Transportation Research Part A 149 (2021) 150–169

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Supplementing transportation data sources with targeted marketing data: Applications, integration, and internal validation

F. Atiyya Shaw *, Xinyi Wang, Patricia L. Mokhtarian, Kari E. Watkins

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

ARTICLE INFO

Keywords:
Consumer data
Targeted marketing data
Travel behavior
Household travel survey
Big data
Third-party data
Travel demand modeling

ABSTRACT

Unlike many third-party data sources, targeted marketing (TM) data constitute holistic datasets, with disaggregate variables – ranging from socioeconomic and demographic characteristics to attitudes, propensities, and behaviors – available for most individuals in the population. These qualities, along with ease of accessibility and relatively low acquisition costs, make TM data an attractive source for the supplementation of traditional transportation survey data, which are facing growing threats to quality. This paper develops a typology demonstrating ways in which TM data can aid in the design of transport studies, as well as in the augmentation of modeling efforts and policy scenarios, allowing for improved understanding and forecasting of travel-related attributes. However, challenges associated with integrating, validating, and understanding TM variables have resulted in only a few transportation studies that have used these data thus far. In this paper, we provide a transportation discipline-specific resource for TM data, informed by our integration of an extensive TM database with both the National Household Travel Survey (Georgia subset) and a statewide travel behavior survey conducted in Georgia on behalf of the Georgia Department of Transportation. Using the resultant datasets, we validate TM data by means of several approaches, and find that the TM dataset reports gender, age, tenure, race, marital status, and household size with match rates ranging from 70% to 90% relative to both transportation surveys. However, we also identify biases in favor of population segments that may have more longstanding financial/transactional records (e.g., males, homeowners, non-minorities, and older individuals), biases comparable but not identical to those of survey data. While this work suggests wide-ranging implications for the use of TM data in transportation, we caution that flexible and responsible approaches to using these data are critical for staying abreast of evolving privacy regulations that govern third-party data sources such as these.

1. Introduction

Declining travel survey response rates coupled with the rapid proliferation of big data have created fertile ground for the exploration of novel third-party data sources to support transportation supply and demand modeling applications. Most prolific have been the use of mobile phone location data to supplement traditional travel diary data, but a wide range of sources, from social media to smart cards, have been effectively used to provide/augment key transport model inputs (Chen, Ma, Susilo, Liu, & Wang, 2016; He,

* Corresponding author.
E-mail addresses: atiyya@gatech.edu (F.A. Shaw), xinyi.wang@gatech.edu (X. Wang), patmokh@gatech.edu (P.L. Mokhtarian), kari.watkins@gatech.edu (K.E. Watkins).

https://doi.org/10.1016/j.tra.2021.04.021
Received 19 November 2019; Received in revised form 1 April 2021; Accepted 30 April 2021
Available online 25 May 2021

Miller, & Scott, 2018; Khan, Ngo, Morris, Dey, & Zhou, 2017; Ma, Li, Yuan, & Bauer, 2013; Ruiz, Mars, Arroyo, & Serna, 2016; Toole et al., 2015; F. Wang & Chen, 2018; Z. Wang, He, & Leung, 2018; Welch & Widita, 2019). These successes make clear that transportation planners, engineers, and researchers must continue to explore effective approaches to utilizing nontraditional data sources in transport modeling and forecasting efforts. However, the sources utilized thus far have tended to entail siloed data that lack linkages to socioeconomic and demographic (SED) indicators, psychometric attributes (e.g., attitudes), and behaviors across different domains. In contrast, targeted marketing (TM) data are largely untapped, low-cost, holistic databases that house hundreds to thousands of diverse variables on individuals and households across the country.

