adults and children of various age categories.
<i>BORNINUS</i>	Respondent was born in U.S.	Binary
<i>CONDNIGH</i>	Medical condition results in limiting driving to daytime	Binary
<i>CONDPUB</i>	Medical condition results in using bus/subway less frequently	Binary
<i>DRIVER</i>	Driver status of respondent	Binary
<i>EDUC</i>	Highest grade completed	Ordinal: 1= <HS, 2=HS, 3=Some college, 4=Bachelor's, 5=Graduate degree
<i>GCDWORK</i>	Great circle distance (miles) between home and work	Continuous
<i>OCCAT</i>	Job category	Categorical: 1=Sales/service, 2=Clerical/admin, 3=Manufacturing, 4=Profess./managerial, 97=Other
<i>R_AGE</i>	Respondent age (years)	Continuous
<i>R_SEX</i>	Respondent gender	Binary: 1=Male, 2=Female
<i>SELF_EMP</i>	Self-employed	Binary
<i>TIMETOWK</i>	Minutes to go from home to work last week	Continuous
<i>TRAVDAY</i>	Travel day – day of week	Categorical: 1=Sunday,..., 7=Saturday
<i>WKFTPT</i>	Work full or part-time	Binary: 1=Full-time, 2=Part-time
<i>WORKER</i>	Respondent worker status	Binary
<i>WRKTRANS</i>	Transportation mode to work last week	Categorical: 1=Car, 2=Van, 3=SUV, 4=Pickup truck, 5=Other truck, 6=RV, 7=Motorcycle, 8=Light EV, 9=Local bus, 10=Commuter bus, 11=School bus, 12=Charter bus, 13=Intercity bus, 14=Shuttle bus, 15=Amtrak, 16=Commuter train, 17=Subway/elevated, 18=Streetcar, 19=Taxi, 20=Ferry, 21=Airplane, 22=Bicycle, 23=Walk, 24=Spec transit, 97=Other
<i>DISTTOWK</i>	One-way distance to workplace (miles)	Continuous
<i>TDAYDATE</i>	Date of travel day (YYYYMM)	Date

TABLE 9.5: Comparison of selected variable distributions in the source (NHTS) and target (MSNCC) domains

Variable	Mean		Test stat. ^a	p-value
	Source	Target		
<i>HH_HISP</i>	0.08	0.08	0.014	0.90
<i>HH_RACE: White</i>	0.66	0.86	694.175	0.00
<i>HH_RACE: Black</i>	0.04	0.05	6.268	0.01
<i>HH_RACE: Asian</i>	0.15	0.03	1072.830	0.00
<i>HHFAMINC: \$0-25k</i>	0.08	0.07	0.065	0.80
<i>HHFAMINC: \$25-50k</i>	0.14	0.20	45.504	0.00
<i>HHFAMINC: \$50-75k</i>	0.19	0.20	1.141	0.29
<i>HHFAMINC: \$75-100k</i>	0.19	0.19	0.001	0.97
<i>HHFAMINC: >\$100k</i>	0.38	0.29	70.913	0.00
<i>DRIVER</i>	0.96	0.98	47.174	0.00
<i>EDUC: less than HS degree</i>	0.00	0.04	79.752	0.00
<i>EDUC: HS degree</i>	0.03	0.23	530.803	0.00
<i>EDUC: less than BS/BA degree</i>	0.24	0.29	29.938	0.00
<i>EDUC: BS/BA degree</i>	0.31	0.24	62.672	0.00
<i>EDUC: graduate degree</i>	0.42	0.18	853.179	0.00
<i>OCCAT: service</i>	0.06	0.25	433.764	0.00
<i>OCCAT: clerical</i>	0.15	0.12	9.878	0.00
<i>OCCAT: manufacture</i>	0.02	0.13	282.541	0.00
<i>OCCAT: professional</i>	0.64	0.44	404.836	0.00
<i>R_SEX: female</i>	0.61	0.49	127.523	0.00
<i>WORKER</i>	0.94	1.00	1821.530	0.00
<i>WRKTRANS: car</i>	0.51	0.94	6204.507	0.00
<i>WRKTRANS: motorcycle</i>	0.01	0.01	0.076	0.78
<i>WRKTRANS: local bus</i>	0.06	0.01	435.505	0.00
<i>WRKTRANS: express bus</i>	0.07	0.00	1969.312	0.00
<i>WRKTRANS: heavy rail</i>	0.08	0.01	1560.817	0.00
<i>WRKTRANS: light rail</i>	0.16	0.00	7757.984	0.00
<i>WRKTRANS: bicycle</i>	0.10	0.01	2485.862	0.00
<i>WRKTRANS: walk</i>	0.02	0.02	0.160	0.69
<i>DRVRCNT</i>	2.20	2.23	0.097	0.00
<i>HHSIZE</i>	2.70	2.93	0.086	0.00
<i>HHVEHCNT</i>	2.04	2.58	0.214	0.00
<i>NUMADLT</i>	1.97	2.23	0.138	0.00
<i>WRKCOUNT</i>	2.19	1.80	0.163	0.00
<i>R_AGE</i>	43.87	47.57	0.139	0.00
<i>TIMETOWK</i>	44.66	24.24	0.343	0.00
<i>DISTTOWK</i>	21.16	14.29	0.150	0.00

