Project Report

An Investigation of the Contribution of Targeted Marketing Data to the Prediction of Attitudes

Prepared for **Teaching Old Models New Tricks (TOMNET) Transportation Center**

by,

Patricia L. Mokhtarian, Ph.D.
Email: patricia.mokhtarian@ce.gatech.edu
ORCID: https://orcid.org/0000-0001-7104-499X

Giovanni Circella, Ph.D.
Email: giovanni.circella@ce.gatech.edu
ORCID: https://orcid.org/0000-0003-1832-396X

Kari Watkins, Ph.D.
Email: kari.watkins@ce.gatech.edu
ORCID: https://orcid.org/0000-0002-3824-2027

F. Atiyya Shaw
Email: atiyya@gatech.edu
ORCID: https://orcid.org/0000-0001-8717-5118

Xinyi Wang
Email: xinyi.wang@gatech.edu
ORCID: https://orcid.org/0000-0002-3564-9147

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

April, 2019

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. N/A	2. Government Accession No. N/A	3. Recipient's Catalog No. N/A	
4. Title and Subtitle An Investigation of the Contribution of Targeted Marketing Data to the Prediction of Attitudes		**5. Report Date** April 2019 **6. Performing Organization Code** N/A	
7. Author(s) Patricia L. Mokhtarian, https://orcid.org/0000-0001-7104-499X Giovanni Circella, https://orcid.org/0000-0003-1832-396X Kari Watkins, https://orcid.org/0000-0002-3824-2027 F. Atiyya Shaw, https://orcid.org/0000-0001-8717-5118 Xinyi Wang, https://orcid.org/0000-0002-3564-9147		**8. Performing Organization Report No.** N/A	
9. Performing Organization Name and Address School of Civil and Environmental Engineering Georgia Institute of Technology 790 Atlantic Drive, Atlanta, GA 30332		**10. Work Unit No. (TRAIS)** N/A **11. Contract or Grant No.** 69A3551747116	
12. Sponsoring Agency Name and Address U.S. Department of Transportation, University Transportation Centers Program, 1200 New Jersey Ave, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Research Report (2017 – 2018) **14. Sponsoring Agency Code** USDOT OST-R	
15. Supplementary Notes N/A			
16. Abstract This project involves the use of machine learning methods to impute attitudes into the Georgia subsample of the 2016-17 National Household Travel Survey, training the algorithms on the responses to a 2017 attitudinal survey administered to a separate statewide sample in Georgia. The "common variables" needed to train the learning function will include socio-economic/demographic and other variables found in both samples, but will be augmented by (1) land use-related variables (obtained from multiple external sources) associated with respondents' residential neighborhoods, and (2) (for the first time) lifestyle-oriented targeted marketing variables associated with the household/respondent that are purchased from a commercial provider. The project evaluates the effectiveness of targeted marketing variables for this purpose. The objectives of this project are (1) to impute attitudes into the Georgia subsample of the 2016-17 NHTS, training the imputation functions using attitudinally-rich data collected in Fall 2017 from a sample that is (reasonably) representative of the urban and small-town population of the state of Georgia; and (2) to augment the set of "common variables" available for training the imputation process with information from targeted marketing databases. Achievement of both objectives involves testing the efficacy of the imputed attitudes for predicting travel-related choices of interest, using a variety of comparisons.			
17. Key Words Attitudes; Travel Behavior; Data Imputation; Georgia		**18. Distribution Statement** No restrictions.	
19. Security Classif.(of this report) Unclassified	**20. Security Classif.(of this page)** Unclassified	**21. No. of Pages** 21	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TABLE OF CONTENTS

1. INTRODUCTION

Travel demand forecasting models are complex systems of choice models that often operate at less than 10% explanatory power, a fact that may be partially attributable to the lack of attitudes, preferences, perceptions, social and personal values, and other such transportation system user traits (i.e. psychometric data) within the models (e.g. Mokhtarian and Salomon, 1997; Kuppam et al., 1999; Domarchi et al., 2008). A primary reason for the lack of psychometric data available for use in forecasting models is the reduced response rate that accompanies longer surveys, thus resulting in a dearth of psychometric survey questions on major transport surveys such as the U.S. National Household Travel Survey (NHTS). Here, we present an approach to addressing this issue by applying machine learning algorithms to impute attitudinal data from regional small-scale surveys into nationwide surveys such as NHTS.

Specifically, this project will impute attitudes into the Georgia subsample of the 2016-17 National Household Travel Survey (NHTS), by applying machine learning (ML) algorithms that are trained based on responses to a 2017 statewide survey administered to a separate sample in Georgia (Georgia Department of Transportation Emerging Technologies Survey – GDOT Survey). For the ML model training and application process, the GDOT and NHTS samples must share "common variables" (CVs) that are present in both datasets. In this study, the CVs are (1) socio-economic/demographic (SED) variables that are present in both NHTS and GDOT surveys; (2) targeted marketing (TM) variables that are purchased for all respondents from a commercial data compiler/provider; and (3) land use (LU) variables associated with respondents' residential locations and derived primarily from five-year American Community Survey (ACS) estimates, among other sources. A significant contribution of this project is to evaluate the effectiveness of TM variables in aiding the imputation of psychometric traits such as attitudes across surveys, as this is the first time that the efficacy of TM data for this purpose has been tested.

This research has major societal implications that center on the potential for improved travel demand forecasting and behavioral predictions, which would directly facilitate more efficient expenditures, improved infrastructure planning and development, and ultimately increased (travel) satisfaction and quality of life for all. Additionally, if TM data proves to significantly improve attitudinal prediction, this could have wide ranging implications for planning agencies across the country, as they can feasibly integrate this data source into their travel behavior datasets, and ultimately their forecasting models. From a different perspective, improved travel behavior models can also benefit supply side models that rely on understanding transportation system user behaviors for optimizing network performance. Even more broadly, if the methods of this research prove successful, they can be applied to enrich many more large-scale behavior-based surveys with psychometric variables – such as the American Time Use Survey, the Residential Energy Consumption Survey, the National Health and Nutrition Examination Survey, and so on. Thus, the research has the potential to benefit a large number of fields of study and areas of public policy.