TM data are typically used to identify and market to individuals likely to be more receptive to a particular product/brand, but due to attributes such as data magnitude and ever-increasing variable richness (supported by continuous technological advances), there is enormous potential in using these data to supplement travel demand modeling and forecasting efforts that currently primarily depend on actively collected survey data. Transportation surveys such as household travel surveys (HHTS) and research-oriented stated and revealed preference surveys are infrequent, expensive, and suffer from continuously declining response rates that can threaten the validity of using these sources independently (PTV NuStats, 2011, National Research Council, 2013). On the other hand, while TM data are available and relatively inexpensive, challenges associated with integrating TM data with transportation survey data, validating acquired TM variables, and further interpreting these variables have meant that only a few transportation studies have successfully used these data. As such, the purpose of this paper is to provide a discipline-specific resource that details potential applications and impacts of TM data in transportation, and provides tools to address some of the barriers that currently hinder the use of TM data.

The remainder of this paper is organized as follows. We begin by examining sources that inform the creation of TM data, and detail some benefits and challenges associated with utilizing these data (Section 2). Next, we review how TM data have thus far been used in the transport domain, and develop a taxonomy of possible transportation applications, outcomes, and research directions that could benefit from the use of TM data (Section 3). Based on the team's integration of a large TM dataset with statewide and national transportation surveys (Section 4.1), we then present a framework for the integration of TM data with existing transportation data sources (framework summarized in Section 4.2, and further detailed in Appendix B). Using the integrated dataset, we examine the quality of TM data relative to comparable self-reported data from travel surveys, and also examine biases of TM data by comparing survey respondents with and without records in the TM database (i.e., we conduct an evaluation of TM data veracity – see Section 5). We then discuss the findings detailed in Section 5, and provide recommendations for how transportation professionals can address TM data biases identified in the paper (Section 6). We close with a summary of contributions and findings (Section 7). Appendix A provides supplementary tables and figures, while Appendix B provides additional data integration details for analysts seeking to enrich their own survey datasets with TM data (Appendices A and B are in an online supplemental file).

2. Exploring targeted marketing data

Since TM data have been little-used by the transportation community as yet, we begin the discussion with a general introduction to this type of data. In this section, we examine TM-related terms and data sources, followed by a summary of benefits and disadvantages of which transportation researchers/ practitioners should be aware when using TM data.

2.1. Defining targeted marketing data

The terms consumer, audience, and/or (targeted) marketing data are often used interchangeably; however, they can refer to different concepts. In this work, we use the term "Targeted Marketing (TM) data", and explain why we make this distinction by presenting a brief overview of the related terms here:

- **Consumer data** are defined by Birkin (2019) as "data arising from the interaction between customers and service providers", and should be the byproduct of a "market-based exchange of value". The most common form of consumer data is transactional data, which are obtained each time a consumer utilizes a credit or debit card to make a purchase. These data are then typically aggregated to yield variables such as the number of purchases made within various consumer categories (e.g., apparel, home, etc.), frequency of purchase, and the medium used for transactions (e.g., online, in-store, etc.). Consumer data may also include less traditional, technologically-enabled transactional interactions such as mobile application use, digital browsing, and smart card usage for fare payment.
- **Targeted marketing data** refers to large databases that house hundreds to thousands of individual- and household-level variables that data providers (often these are credit reporting firms) either directly collect, purchase, or develop. TM data are developed with the explicit purpose of being re-sold to businesses who use selected variables to aid in marketing campaigns that target their specific audience. In some contexts, TM firms use the term "consumer data" to indicate that TM variables represent profiles of *consumers* in the marketplace. As such, the term "TM data" is often conflated with the term "consumer data"; however, while TM databases often include many variables that are derived from consumer data, they also include other types of variables/data.
- **Audience data** is a term used by marketers/business strategists to represent variables that are specific to a business's target base of consumers, i.e., its *audience*. Business entities may select from already developed audience segments present in TM databases, or alternately, may request TM providers to develop personalized segments that are relevant to their services. Thus, audience data/ segments can be derived from TM databases, although businesses also often collect their own internal audience data.