^a Kolmogorov-Smirnoff and chi-square statistics are used for continuous and categorical variables, respectively.

TABLE 9.6: Descriptive statistics of the observed continuous attitudes for the MSNCC (N=2,352)

Variable	Mean	SD	Median	Min	Max	Skew	Kurtosis
<i>Pro-transit</i>	-0.01	1.00	-0.05	-3.06	2.76	0.03	-0.46
<i>Travel is wasted time</i>	0.01	1.00	-0.15	-2.39	3.62	0.46	0.03
<i>Pro-technology</i>	-0.01	1.01	-0.04	-2.85	2.95	0.01	-0.26
<i>Commute benefit</i>	0.00	1.00	0.23	-3.60	2.00	-0.76	0.47
<i>Time pressure – reality</i>	0.01	1.00	0.01	-3.06	3.94	0.03	-0.34
<i>Time pressure – preference</i>	0.02	0.99	-0.01	-2.85	3.02	0.03	-0.32
<i>Pro-active transportation</i>	0.00	1.00	0.15	-2.94	1.80	-0.45	-0.38
<i>Satisfaction</i>	0.01	1.00	0.09	-4.12	1.96	-0.89	1.27
<i>Pro-density</i>	0.00	1.01	-0.07	-3.02	2.77	0.16	-0.30

TABLE 9.7: Descriptive statistics of the predicted continuous attitudes for the MSNCC (N=2,352)

Variable	Mean	SD	Median	Min	Max	Skew	Kurtosis
<i>Pro-transit</i>	-0.01	0.46	-0.04	-1.12	1.21	0.15	-1.11
<i>Travel is wasted time</i>	0.01	0.10	0.01	-0.39	0.27	-0.33	-0.11
<i>Pro-technology</i>	-0.01	0.25	-0.02	-0.63	1.42	0.38	0.22
<i>Commute benefit</i>	0.00	0.34	0.00	-1.14	1.15	0.05	-0.37
<i>Time pressure – reality</i>	0.01	0.11	0.01	-0.38	0.55	0.06	-0.11
<i>Time pressure – preference</i>	0.02	0.21	0.04	-1.05	0.48	-0.28	-0.14
<i>Pro-active transportation</i>	0.00	0.45	-0.09	-1.05	1.53	0.96	0.53
<i>Satisfaction</i>	0.01	0.20	0.02	-1.06	1.17	-0.47	1.36
<i>Pro-density</i>	0.00	0.48	-0.05	-1.32	1.81	0.42	0.16