2. METHODOLOGY

This research draws upon multiple data sources to inform a final enriched, fused dataset. First, we provide a brief overview of the study process (Section 2.1), followed by a summary of data processing and manipulation procedures (Section 2.2).

2.1 Overview of Study Process

Figure 1 summarizes the methodological process of this study using the transfer learning framework developed in Pan and Yang (2010), and first applied within this particular research context by Malokin et al. (under review). Here, the source domain represents the GDOT dataset (D_S), which is holistically defined as a set of input variables χ_S with probability distributions P(X_S), and output variables to be transferred $\mathcal{Y}_S$ (i.e. the attitudinal variables in this research study). We note here that variable space X is a subset of a larger p-dimensional space χ, and similarly the variable space Y is a subset of a larger q-dimensional space $\mathcal{Y}$. The source domain input variables to be used as part of the algorithm training process represent the common variables between source and target datasets (X_S') and is thus a subset of the total available input variables (X_S). The target domain represents the NHTS dataset (D_T), which is defined holistically as a set of input variables χ_T with probability distribution P(X_T). The target domain input variables that are common to the source and target domains are denoted as X_T', and the additional variables, X_T'', represent variables unique to the target domain (ex. travel behaviors derived from detailed travel diary data not present in the source domain). As indicated in the figure, common variables (X_S' and X_T') will be composed of socio-economic/demographic variables that are equivalent in content (or which can be made equivalent) between the GDOT (source) and NHTS (target) datasets, targeted marketing data that covers a broad array of topic areas, and land use characteristics. Given these definitions, we develop a learning function $f(\cdot)$ that learns to predict Y_S based on X_S', and then we apply this function to X_T' to predict $\hat{Y}_T$. Thus, $Y_S = f_S(X_S') + \varepsilon_S$, and $\hat{Y}_T = f_S(X_T')$, where the learning function f_S is invariant between the source and target domains. The performance of the learning function will be checked using cross-validation, a staple tool for assessing machine learning algorithms. Following this, an external validation procedure will be implemented to assess the added value of the transferred attitudes into the target domain.

2.2 Data Acquisition and Processing

A key challenge and contribution of this study lies in the data processing and manipulation required to effectively work across the four datasets utilized: (1) the GDOT survey dataset, an attitudinally-rich statewide survey conducted by the research team (N ~ 3000); (2) the Georgia subsample of the NHTS dataset, a nationwide travel behavior-focused survey conducted by the U.S. Department of Transportation (N ~ 8000); (3) a Targeted Marketing dataset (p ~ 5500) purchased for all respondents in the NHTS and GDOT survey samples; and (4) a large amalgamation of land use-related variables based upon respondents' addresses, acquired across several data sources (p ~ >10000). The following subsections discuss each of these datasets in turn.

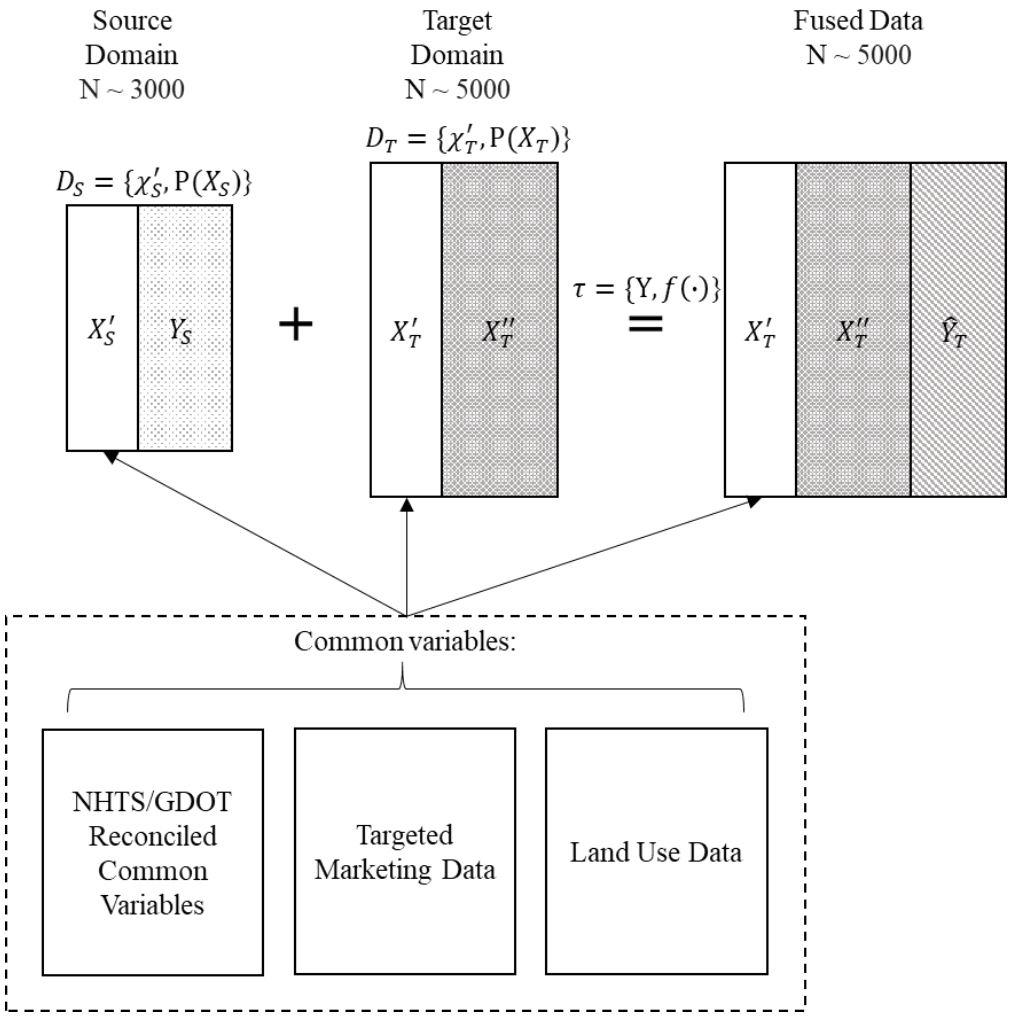

Figure 1. Methodological Overview of Study Process

Source: Derived from van der Putten et al. (2002) and Malokin et al. (under review)