As can be seen, there is significant overlap between these terms. We recommend the use of the term "TM data" for datasets purchased from TM and/or credit reporting firms or other large third-party data providers/ compilers, as it is likely that many of the variables in such databases have been developed and/or imputed based on a host of other variables. For example, while a variable description may suggest that a variable is a "pure" consumer variable (i.e., directly collected by a service provider), it is likely that this variable was modified using information from other sources (e.g., from public records or survey data) in the TM database, and thus the use of the term "TM data" aids in clarifying the source of the variables being used.

Fig. 1 provides a non-exhaustive organizational structure for sources that typically inform TM databases. Shown first is the most established source, that of administrative data such as births, deaths, and property ownership captured in public records, or birth dates and address information captured in customer records (Connelly, Playford, Gayle, & Dibben, 2016). The next most entrenched/ longstanding form of TM data is consumer data that can be obtained from a wide range of transactional records, such as purchase details, loyalty cards, and product/service usage (Birkin, 2019). In recent developments, some TM databases are integrating digital data that track individuals' online browsing patterns and access. Relatedly, another form of online data is derived from social network platforms, and may include information ranging from contact networks to taste preferences regarding movies, news content, music, etc.

In addition to these passive data sources, TM data may also include active data sources from surveys that are typically conducted by consumer research firms, but which can also come from individuals' responses to online quizzes/games/ questionnaires. For clarity, we note here that to qualify as active data, the individual must choose to relay the information being obtained, while with passive data, the individual may not even be aware that information is being collected. TM databases often comprise information from both active and passive data sources, a characteristic differentiating them from traditional third-party and/or big data, which are typically entirely derived from passive data sources. The TM variables that are derived from active data sources like surveys typically include individuals' preferences and opinions toward specific products and/or services (e.g., the importance of post-purchase customer service in selecting a specific type of service), but can also include more general preferences. Examples of the types of variables present in TM databases can be found in Section 4.1.3 and Table A2 of online Appendix A.

2.2. *Benefits of targeted marketing data*

The most significant benefits of TM data within a transportation context are the volume and disaggregate nature of the data. TM datasets are extremely large because they are available for almost all individuals/households in the population, allowing for the possibility of using TM data to enrich other data sources at a disaggregate level for most individuals in a typical transportation study. This contrasts with most publicly available data (e.g., Census and American Community Survey), which are commonly used for transportation data validation, but which report only aggregate-level cross tabulations (e.g., block groups, census tracts) or are available only for a small fraction of the population (e.g., the Public Use Microdata Sample). The resulting potential magnitude of TM data may also facilitate the use of efficient artificial intelligence approaches for researchers and practitioners interested in using these methods.

Further, in contrast to traditional large-scale household travel surveys which tend to occur once every 10 years or so, TM data are dynamic, meaning that the values of many TM variables are updated on a monthly, quarterly, or yearly basis. In addition, TM data comprise a range of diverse variables, many of which are not available through traditional or novel data sources currently used in transportation. Thus, the primary overall benefits of TM data lie in the overall magnitude/size of the data, the diversity/richness of TM variables, and the rapidity of TM data generation and renewal (Erevelles et al., 2016; Sivarajah, Kamal, Irani, & Weerakkody, 2017). These three attributes are respectively known as volume, variety, and velocity, and also happen to be considered the original three defining attributes of big data (Laney, 2001). "Value" and "veracity" were added later, with these five attributes collectively being known as the "5 Vs" commonly used to characterize and evaluate big data (although we note that one could find lists ranging from seven to forty-two Vs that are used in various contexts to further describe big data; Sivarajah et al., 2017). However, while TM