TABLE 9.8: Cross-validation results for the regression problem

Variable	MSE											Best learner	Δ MSE (mean assignment vs. best learner)
	RHD	Assigning the mean	Forward stepwise linear regression	CART	Evolutionary regression tree	Recursive tree	Bagging	Random forest	LASSO	SVM	AdaBoost		
<i>Pro-transit</i>	2.021	0.993	0.895	0.829	0.815	0.816	0.826	0.823	0.757	0.801	0.771	LASSO	-0.236
<i>Travel is wasted time</i>	1.952	1.001	1.146	1.003	1.000	0.992	1.071	1.078	0.985	1.005	1.012	LASSO	-0.016
<i>Pro- technology</i>	2.035	1.017	1.110	0.971	0.965	0.962	1.022	1.026	0.951	0.968	0.971	LASSO	-0.066
<i>Commute benefit</i>	2.080	1.008	1.279	0.932	0.930	0.904	0.959	0.963	0.898	0.961	0.919	LASSO	-0.110
<i>Time pressure – reality</i>	2.003	1.009	1.169	1.002	0.994	1.002	1.075	1.083	1.003	0.995	1.022	Evolutionary regression tree	-0.015
<i>Time pressure – preference</i>	1.903	0.994	1.112	0.963	0.968	0.955	1.004	1.005	0.936	0.953	0.954	LASSO	-0.058
<i>Pro-active transportation</i>	2.009	1.009	0.923	0.848	0.847	0.854	0.863	0.866	0.789	0.842	0.811	LASSO	-0.220
<i>Satisfaction</i>	1.904	1.004	1.156	0.995	0.991	0.994	1.045	1.046	0.976	0.993	0.996	LASSO	-0.028
<i>Pro-density</i>	1.970	1.005	0.848	0.847	0.828	0.829	0.831	0.837	0.748	0.762	0.761	LASSO	-0.257

TABLE 9.9: Cross-validation results for the classification problem

Variable	MCE							
	RHD	Assigning the median	Forward stepwise linear regression	C5.0	Evolutionary classification tree	Recursive tree	Bagging	Random forest
A1a_goodcommute	0.615	0.433	0.526	0.508	0.437	0.434	0.496	0.506
A1b_jobmoney	0.744	0.628	0.658	0.676	0.641	0.629	0.663	0.666
A1c_closestore	0.703	0.537	0.638	0.627	0.547	0.537	0.602	0.601
A1d_prefdrive	0.775	0.751	0.690	0.712	0.679	0.686	0.700	0.715
A1e_boring	0.705	0.577	0.647	0.667	0.608	0.582	0.646	0.648
A1f_deadline	0.703	0.557	0.691	0.655	0.576	0.563	0.617	0.630
A1g_yards	0.725	0.619	0.648	0.663	0.637	0.619	0.638	0.650
A1h_newtech	0.717	0.590	0.632	0.642	0.604	0.596	0.614	0.636
A1i_traffic	0.686	0.602	0.629	0.641	0.618	0.603	0.643	0.652
A1j_transit	0.709	0.583	0.591	0.612	0.601	0.586	0.598	0.614
A1k_trendset	0.722	0.663	0.648	0.685	0.676	0.656	0.664	0.663
A1l_dayoff	0.712	0.547	0.710	0.612	0.562	0.549	0.602	0.604
A1m_grocery	0.672	0.546	0.519	0.554	0.521	0.511	0.550	0.553
A1n_timetowork	0.688	0.575	0.631	0.660	0.590	0.575	0.630	0.631
A1o_eproducts	0.739	0.630	0.672	0.709	0.643	0.620	0.693	0.691
A1p_travelwaste	0.703	0.569	0.656	0.645	0.573	0.570	0.620	0.629
A1q_stresscommute	0.679	0.511	0.585	0.600	0.516	0.514	0.564	0.574
A1r_goodjob	0.585	0.407	0.493	0.483	0.412	0.408	0.475	0.479
A1s_walkbike	0.766	0.786	0.680	0.674	0.679	0.660	0.663	0.683
A1t_closetransit	0.775	0.764	0.655	0.672	0.647	0.638	0.653	0.656
A1u_liketravel	0.645	0.493	0.601	0.596	0.514	0.492	0.537	0.538

(Table 9.9 is continued on the next page)

TABLE 9.9: Cross-validation results for the classification problem (CONT'D)