2.2.1 GDOT Survey Dataset

The GDOT survey was conducted from September 2017 to January 2018 (with a minimal number of surveys received after that time period). The comprehensive 14-page survey obtains general attitudes and preferences, technology use, lifestyle-related variables such as employment and relationship status, a wide array of current and future travel-related attitudes, behaviors, and preferences, and socio-economic/demographic characteristics. Such long form, intensive surveys are often limited to regional, small-scale studies, thus motivating the potential of this work to inform larger-scale survey efforts using richer datasets that are more limited in geographic scope. Invitations to complete the GDOT survey were mailed to two groups of respondents: (a) a randomized set of 30,000 names/addresses selected from across 14 Metropolitan Planning Organization (MPO) areas in Georgia (this randomized set of names/addresses was purchased from InfoGroup, a mailing list provider, in Fall 2017), and (b) ~5000 respondents who responded to the NHTS and agreed to be contacted for a follow up survey. Approximately ~1800 of the original

30,000 sampled returned a completed (usable) GDOT survey, and about ~1500 of the ~5000 NHTS respondents sampled returned a usable GDOT survey. At the time of this report, ~3300 valid respondents are retained in the dataset, and TMD data have been purchased for each of these respondents (Figure 2). As always, the number of valid respondents is subject to reductions/ fluctuations as data processing continues; and in fact, this number changes over the course of the data preparation for this particular project. A comprehensive report for the GDOT Emerging Technologies Survey is currently in progress, and is available upon request, should the reader be interested in additional details regarding sampling method, response rate, variable selection, etc. for this survey. The GDOT survey respondents constitute the source dataset (also commonly known in machine learning parlance as the training dataset) in the study methodology (Figure 1).

2.2.2 NHTS Dataset

The NHTS is a repeated cross-sectional nationwide travel behavior survey conducted by the Federal Highway Administration, and deemed by the agency as the "authoritative source on travel behavior of the American public." The NHTS used in this study was the most recent wave, conducted from March 2016 to May 2017, and includes both individual and household-level modules that cover general household characteristics, vehicle ownership attributes, long distance travel behavior, and person-level characteristics including person trips and health. Each survey question is therefore linked to a household ID, with a subset of questions having a person ID that indicates which member of the household responded to the respective person-level question being reported. While the NHTS is nationwide, states and regions are given the opportunity to purchase add-on samples (and/or add additional region-specific questions), thus increasing the number of respondents for which data is available in the respective jurisdiction. The state of Georgia, through GDOT, purchased additional respondents and correspondingly the data used in this study comes from what is referred to in this document as the Georgia subsample of the NHTS (which includes *all* respondents in Georgia: both the Georgia respondents in the core national NHTS sample, and the auxiliary Georgia add-on sample purchased by GDOT). Additional details regarding the NHTS can be accessed at the following repository: https://nhts.ornl.gov/documentation.

We obtained TM data for all respondents in the Georgia subsample of the NHTS. Of these, ~5000 respondents indicated that they could be contacted for a follow-up survey, and as such, these respondents also received the GDOT survey discussed in Section 2.2.1. Approximately ~1500 (of the ~5000) returned usable GDOT surveys, and thus these respondents comprise a subset for whom we have both GDOT survey and NHTS responses; the remainder of these respondents (i.e. those who agreed to be contacted again, but who did not respond to the GDOT survey) are classified as NHTS-only, as only NHTS data for these remaining respondents is available. Another subset of the NHTS respondents (N ~3500) indicated from the inception that they did not want to be contacted again for follow up surveys, and as such, are also retained in the NHTS-only sample (Figure 2). The NHTS-only respondents thus constitute the target dataset (also known as the recipient dataset) in the study methodology (Figure 1).

2.2.3 Targeted Marketing (TM) Dataset

As before discussed, a primary contribution of this study is the use of TM data to expand the common variable space needed for the development of machine learning algorithms that can effectively transfer knowledge between survey datasets. Given the important role of TM in this study, here we provide a detailed overview of the process undertaken to select and acquire TM data across all respondents.

2.2.3.1 Selection of TM Data Provider

More than ten targeted marketing data providers were investigated (ex. Equifax, Experian, Acxiom, GeoSelector, InfoGroup), and of these, four were contacted, engaged, and further explored prior to selecting the chosen provider. For two of these firms, the team purchased small "test" sets of variables for randomly selected samples of respondents and examined the consistency of TM data relative to GDOT and NHTS data. Results from these inquiries made it apparent that TM firms generally cater to a different client-base than academic researchers, and as such, researchers seeking to utilize TM data are advised to anticipate additional time for data acquisition, investigation, cleaning, and processing prior to data use. The TM data provider selected for use in this project works with smaller data purchases but is affiliated with one of the largest TM data providers, and thus accesses the larger firm's database. We note that size of data purchase in TM terms is often assessed based on the number of cases for which the data is being purchased, and not the number of variables – an important point, as we were interested in purchasing the largest number of variables available, but for a limited sample size of roughly 10,000 cases (GDOT and NHTS samples combined, with some duplicates for cases on which name/address confidence were low). This apparent contradiction in needs spawned numerous hiccups in the data acquisition stage, as larger providers were unwilling to provide *all* of their respective database variables given the limited sample size of this study.

Accordingly, several reasons informed the decision to select the chosen TM provider, the most pertinent of which were the firm's willingness and ability to provide a rich selection of variables for the smaller sample size and nontraditional (exploratory) data needs of this research project. In addition to the TM firm's consumer variables, the database acquired also houses supplementary variables purchased from an array of corporations such as Claritas, SEMcasting, etc. Finally, in addition to the apparent richness of variables available for data augmentation, the selected firm provided excellent data documentation and customer service for the duration of the project, in marked contrast to some of the other providers investigated. At the time of purchase, the firm's database housed a total of 5582 variables, all of which were purchased across all cases for this project. The cost of purchase was $1500 per matched one thousand cases, for all 5582 variables.