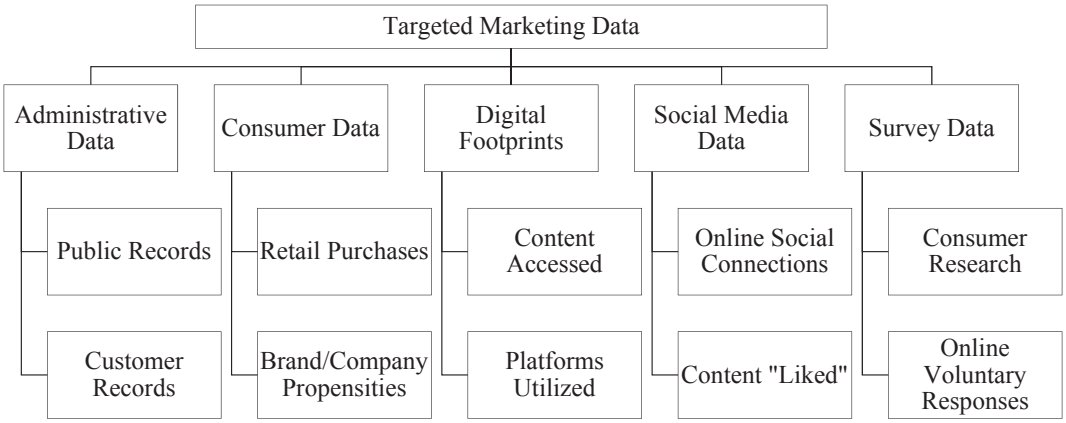

Fig. 1. Examples of Targeted Marketing Data Sources.

databases have increased volume, variety, and velocity relative to transportation survey data, they are generally smaller in size and are generated more slowly than traditional big data, which tend to be purely passive data such as second-by-second GPS location traces. Nonetheless, relative to traditional transportation survey data, TM data can be considered to meet the loosely defined and broadly applied definition of big data (Macfarlane, 2014).

Regarding utilitarian benefits, TM data are inexpensive and easily accessible relative to traditional survey data collection. In recent projects executed by our team, TM data cost approximately US$1.50 per person, while a statewide transportation survey that obtained rich attitudinal and behavioral variables (see Section 4.1.1) cost an estimated US$20 per person. Further, the purchased TM dataset contained 5583 variables, while the transportation survey dataset contained 200 – 400 unique variables (using varied coding techniques). The overall cost of the transportation survey was ~US$65,000, while the overall cost of purchasing the TM dataset for more than three times as many respondents (~10,000 cases) as were contained in the survey final sample (~3,000 cases) was ~ US$15,000, not counting graduate student/faculty time costs for either. At a household level, Kressner and Garrow (2014) reported that in their estimates, the cost of obtaining a completed travel survey for one household in Atlanta is around US$200, relative to five cents for obtaining a set of TM variables for that household (the study did not detail *how many* TM variables were obtained). Finally, a significant benefit for transportation professionals is that TM data have widespread availability, meaning that any entity, from academic researchers to governmental agencies, could purchase TM data from marketing firms (after agreeing to legally mandated privacy restrictions). This accessibility means that if TM data are shown to improve modeling/forecasting, transportation agencies can feasibly acquire TM data and integrate them into their operations; however, as will be discussed in more detail next, with increased privacy restrictions, this availability may be moderated in the future.

2.3. Challenges of targeted marketing data

Before TM data can achieve widespread utilization in transportation, it is important to assess the veracity (i.e., accuracy) and value (i.e., worth/usefulness) of these data within the context of intended applications (Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011; Lovelace, Birkin, Cross, & Clarke, 2016; Lukoianova & Rubin, 2014; Sivarajah et al., 2017). In the literature, *veracity* is typically evaluated through comparing TM data against benchmarks from other sources such as the Census or other surveys – termed internal validation in this paper. Such comparisons assess both the accuracy of the data that are *included*, (i.e., evaluate *observational* errors, and particularly biases) and the extent and nature of the biases generated by *excluded* data (*nonobservational* biases) – since, as with all data sources, TM data have inherent biases that may disproportionately affect underrepresented/vulnerable populations, among others. Assessing the *value* of the TM data can be achieved by integrating the TM data into transportation applications and observing the model predictions or performance with the new data source(s) – termed external validation in this paper. While a handful of studies have shown the value of TM data in transportation (detailed in Section 3), to date, the authors are aware of only three studies that have sought to examine the veracity of TM data from a transport perspective (Kressner & Garrow, 2014; Kressner, Carragher, & Watkins, 2014; Lovelace et al., 2016). This may be partially due to challenges associated with integrating TM variables with traditional travel datasets, namely that names and addresses are needed to obtain TM data for individual-level validations; however this does not restrict aggregate level validations, which are similarly rare.