Variable	MCE							
	RHD	Assigning the median	Forward stepwise linear regression	C5.0	Evolutionary classification tree	Recursive tree	Bagging	Random forest
A1v_useminute	0.729	0.720	0.696	0.691	0.686	0.644	0.684	0.683
A1w_techproblems	0.728	0.709	0.687	0.663	0.678	0.649	0.665	0.667
A1x_destination	0.626	0.451	0.583	0.528	0.458	0.454	0.507	0.516
A1y_payquicktrip	0.711	0.683	0.658	0.687	0.675	0.668	0.682	0.666
A1z_transitovercar	0.762	0.766	0.673	0.678	0.642	0.609	0.650	0.678
A1aa_noisysshops	0.736	0.749	0.674	0.685	0.627	0.618	0.647	0.651
A1ab_hurry	0.707	0.737	0.703	0.704	0.682	0.650	0.692	0.695
A1ac_carjustmove	0.674	0.500	0.663	0.594	0.509	0.500	0.572	0.572
A1ad_wanttravel	0.744	0.823	0.769	0.688	0.653	0.653	0.685	0.688
A1ae_likeinternet	0.618	0.467	0.485	0.533	0.484	0.467	0.520	0.524
A1af_driving	0.676	0.546	0.608	0.598	0.564	0.554	0.575	0.586
A1ag_busy	0.692	0.535	0.656	0.618	0.538	0.537	0.613	0.610
A1ah_impressivecar	0.740	0.645	0.665	0.691	0.664	0.648	0.698	0.696
A1ai_fewtrips	0.563	0.494	0.525	0.533	0.535	0.496	0.543	0.550
A1aj_welcomecommute	0.711	0.628	0.673	0.671	0.626	0.594	0.653	0.648
A1ak_wbovercar	0.705	0.553	0.581	0.596	0.517	0.513	0.560	0.592
A1al_neverbehind	0.723	0.782	0.735	0.678	0.663	0.662	0.667	0.666
A1am_goodlife	0.578	0.416	0.479	0.489	0.426	0.416	0.469	0.472

(Table 9.9 is continued on the next page)

TABLE 9.9: Cross-validation results for the classification problem (CONT'D)

Variable	MCE							Best learner	Δ MCE (median assignment vs. best learner)
	MNL	LASSO	SVM	AdaBoost	kNN	XGB	ANN		
<i>A1a_goodcommute</i>	0.487	0.433	0.433	0.433	0.433	0.434	0.434	Median/ LASSO/ SVM/ AdaBoost/ kNN	0.000
<i>A1b_jobmoney</i>	0.682	0.624	0.639	0.627	0.623	0.627	0.627	kNN	-0.005
<i>A1c_closestore</i>	0.594	0.537	0.545	0.536	0.536	0.537	0.538	AdaBoost/ kNN	-0.001
<i>A1d_prefdrive</i>	0.699	0.680	0.692	0.674	0.673	0.675	0.673	kNN/ ANN	-0.078
<i>A1e_boring</i>	0.639	0.583	0.588	0.579	0.577	0.577	0.577	Median/ kNN/ XGB/ ANN	0.000
<i>A1f_deadline</i>	0.593	0.560	0.560	0.562	0.556	0.555	0.556	XGB	-0.002
<i>A1g_yards</i>	0.642	0.609	0.612	0.602	0.597	0.590	0.614	XGB	-0.029
<i>A1h_newtech</i>	0.617	0.590	0.599	0.594	0.589	0.590	0.590	kNN	-0.001
<i>A1i_traffic</i>	0.611	0.588	0.604	0.617	0.595	0.592	0.598	LASSO	-0.014
<i>A1j_transit</i>	0.611	0.560	0.574	0.581	0.575	0.577	0.571	LASSO	-0.023
<i>A1k_trendset</i>	0.678	0.642	0.659	0.649	0.630	0.629	0.643	XGB	-0.034
<i>A1l_dayoff</i>	0.601	0.547	0.552	0.547	0.546	0.546	0.546	kNN/ XGB/ ANN	-0.001
<i>A1m_grocery</i>	0.524	0.487	0.505	0.501	0.495	0.500	0.514	LASSO	-0.059
<i>A1n_timetowork</i>	0.627	0.575	0.587	0.585	0.576	0.572	0.573	XGB	-0.003
<i>A1o_eproducts</i>	0.658	0.629	0.635	0.625	0.629	0.626	0.631	Recursive tree	-0.010
<i>A1p_travelwaste</i>	0.619	0.566	0.578	0.564	0.562	0.557	0.566	XGB	-0.012
<i>A1q_stresscommute</i>	0.563	0.509	0.517	0.515	0.510	0.510	0.511	LASSO	-0.002
<i>A1r_goodjob</i>	0.449	0.407	0.407	0.408	0.407	0.409	0.409	Median/ LASSO/ SVM/ kNN	0.000
<i>A1s_walkbike</i>	0.649	0.627	0.678	0.649	0.658	0.677	0.674	LASSO	-0.159
<i>A1t_closetransit</i>	0.652	0.615	0.621	0.620	0.614	0.625	0.645	kNN	-0.150
<i>A1u_liketravel</i>	0.528	0.493	0.498	0.492	0.492	0.492	0.492	Recursive tree/ AdaBoost/ kNN/ XGB/ ANN	-0.001