Of the 5582 total variables available, 1508 represent a set for which there are no variable name release restrictions (i.e. the full names of these variables can be shared publicly); this is the variable set that most marketers (i.e. typical clients for TM firms) select from when purchasing data augmentation services. The additional 4074 variables available for purchase are termed audience propensity variables, and are developed on contract to be sold to certain corporations or

companies, and thus might be updated/changed on a monthly basis. Prior to obtaining the audience propensity variables, it was necessary to sign a legally binding non-disclosure agreement barring disclosure of these variable names. Further, to obtain the full set of variables, we were required to provide an official statement of use explaining the purpose for which these variables would be used, and upon the approval of this statement, to sign additional legal documents regarding the terms of use for these variables. Certain variable subsets (such as sensitive financial variables) required the TM provider to obtain specific approval from the firms that generated those variables before they could be included in the overall purchase for this study. Thus, as can be seen, the process of obtaining a full variable set is a non-trivial undertaking that requires months of communication prior to final approval and variable transmission.

2.2.3.2 TM Data Acquisition

Each TM provider requires a list of names and addresses across all cases to facilitate the purchase and subsequent appending of variables – submitted names/addresses are matched against names and addresses on file in the provider's database, and if the exact name cannot be matched, results based on the address are often returned (with some variable matches degenerating into zip+4 and zip code matches if the attribute is not available for an exact address). It is therefore important to obtain and submit as complete and accurate a list of names/addresses as possible, which is another critical point to note, as this can be a limitation for other researchers hoping to purchase TM data to augment their samples. The need to obtain names and addresses for respondents in both the GDOT and NHTS-only samples resulted in a substantial increase in data manipulation required. As alluded to already in Sections 2.2.1 and 2.2.2, there are various subsets of respondents within the GDOT and NHTS samples; these are described as follows, and shown schematically in Figure 2:

1. NHTS respondents who agreed to be contacted again, and thus received a copy of the GDOT survey, but *did not respond* (NHTS_Agree_DNR).
2. NHTS respondents who did not agree to be contacted again for a follow-up survey, and as such did not receive a copy of the GDOT survey (NHTS_DNAgree_DNR). The union of these two subsamples constitutes the NHTS-only sample.
3. GDOT survey respondents who are not in the NHTS respondent pool (i.e. are unique to the GDOT survey – GDOT_R). Recall that approximately 30,000 respondents **who did not participate in the NHTS** were selected for sampling in the GDOT survey, but only 1808 of these 30,000 sampled returned a usable copy of the GDOT survey.
4. NHTS respondents who agreed to be contacted again, and who *did respond* to the GDOT survey (NHTS_Agree_R). Hence, both GDOT_R and NHTS_Agree_R responded to the GDOT survey, but an overlap between the GDOT and NHTS respondents occurs only for those respondents in the NHTS_Agree_R subset. NHTS_Agree_R respondents are grouped with GDOT_R respondents for the purposes of this particular study, and in some contexts the union of these two groups

may be referred to as the GDOT sample, since all of these respondents completed the GDOT survey.

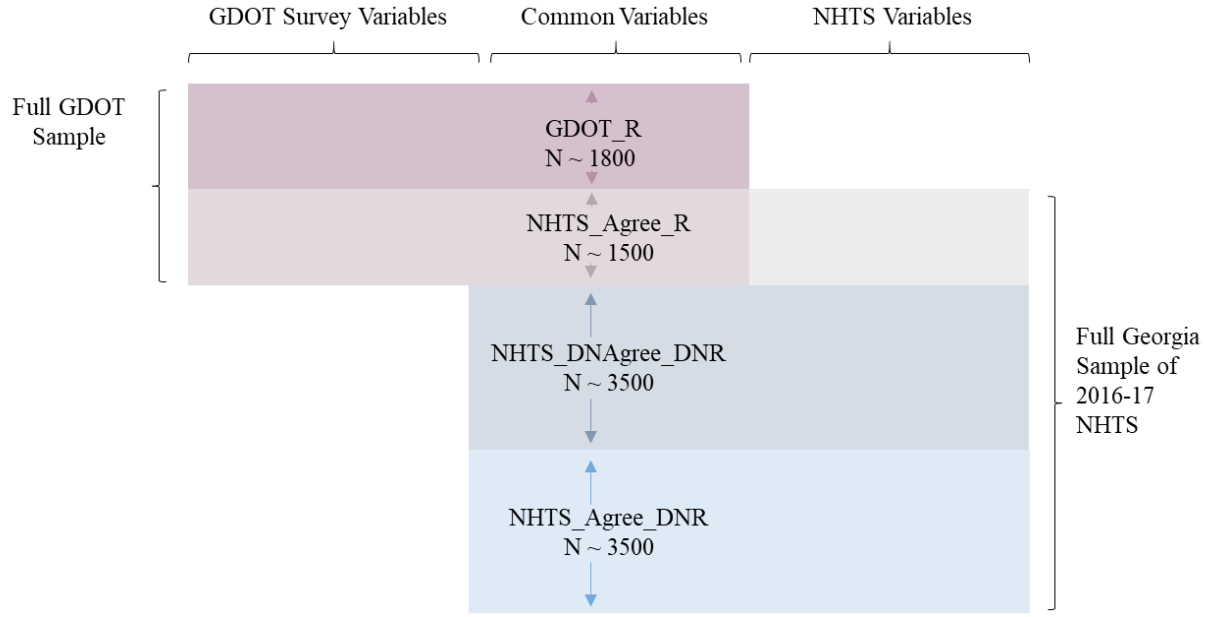

Figure 2. Schematic Representation of GDOT and NHTS Data Subsets

Differentiating among these respondent subsets is critical because each subset has varying name and address information available. For NHTS respondents who had agreed to be contacted again (NHTS_Agree_DNR and NHTS_Agree_R), the names and addresses provided by these respondents had been made available to GDOT and subsequently our research team (so that they could be invited to take the GDOT survey). However, the names and addresses of NHTS respondents who did not agree to be contacted again (NHTS_DNAgree_DNR) were understandably not made available, necessitating several additional steps in order to deduce them (while, of course, honoring their request not to be contacted further).