To begin the process of mitigating these challenges, this paper provides a guide to integrating TM data with existing transportation datasets (Section 4), and further presents a veracity assessment for a specific TM dataset. This assessment involves both an individual/household-level pairwise validation through comparison with survey data from (theoretically) the same individuals (Section 5.1), and an examination of TM data biases and representativeness (Sections 5.2 and 5.3). Further, Section 6 provides a brief discussion of methods for ameliorating dataset biases that may be useful in the specific context of the TM data being examined in this study. Given the significant length of the present paper in aiming to achieve the afore-stated goals, we reserve an assessment of value for a separate study of its own (Shaw, 2021; Shaw, Wang, Mokhtarian, & Watkins, in-progress; Shaw & Mokhtarian, in-progress).

A second set of challenges in working with TM data lies in the development of the variables. TM providers often use proprietary algorithms to develop, impute, and/or model many variables, not only making it difficult to evaluate the robustness of TM variables, but further clouding the interpretation of these variables if they are to be used in transport models. We emphasize here that this constitutes a significant disadvantage of third-party data sources like TM data relative to first or second-party data that are often more transparent regarding variable development procedures. In addition, modeled TM variables may be relatively unstable as the algorithms may be tweaked over time, thus precluding consistent definitions of the variables. Furthermore, variables themselves may become obsolete as the data sources used to inform the TM databases ebb and flow, in part in response to the commercial demand for the associated information. Moreover, variables are both measured on different time frames and updated on a schedule that differs across variables and which may not be transparent to the user. For example, a variable indicating whether the individual has purchased a car within the past 12 months may have been last updated 11 months ago (and therefore be almost a year out of date), while a variable indicating whether the individual has had food delivered to the house within the past month may have been last updated six months ago. Our team investigated the purchase of TM data for a study of how consumer behavior changed during the Covid-19 pandemic, but ultimately concluded that we did not have precise enough information about the dates to which key variables pertained to be confident in such an analysis.

Nonetheless, such issues are present in most external data sources, as we see variable definitions and included variables changing even across national data sources such as the U.S. Decennial Census and National Household Travel Survey. Moreover, these challenges do not detract from the richness of the information that TM data have to offer, and in reality, there are numerous consistent TM variables that users can rely on while avoiding variables that may be unclear or unstable. Furthermore, as with most big data, when methodologies like machine learning are used, the stability and interpretation of variables are arguably less important than their

contribution to an overall improvement in forecasting that facilitates more accurate decision making. In Section 3.1 we show that in post-model development, TM can be used to develop policy scenarios, thereby compensating for the reduced interpretability of some variables in model development.

From another perspective, the quantity and richness of TM variables provide an added challenge. Specifically, since TM data come from a large array of sources, there may be reduced consistency in data scales and definitions *across variables*, as our research team experienced. As such, users may have to spend additional time processing the received data, and in some cases, building their own data dictionaries. Thus, as acquired data become increasingly voluminous and diverse, the potential to obtain value is moderated by the available physical, human, and organizational capital (Sivarajah et al., 2017). We note as well that since TM data are collected and aggregated for marketing purposes, the resultant databases do not contain the same breadth of general and transport-related preferences and opinions that can be obtained using transportation survey data. We believe that the challenges discussed here are likely some of the major reasons slowing the use of TM in transportation, and we hope that this work, in combination with additional efforts from other TM data users, will serve to introduce the requisite outlook and approaches needed to overcome these challenges.