(Table 9.9 is continued on the next page)

TABLE 9.9: Cross-validation results for the classification problem (CONT'D)

Variable	MCE							Best learner	Δ MCE (median assignment vs. best learner)
	MNL	LASSO	SVM	AdaBoost	kNN	XGB	ANN		
<i>Alv_useminute</i>	0.706	0.645	0.654	0.647	0.648	0.638	0.642	XGB	-0.082
<i>Alw_techproblems</i>	0.664	0.644	0.670	0.656	0.651	0.655	0.654	MNL/ LASSO	-0.065
<i>Alx_destination</i>	0.511	0.452	0.458	0.450	0.450	0.453	0.452	AdaBoost/ kNN	-0.001
<i>Al_y_payquicktrip</i>	0.665	0.647	0.645	0.657	0.654	0.653	0.662	SVM	-0.038
<i>Alz_transitovercar</i>	0.656	0.602	0.654	0.607	0.652	0.673	0.628	LASSO	-0.164
<i>Alaa_noisysshops</i>	0.651	0.602	0.603	0.593	0.601	0.598	0.605	AdaBoost	-0.156
<i>Alab_hurry</i>	0.682	0.660	0.677	0.670	0.664	0.670	0.670	Recursive tree	-0.087
<i>Alac_carjustmove</i>	0.541	0.498	0.507	0.501	0.499	0.498	0.498	LASSO/ XGB/ ANN	-0.002
<i>Alad_wanttravel</i>	0.669	0.660	0.639	0.651	0.640	0.642	0.650	SVM	-0.184
<i>Alae_likeinternet</i>	0.489	0.467	0.471	0.469	0.465	0.466	0.466	kNN	-0.002
<i>Alaf_driving</i>	0.586	0.550	0.557	0.556	0.546	0.543	0.546	XGB	-0.003
<i>Alag_busy</i>	0.580	0.535	0.543	0.535	0.535	0.535	0.532	ANN	-0.003
<i>Alah_impressivecar</i>	0.648	0.639	0.652	0.637	0.641	0.636	0.638	XGB	-0.009
<i>Alai_fewtrips</i>	0.522	0.496	0.510	0.524	0.502	0.487	0.489	XGB	-0.007
<i>Alaj_welcomecommute</i>	0.639	0.597	0.608	0.601	0.594	0.596	0.597	Recursive tree/ kNN	-0.034
<i>Alak_wbovercar</i>	0.536	0.503	0.547	0.513	0.538	0.553	0.517	LASSO	-0.050
<i>Alal_neverbehind</i>	0.667	0.632	0.650	0.640	0.630	0.634	0.648	kNN	-0.152
<i>Alam_goodlife</i>	0.448	0.417	0.415	0.415	0.417	0.414	0.415	XGB	-0.002

TABLE 9.10: External validation framework full results: linear regression VO model results