The home addresses of this latter subset of individuals were obtained using trip diary data from the confidential version of the NHTS (the confidential version can be made available for legitimate research purposes, with appropriate safeguards of respondent privacy). To ascertain their names, we paid the selected TM provider to execute a name/address "append," whereby the TM firm provides names and the corresponding gender for the first three individuals at a particular address. The typical match rate (i.e., the share of addresses for which names are present in the TM database, signifying that the firm has at least some information about one or more people living there) for a name append service, if the firm is appending using its consumer database (and NOT the US Postal Service names and addresses database, which by law is not permitted to be used except for imminent mailings) is 40-60%, according to our TM provider. However, the addresses we provided to our TM data vendor had a 75% match rate, suggesting that the NHTS survey respondents had a higher probability of being included in general consumer-oriented databases

than members of the adult population at large. We acknowledge that this is a potential bias of these respondents, and further make the connection that the names for the GDOT survey (N=30,000) had been purchased from another targeted marketing data firm in August 2017, perhaps suggesting that this group of respondents (GDOT_R) also had the bias of being more likely to be included in consumer databases. We note this here for readers to consider when making decisions about name/address lists to be used for inviting potential respondents to participate, and (when population representativeness is important) recommend ensuring that name/address lists purchased for survey recruitment purposes come from the US Postal Service (USPS) mailing records, and not from consumer databases.

The name append service offered by the TM firm provides names and genders for up to the first three household individuals at each address. As noted, the submitted addresses had no name/gender matches for 25% of the addresses. Thus, these cases (~ 1000) are discarded, since the firm cannot provide data augmentation services for any cases that do not have names. For cases that received at least one individual/address (N ~ 2500) in the original name append, we then obtained the first name, gender, and age for the first, second, third, fourth and fifth individuals in those households through an additional "variable" append (recall that we already had name and gender for the first, second, and third individuals in those households through the "name" append, but the follow-on variable append offered more individuals/household and thus, we used the information from the variable append for all individuals for uniformity). To provide a deeper understanding of the complexity faced throughout this process, we note that the order of the individuals differed between the initial "name" append and follow-on "variable" append (ex.: individual 1 at address x in the name append may be individual 5 at address x in the variable append), thus requiring the team to examine both sets of records to ensure that the correct age from the variable append was matched with the correct individual. Such inconsistencies in the TM data accumulated over the course of this project, and ultimately resulted in significant additional time needed to process all datasets received from the TM firm.

At this point, we investigated gender/age matches between the NHTS households and the first through fifth individuals' data obtained from the TM append services. The purpose of this process was to find the *best* individual match, by age and gender, between the household in the consumer database and all members of each household in the NHTS. We did not want to ultimately retain/use targeted marketing data for *multiple* individuals in a household because that would represent a hierarchical structure, because individuals within the same household are not independent. Accordingly, for households that had two or more individuals with equivalent match levels, we selected a single household member name at random. For cases where the gender matched and the age was within 5 years of the corresponding NHTS person, we assigned a quality flag of 1. For cases where the gender matched, but the age was unknown for all household members, we assigned a quality flag of 2. For cases where there was no gender match in the household, we selected a household member at random and assigned a quality flag of 3.

For the GDOT survey respondent subsets (GDOT_R and NHTS_Agree_R), there are multiple sources of name/address information: (1) from the original mailing list purchased for the GDOT survey; (2) from the home address portion of the survey (Section C); and (3) from the final

page of the survey where respondents indicated their name and address to receive a small token of appreciation (denoted Section L). We note that Section C asked for either an address or intersection, and a portion of respondents opted to report a nearby intersection rather than sharing their exact address. Specifically, the address question obtained in Section C (i.e. the "key aspects of lifestyle" section) of the survey was as follows:

> Knowing more about your neighborhood will help us put your transportation choices and opinions in context. Please give your address or, if you prefer, an intersection (two streets that cross) near your home.
>
> Street address: _____
> City: _____ Zip Code: _____

> The address information obtained in Section L was related to the respondent providing contact information to be used for some, none, or all of three different purposes (receive a token of appreciation, answer questions regarding their survey, or for a follow up survey), and was worded as follows:
>
> In what ways may we contact you? Please provide ALL that apply.
>
> Name: _____
> Telephone:_____ or _____
> Email: _____
> Mail: _____

For the NHTS_Agree_R respondents, as with the other NHTS respondents who agreed to be contacted again, the names and addresses of these respondents were made available through NHTS, and these were the names and addresses to which the GDOT survey was mailed. However, given that there are numerous situations in which the NHTS name/address to which the GDOT survey was mailed may differ from the name/address of the actual eventual respondent, we did not use NHTS-provided names and addresses for the NHTS_Agree_R respondents. Thus, we proceed with discussing and processing the GDOT_R and NHTS_Agree_R respondents' address lists together here.

We developed detailed name and address flags to cross check names and addresses from the three sources for the GDOT respondents. The flags were developed to code respondents depending on whether their self-reported names and/or addresses differed from the mailing list names and addresses to which that unique survey (uniqueness identified by access code and preserved in another field, as detailed later in this report) was delivered. For example, if a survey addressed to person x was delivered to y address, and filled out by person z at y address, and the respondent reported her name and address accurately on the filled out survey, this flag would capture the fact that the person is different, but the address to which the survey was mailed is consistent. We developed this system to help in selecting which name and address combination

should be submitted for TM data augmentation. For ~450 respondents, duplicate cases (either two or three) were submitted for TM data augmentation due to uncertainties or differences regarding the self-reported name/address and the database name/address. Duplicate cases represent a form of insurance for obtaining the best possible match rate in the data augmentation process.