The final group of challenges for using TM data comprises evolving privacy regulations and concerns that are increasingly salient to researchers, regulatory agencies, and the public. The European Union (EU) General Data Protection Regulation (GDPR), introduced in 2018, represents the strictest data protection law in the world to date. Even though the U.S. as a whole is currently far from this level of regulation, some states have expressed interest in emulating the E.U., such as California, which instituted the California Consumer Privacy Act (CCPA) at the start of 2020. While the specifics of these laws are complex, the most relevant detail in the context of this paper is that both laws aim to provide consumers with the ability to opt out of the collection and sharing of their personal information, and/or to edit the consumer records that are available to them. In practice, this means that TM data providers will still retain existing databases identical to those described in Section 2.1; however as mentioned, consumers can request corrections or deletions made to their records that are present in those databases (Acxiom, 2020). At this point it remains to be examined how this new provision will affect TM databases for European and Californian consumers in the future, a point that of course rests on how many consumers take advantage of the policies to remove/edit their records in the database. Future research should seek to explore the changes that have occurred in the databases as a result of new privacy laws.

Thus, overall, from a data availability standpoint, evolving privacy regulations may threaten the stability and reliability of TM data for long-term transportation applications, particularly those that require TM records to be matched at a disaggregate level. Despite these complications, third-party data such as TM data are expected to continue to be critical supplementary data sources for a wide array of fields, and as such, this paper aims to provide a stimulus for transportation professionals to explore compliant and ethical approaches to using these diverse data sources to improve transportation modeling and forecasting efforts. One potential solution may lie in the use of data agencies that can serve as intermediaries between data providers and researchers, thus ensuring that the data provided to individual research teams has been appropriately processed to prevent any potential privacy incursions (examples of agencies that could/ already serve this purpose are the United Kingdom Administrative Data Research Network, the Consumer Data Research Centre, and the University of Washington Transportation Data Collaborative). Regardless of how the data are acquired, we recommend that analysts meet with appropriate institutional research ethics personnel prior to beginning any project that uses third-party data sources, and once the data have been acquired, to work toward timely de-identification of the datasets being used.

3. Using targeted marketing data

TM data can be used in each stage of travel demand modeling and forecasting efforts, beginning with survey design and sampling, extending to model prediction and accuracy, and even having implications for result interpretation and application (see typology in

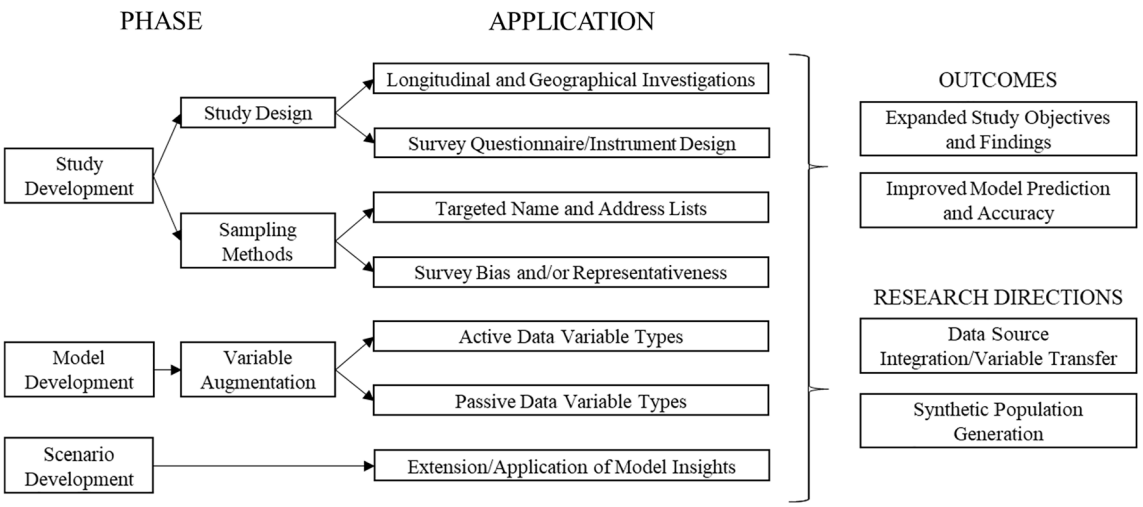

Fig. 2. A Typology of TM Applications in Transportation.