Model specification^a	1	2	3	4	5	6	7
<i>Dataset</i>	Source	Source	Source	Source	Target	Target	Target
<i>Attitudes</i>	Observed	N/A	Predicted	Predicted	Predicted	Predicted	N/A
<i>Specification</i>	Best	1 w/o atts.	1	New best	1	New best	6 w/o atts.
<i>Adjusted R-squared</i>	0.4544	0.4209	0.4559	0.4565	0.3849	0.3894	0.3848
Variable^b							
<i>Intercept</i>	0.890***	0.944***	0.754***	0.579***	0.576***	0.670***	0.756***
<i>Pro-transit</i>	-0.106***	0	-0.190***	-0.194***	0.025	0.008	0
<i>Pro-active transportation</i>	-0.156**	0	-0.072	-0.086	-0.082***	-0.110***	0
<i>Pro-density</i>	-0.160***	0	-0.461***	-0.471***	-0.291***	-0.228***	0
<i>HH_HISP</i>	0	0	0	0	0	-0.057***	-0.048**
<i>HH_RACE: Black</i>	-0.269**	-0.273**	-0.304**	-0.318***	-0.073***	-0.077***	-0.070***
<i>HH_RACE: Asian</i>	0	0	0	-0.124*	0	-0.064**	-0.062**
<i>HH_RACE: Multi</i>	-0.838**	-0.842**	-0.906***	-0.893***	0.004	0	0
<i>HH_RACE: Other</i>	-0.355*	-0.364*	-0.444*	-0.457**	-0.012	0	0
<i>HHFAMINC: \$0-25k</i>	-0.706***	-0.706***	-0.687***	-0.497***	-0.550***	-0.538***	-0.526***
<i>HHFAMINC: \$25-50k</i>	-0.470***	-0.444***	-0.475***	-0.291***	-0.316***	-0.312***	-0.296***
<i>HHFAMINC: \$50-75k</i>	-0.356***	-0.339***	-0.356***	-0.168**	-0.156***	-0.153***	-0.139***
<i>HHFAMINC: \$75-100k</i>	-0.182***	-0.156**	-0.193***	0	-0.074***	-0.071***	-0.057***
<i>HHFAMINC: >\$100k</i>	0	0	0	0.201***	0	0	0
<i>Was born in the US?</i>	0.117**	0.123**	0.131**	0.107*	0.129***	0.101***	0.114***
<i>Condition preventing using public transit</i>	0	0	0	0	0	-0.261***	-0.246***
<i>EDUC: less than HS degree</i>	0	0	0	0	0	0.077***	0.112***
<i>EDUC: HS degree</i>	0	0	0	0	0	0.041***	0.153***
<i>EDUC: less than BS/BA degree</i>	0	0	0	0	0	0.072***	0.137***
<i>OCCAT: service</i>	0	0	0	0	0	0.060***	0.054***
<i>OCCAT: clerical</i>	0	0	0	0	0	0.042*	0.033
<i>OCCAT: manufacture</i>	0.509***	0.614***	0.465**	0.461**	0.163***	0.184***	0.185***
<i>OCCAT: professional</i>	0	0	0	0	0	0.062***	0.046**

(Table 9.10 is continued on the next page)

TABLE 9.10: External validation framework full results: linear regression VO model results (CONT'D)

Model specification^a	1	2	3	4	5	6	7
<i>Dataset</i>	Source	Source	Source	Source	Target	Target	Target
<i>Attitudes</i>	Observed	N/A	Predicted	Predicted	Predicted	Predicted	N/A
<i>Specification</i>	Best	1 w/o atts.	1	New best	1	New best	6 w/o atts.
<i>Adjusted R-squared</i>	0.4544	0.4209	0.4559	0.4565	0.3849	0.3894	0.3848
Variable^b							
<i>R_SEX</i>	0	0	0	0	0	-0.025**	-0.008
<i>SELF_EMP</i>	0	0	0	0	0	0.160***	0.136***
<i>Works full time?</i>	0	0	0	0	0	-0.060***	-0.060***
<i>DRVRCNT</i>	0.514***	0.530***	0.511***	0.514***	0.780***	0.765***	0.781***
<i>WRKCOUNT</i>	0.314***	0.300***	0.319***	0.323***	0.128***	0.109***	0.111***
<i>R_AGE</i>	0	0	0	0	0	-0.001*	-3e-04
<i>DISTTOWK</i>	0.001*	0.001*	0.002*	0.001*	0.001***	0.001***	0.001***
<i>Population density</i>	-0.003*	-0.005***	-0.002	0	-0.004***	-0.003***	-0.005***
<i>Activity density</i>	0	0	0	0	0	-0.001***	-0.001***
<i>Jobs per HH</i>	-0.009*	-0.011**	-0.008*	0	2e-05	0	0
<i>Road network density</i>	0	0	0	0	0	-0.007***	-0.013***
<i>Jobs within 45 mins</i>	-2e-6***	-2e-6***	-7e-07	-9e-07**	-2e-07***	0	0
<i>Number of children</i>	-0.038*	-0.018	-0.052**	-0.051**	-0.056***	-0.036***	-0.034***
<i>Presence of children</i>	0	0	0	0	0	-0.052***	-0.042**
<i>DRVRCNT*WRKCOU NT interaction</i>	-0.041***	-0.037***	-0.043***	-0.044***	-0.009	0	0

^aNumbering corresponds to Table 3.5.

^bSignificance is represented by asterisks: * for $p < 5\%$, ** for $p > 1\%$, *** for $p > 0.1\%$). Zeros indicate the coefficient's absence from the model specification.