Thus, as can be seen, the process of developing name and address lists for TM augmentation of existing survey respondents was a complicated undertaking. The team has prepared an internal memo with further technical details on this process, and this memo is available upon request. We have provided extensive information on this portion of the data processing, as we acknowledge that it may constitute a limitation of using TM data for large-scale survey enhancement in the future.

2.2.4 Land Use Dataset

Data acquisition for this portion of the project is still in progress. We have obtained data from the U.S. Environmental Protection Agency Smart Location Database 2013 and the five-year estimates (2013-2017) of the American Community Survey at the block group or tract levels, depending on the respective level of the variables. Additional data sources for land use-related variables are currently being explored.

2.2.5 Overall Data Preparation

Following the acquisition of the four primary datasets, further efforts were required to clean and merge these datasets prior to analysis. Here we detail the progress currently completed, but again note that this portion of the project is still in progress.

2.2.5.1 Data Cleaning and Matching

As summarized in Figures 1 and 2, the common variable space in part comprises variables that are present in both the GDOT and NHTS surveys. Therefore, the next step after obtaining TM data was to match each NHTS respondent in our dataset with the appropriate person ID in each household. This is because we have only names and addresses for these individuals, and in order to use the NHTS data to derive common variables, we need to be able to link the names and addresses to the correct record in the Georgia subsample of the NHTS. The three variables selected for use in the matching are: gender, age, and education level, in order of importance. As such, cases for which gender, education, and age level are all missing were removed from the dataset, as appropriate matches between the TMD data and person IDs cannot be established without this information. For the NHTS-only samples (NHTS_Agree_DNR and NHTS_DNAgree_DNR), TMD data was checked relative to the NHTS data to determine person ID matches. For the NHTS_Agree_R sample, GDOT data was checked relative to the NHTS data to determine person ID matches. In both instances, gender mismatches were first identified, and these cases removed. We retained cases for which gender was missing in either the TMD or the NHTS sample, as these could not yet be definitively ruled out. Next, we retained person IDs in each household with the minimum age and education difference between the respective data sources being used for cross checking. Following this, we removed cases that fell outside of an age tolerance of +/- four years,

and outside of an education level tolerance of +/- two years. Finally, we applied manual editing to remove any cases for which there was still more than one equivalent match per household between the TM and NHTS data (as each case must be matched to only one record in the NHTS data).

The second step was to remove GDOT survey cases that either did not match well between the TM data and GDOT data or for which duplicate records were submitted for TM acquisition. Similar steps were followed as reported in the preceding paragraph, with gender mismatches first removed, followed by the institution of age and education minimums and tolerances. However, more than 100 duplicates (i.e. two or more records per unique case) remained following these steps, and as such, significant manual cross checking was required to eliminate duplicate cases. Specifically, we used other variables such as TM segmentation variables to identify which cases represented the most tenable match between the TM and GDOT data. The data processing steps described in this section are summarized in Figure 3.

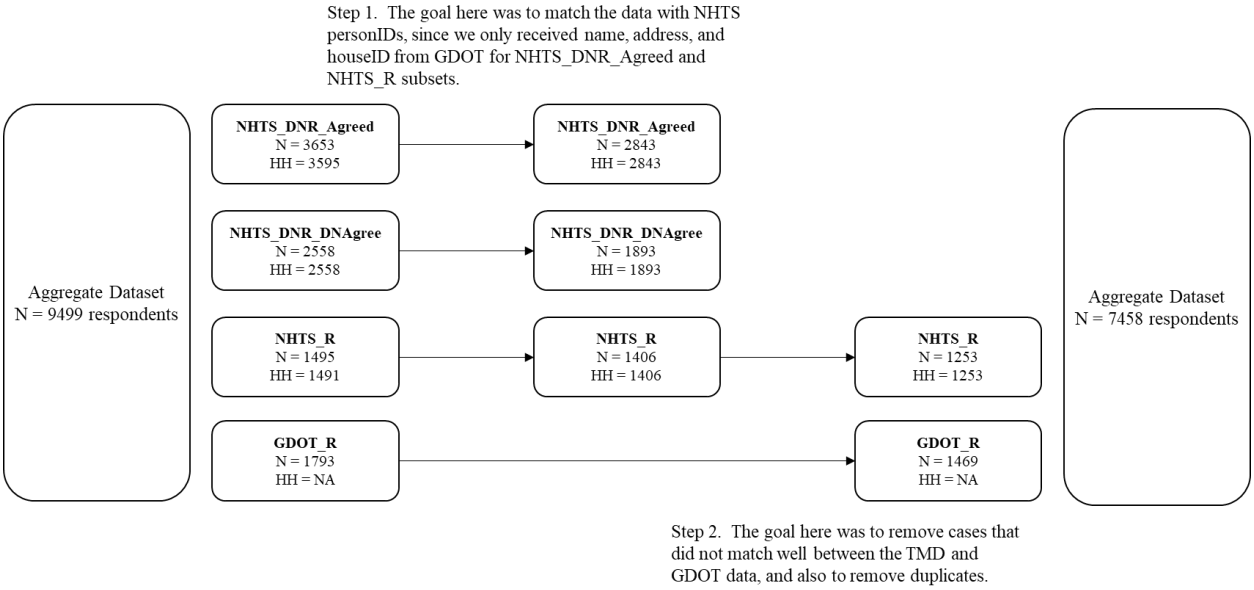

Figure 3. Data Matching for NHTS IDs and Duplicate Case Removal

N = number of cases and HH = total of unique households

* Note: numbers are subject to fluctuation in future stages of this work

As this section details, over the course of the data matching process, the total number of respondents was reduced by approximately 21%, with cases that did not match well with the TM data being screened out during either the first or second steps. Removal of cases due to the institution of various thresholds is in general known to introduce bias into the remaining datasets; and in the case of this particular study, we may anticipate that the retained cases are those that are more likely to have records in consumer databases. These may be individuals/households with increased spending footprints, and/or those more likely to have credit cards/credit histories, among other possibilities. However, given that a primary goal of this project is to investigate the contributions of targeted marketing data to cross-survey imputation, these biases are an unfortunate

inherent side-effect of the study's goals; therefore, while we cannot ameliorate this limitation, we strive to not lose sight of it in the overall study context.