Fig. 2). In this section, we show that the outcomes and research directions associated with TM data in transportation are non-trivial, and have the potential to significantly improve transport planning in the future. Where appropriate, we cite examples of known TM applications in the transportation literature.

3.1. Transport applications and outcomes

TM data have significant implications for shaping transport study development, from both study design and sampling perspectives. The dynamic nature of TM data is one of its most significant benefits with regard to study design, as this allows for ease of data collection at multiple time points. Documented changes in individual/household characteristics over time may allow for improved understanding and forecasting of how these changes influence travel behaviors. Macfarlane et al. illustrated the benefit of the longitudinal nature of TM data by using address histories from TM data to examine how prior places of residence could influence vehicle ownership, a study objective that would not be possible with traditional cross-sectional survey data (Macfarlane, Garrow, and Mokhtarian, 2015). Birkin (2019) later similarly suggested that consumer data (in this case, from online real estate agents) is unique in providing the level of spatial detail (i.e. origins and destinations) and rapid updating necessary for the study of geodemographic mobility, a key transport geography study objective that previously required the use of longitudinal data. In the same way that TM data are present across time, they are also available across regions, facilitating geographic and land use comparisons for transportation attitudes and behaviors, and further, providing the ability to validate/ segment models along those lines.

From a survey instrument design perspective, the presence of TM data for respondents being sampled could theoretically aid in the reduction of the number of questions necessary on travel behavior surveys, thus resulting in shorter surveys and thereby, potentially improved survey response rates. Alternatively, if some variables are able to be reliably sourced from TM data, then the corresponding survey questions may be *replaced* with other questions, yielding a richer set of variables for use in model building, and thus potentially improving model predictions.

TM variables can also aid sampling efforts for traditional transportation data collection. TM databases are already widely used to obtain names and addresses for use in travel behavior survey sampling (e.g., Handy, Cao, & Mokhtarian, 2005; Kressner et al. 2014). Building on that, the SED characteristics present in TM data (e.g., gender, income) can allow analysts to sample socioeconomic and demographic (SED) groups of interest in greater proportions relative to other groups. For example, it is known that individuals in certain SED groups have lower or higher response rates relative to other groups, and respondent information based on the TM variables could aid in over-/under-sampling as appropriate. TM data could also be used to examine survey biases and representativeness by providing an additional source of information that could be compared with traditional survey data, although we note that traditional survey data and TM data will each have their own inherent biases, a point further examined in Section 5.

TM data have perhaps the greatest potential to benefit transportation model development through the augmentation of transportation datasets with variables that are not possible to obtain through traditional transportation surveys (i.e., passive data), as well as variables that are obtained through active data collection methods. As a result, given appropriate prior hypotheses, TM data can facilitate the testing of a larger range of variables in predictive models, leading to enhanced conceptual understanding of travel-related attributes, as well as potential improvements to model performance. Some transport studies have already shown that unique TM variables can improve model accuracy; for example, Kressner showed that using TM lifestyle segments improves prediction for air passenger trip models and residential location choice models (Kressner and Garrow, 2012; Kressner, 2014). In general, transport researchers have long studied the connection between lifestyle and travel behavior, showing that lifestyle segment stratification can improve the predictive accuracy of transport behavioral models (Kitamura, 2009; Salomon and Ben-Akiva, 1983; Van Acker, Goodwin, & Witlox, 2016). This bodes well for the value of TM data, given that TM databases are best known for having robust lifestyle, financial, and technology-based segmentation variables. In addition to serving as explanatory variables, in some cases active and passive TM variables may also be of interest as dependent variables (i.e., variables to be modeled in their own right) within larger travel demand models/systems. Furthermore, the distributions or frequencies of TM variables of interest may also be used to aid in model development, serving to provide marginal distributions or probabilities that could potentially aid in model calibration (although of course it will be critical to first ensure representativeness of the data being used – see Sections 5 and 6 for more on TM data representativeness).