2.2.5.2 Missing Value Imputation

Following the completion of data cleaning, missing values in each of the datasets needed to be independently handled to avoid listwise deletion that would significantly reduce the sample size and (further) bias results. We note that while some machine learning algorithms can handle missing values, this is not the case for all algorithms applied in this study. The variables to be imputed include many of those in the common variable set (X'_S and X'_T) used for training the algorithms. Some of the common sociodemographic variables, as well as some of the items associated with the attitudinal variables that represent the output variables to be transferred (Y_S), had already been imputed for the GDOT sample, as part of the separate analysis of that survey which is underway in parallel with this TOMNET project.

At this stage in the process, the common variables that are derived from the TM dataset have been imputed, and here we provide a brief overview of that process. The TM dataset (original number of variables: p ~ 5500) comprises numeric, ordinal, and nominal (dichotomous and polytomous) variables, and within these, there are a wide range of variable types. For example, some numeric variables are dates, while others are model scores ranging from 1 to 20 or 1 to 100, others such as home prices have unknown upper limits, and still other variables (such as net worth) may have negative values. Given that the TM dataset has many occurrences of variables that are shown to be Not Missing at Random (NMAR), variables and cases with greater than 5% missing values, as well as variables with zero variance (i.e. constant variables) were removed from the dataset (~800 variables, ~10 cases). Following this, missing data across all numeric, nominal, and ordinal variables were imputed using a single imputation based on the Random Forest algorithm, implemented using R 3.5.2 (R Core Team, 2018) with package "missForest" version 1.4 (Stekhoven & Bühlmann, 2012; Stekhoven, 2013). After imputation, variables with correlations of 1 were identified, and the variable in the pair that had the largest mean absolute correlation (i.e. the overall correlation with the other variables in the dataset) was dropped, resulting in ~ 4800 variables and ~7450 respondents in the final TMD dataset (R package "caret" is used; Kuhn, 2018).

Variable imputation for common variables in the NHTS, GDOT, and land use datasets is currently underway, and is expected to proceed similarly to the method used for TM imputation.

3. OVERVIEW OF DATA

Having completed the data acquisition and processing procedures, we now describe the common variables present in the NHTS, GDOT, and TMD datasets.

3.1 NHTS and GDOT Survey

While the NHTS and GDOT survey have substantially different purposes and designs, there is a core set of socio-economic/demographic and travel-related survey questions that overlap between the two datasets with regard to content obtained. At this stage, the socio-economic/demographic variables will be included as part of the common variable set for training the machine learning

algorithms, while the travel-related questions will be retained for potential use in the external validation step. Table 1 summarizes the common variables derived from the GDOT and NHTS surveys, along with their corresponding values, many of which have been manipulated from their original formats to maximize the congruence between datasets. While we do not include final descriptive statistics on these variables for the source and target domains (as these numbers are not yet final), preliminary examination of the distributions of responses for the common socio-economic/demographic variables indicates a high level of similarity between the NHTS and GDOT datasets, which is promising for the cross-application of the invariant learning function that will be developed as part of the methodology for this study.

Table 1. Common variables between NHTS and GDOT datasets

Variable name	Value labels
Household income	Ordinal: < $25K; $25K to $49,999; $50K to $74,999; $75,000 to $99,999; $100K to $149,999; ≥$150K
Household size	Numeric
Household member age group[1]	Numeric sub-variables: No. within age range: 0-6; 6-14; 15-17; 18-26; 27-34; 35-50; 51-65; >65 years
Gender	Nominal: Male; Female
Age	Numeric
Race[2]	Nominal: Asian/Pacific Islander; Black/African American; Native American; White/Caucasian; Other
Ethnicity	Nominal: Not Hispanic/Latino; Hispanic/Latino
Education	Nominal: Some grade school/high school; Completed high school or GED; Come college/technical school; Bachelor's degree; Completed graduate degree
Worker	Nominal: Worker; Nonworker
Employment situation[2]	Nominal: Two or more jobs; Homemaker/caregiver; Student; Retired; Unemployed; Other
Main occupation	Nominal: Professional, managerial, or technical; Sales/service; Manufacturing, construction, maintenance, or farming; Clerical or administrative support; Other
Driver	Nominal: Yes/No
Number of drivers in household	Numeric
Number of vehicles in household	Numeric
Vehicle year	Numeric
Vehicle make	Nominal: 53 categories[3]
Vehicle type	Nominal: Automobile/car/station wagon; Van (mini/cargo/passenger); SUV; Pickup truck; Other truck; RV (recreational vehicle); motorcycle/motorbike; Other
Mode choice to work/school	Nominal: Car driver (alone); Carpool (driver or passenger); Bus/train: Walk; Bicycle; Flight; Other
Network distance from home to work/school	Numeric
Circular distance from home to work/school	Numeric

Minutes from home to work/school	Numeric
Telecommute	Numeric
No. of rideshare trips	Numeric
No. of carshare trips	Numeric
No. of bike trips	Numeric
No. of public transit trips	Numeric
Medical condition that prevents travel	Nominal: Yes; No

[1]Each of these categories is in and of itself a separate numeric variable. For example, age range 0-6 would constitute one variable that contains a count of the number of household members in that age range.
[2]Because these nominal variables are not mutually exclusive, dichotomous (i.e. dummy) variables for each nominal category will be used instead.
[3]While 53 represents the total number of possible categories, the actual number of categories in the dataset may differ.

3.2 Targeted Marketing Data

Table 2 summarizes the variable classification distribution across the TM variables retained in the common variable set after data cleaning and imputation. The variables are classified into the following topic areas: sociodemographic, land use, attitudes, lifestyle, financial, technology, and transport variables. Given traditional TMD sources of credit card and shopping records, it is intuitive that approximately 60% of the TMD variables are consumer-related variables such as purchase behavior, while 18% are financial variables related to investment, income, and insurance, among others. Given the large number of variables in this dataset, dimension reduction procedures are being explored to address the potential curse of dimensionality that can arise in some algorithms when using large numbers of variables to train the models.

4. SUMMARY OF WORK

Over the first year of this project, the research team completed data acquisition and preparation for the NHTS/GDOT and TM common variables. As before discussed, working with TM data provided a plethora of unexpected challenges that were time-intensive to address, and which ultimately indicated that TM data, as we expect is the case with the majority of external passive sources of data, requires extensive cleaning and classification before use in modeling efforts. We hope that the details provided here, as well as the methodological approaches being developed, will provide a valuable roadmap forward for other researchers who hope to utilize TM data in their respective efforts. At this stage in the project, we are in the process of completing data acquisition for the land use variables, which comprise the final component of the common variable set that will be used in this study. We have begun testing standard, base models (such as stepwise linear regression) as well machine learning algorithms (such as extreme gradient boosting and random forest) using the already-processed CVs. Preliminary results indicate that for some factors there are substantial correlations between predicted and observed attitudinal factors in the source domain, particularly for topical areas that have high proportions of TMD variables (ex. technology). These results are promising, and the team looks forward to forthcoming findings.

Table 2. Classification of Targeted Marketing Variables (p = 4813[1])

Section	p	Category	p	Subcategories
Sociodemographic	246	Composition	186	HH Structure, Age, Gender, Life stage, Background
		Education	13	Level, Background
		Life event	11	Move, Divorce, Home buyer, Relationship
		Work	11	Occupation, Employment status
		Housing	11	Length of residence, Home owner, Codes, Density, Dwelling
		Political Indicators	14	Current affairs, Party membership, Political districts, Political views
Consumer-related	2941	Consumer Behavior	572	Home, Food, Automotive, Arts/Antiques, Clothing, Cause-related donations, Tobacco, Green Living, Leisure, Baby/Children, Books/magazines, Business, Channel, Classic car owner, Cost, Transaction, Home/home appliances, TV/Movie/Video, Holiday, Gift, Collectibles, Crafts, Home office/stationary, Health, Personal care, Lifestyle, General merchandise, Electronics, Novelty, Pets, Travel, etc.
		Consumer Propensity	2203	
		Consumer Interests	146	
		Consumer Attitudes	20	Saving, Consumerism, Shopping, Personal interest, Health, Environment
Financial	907	Financial Behavior	56	Assets, Cash, Credit risk, Income, Insurance, Economic stability, Credit/Debit card, Mortgage, Investment, Race, Spending, Services
		Financial Propensity	819	Credit/Debit Card, Account, Assets, Bank, Bill, Channel, Check, Spending, Other card, Insurance, Investment, Mortgage, Offer, Service, Specification, Tax
		Attitude	32	Bank, Tax, Service, Investment, Economy, Financial publications
Technology[2]	193	Technology Behavior	10	Computer, Internet, Services, Other devices
		Technology Propensity	183	Email, Mobile phone, Mobile wallet, Service, Smart home, Channel, DVR, Social Media, Specification, Wearables
Transport[2]	354	Travel Behavior	23	

				Business, Vacation, Activity, Mode, Travel purchase
		Travel Propensity	120	Activity, Lodging, Spending, Trip purpose, Channel, Mode, Type, Vacation
		Vehicle Behavior	25	Payment, Vehicle ownership, Vehicle purchase
		Vehicle Propensity	186	Vehicle ownership, Vehicle purchase, Vehicle rent, Loyalty, Payment, Specification, Auto club
Segmentation	172	Lifestyle	75	General segmentation[3], Health, Leisure, Shopping, Sports, Media, Food, Privacy
		Sociodemographic	19	Composition, Occupation, Life event
		Financial	30	Banking, Investment, Insurance, Affordability, Income
		Technology	29	Technographic[4], Technology adoption, Applications, Attitude
		Transport	19	Travel, Vehicle, Attitude

[1]Total number of variables (p) is subject to fluctuation in future stages of this work.
[2]Because technology and transportation are highly-populated categories of interest in this research domain, they are classified separately from the other consumer behavior/propensity variables.
[3]General lifestyle segmentation variables are developed based on demographic, socioeconomic, *and* consumer behaviors and are among the most well-recognized and prototypical TMD variables since they capture many dimensions within one variable (ex. Mosaic, Personicx).
[4]The term technographic refers to general technology segmentation; in fact, the term was initially introduced in the Marketing domain to characterize consumer segmentation based on attitudes, behaviors, and preferences towards technology. In addition, there is an entire lexicon devoted to technology segmentation; for example, Mobirati – representing the generation that cannot imagine life without mobile phones.

5. ACKNOWLEDGEMENTS

This work was jointly sponsored by the Georgia Department of Transportation and the Teaching Old Models New Tricks (TOMNET) University Transportation Center. This study has also profited from discussions with Farzad Alemi, Ali Etezady, Sung Hoo Kim, Yongsung Lee, Aliaksandr Malokin, and other members/visitors of the TOMNET team. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor organizations. This paper does not constitute a standard, specification, or regulation.

6. REFERENCES

Domarchi, C., Tudela, A., & Gonzalez, A. (2008). Effect of attitudes, habit and affective appraisal on mode choice: An application to university workers. *Transportation, 35*(5), 585-599.

Kuhn, M. (2018). Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt (2018) caret: Classification and Regression Training. R package version 6.0-81.

Kuppam, A.R., Pendyala, R.M., & Rahman, S. (1999). Analysis of the role of traveler attitudes and perceptions in explaining mode-choice behavior. *Transportation Research Record, 1676,* 68-76.

Malokin, A., Circella, G., & Mokhtarian, P.L. (under review, available from authors upon request) A Transfer Learning-Based Framework for Enriching National Household Travel Survey Data with Attitudinal Variables.

Mokhtarian, P.L., & Salomon, I. (1997). Modeling the desire to telecommute: The importance of attitudinal factors in behavioral models. *Transportation Research A, 31*(1), 35-50.

Pan, S.J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345-1359.

Stekhoven, D.J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

Stekhoven, D.J. (2013). missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

van der Putten, P., Kok, J.N., & Gupta, A. (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.