On the other hand, Binder, Macfarlane, Garrow, and Bierlaire (2014) showed that TM variables typically obtained through survey data collection, such as ethnicity, income, gender, and age, are able to support residential location choice models without depending on HHTSs. This significant finding could allow researchers not only to shorten their surveys, but also to remove more sensitive questions (e.g., income) from survey instruments, both actions which could allay some of the factors contributing to declining survey response rates. In a similar example, Macfarlane, Garrow, and Moreno-Cruz (2015) used SED traits and home prices derived from TM data to model willingness to pay for proximity to public transit. In addition to these examples in the literature, many regional transportation planning agencies also currently obtain employment statistics (for use in their regional models) from business list data acquired through TM firms. Overall, the outcomes possible from augmenting traditionally available travel datasets with TM data offer significant implications for the field, and it is for this reason that Section 4 of this paper provides a generalized framework that can aid in pursuing this application.

While the foregoing discussion has pointed to the exciting *potential* of TM data for the replacement and/or augmentation of variables drawn from traditional transportation data sources, such as those of surveys, our view, given the challenges discussed in Section 2.3 and the internal validation shown in Section 5, is that the current form and state-of-knowledge regarding the accuracy of TM data preclude them from being a suitable *replacement* for conventionally-sourced data. More realistically, as specified in the typology, at least for now, TM data should be looked upon as a source of variable *augmentation* for transportation data sources. Furthermore, as

discussed in the directly preceding paragraphs to this one, the TM variables used to augment datasets in the literature thus far have been SED variables as well as TM data's famous segmentation variables (which are often based on clustering algorithms that combine hundreds of variables). Accordingly, there remains significant work to be done on better understanding which TM variables can be used for data augmentation (i.e., which variables have sufficient *veracity*, as assessed through internal validation), and subsequently whether these variables bring *value* to the transportation applications for which they are used (as assessed through external validation). These comments also apply to the transportation variables present in TM data, examples of which are provided for the specific TM dataset used in this paper in Section 4.1.3 and Table A2 in online Appendix A.

Lastly, TM data are ripe for use in the development and testing of policy scenarios, applications that can expand the insights gleaned from analyses, while potentially clarifying decision-making based on transport study findings. Specifically, TM data can facilitate the post-hoc application of models and/or proposed policies to various segments of the population, allowing for an understanding of how proposed scenarios may affect individuals, demographic groups, overall transport choices, and infrastructure operations. Furthermore, TM data can be purchased for this purpose even after the completion of a study, thus making TM integration at this stage more accessible. In one example from the literature, Binder et al. (2014) used data derived from TM records to examine the effects of three proposed emissions policy scenarios on various SED groups, finding that the suggested and commonly used strategies for reducing the cost of indiscriminate emission testing are inequitable and/or ineffective, and suggesting that other transportation policy tools may be needed to address the issue.

3.2. Transport research directions

The preceding section highlighted the potential for TM data to expand transportation study objectives and improve model predictions. Beyond these outcomes, there are many transportation research directions that could benefit from the use of TM data. Two such examples involve the use of methodological tools like machine learning and discrete event simulation to aid in: (1) the integration of multiple data sources through variable transfer; and (2) the generation of synthetic populations based on disaggregate TM data.

The first initiative is being concurrently developed by the authors of this paper, using the integrated dataset described in Section 4. In this effort, a range of algorithms are trained using an integrated dataset that combines statewide and nationwide transportation surveys with TM data appended at an individual/ household level. High performing algorithms that are able to predict selected variables (e.g., attitudes) may facilitate the transfer of variables that are unique to one data source into a recipient data source. This approach paves the way for data source linkage, with TM data operating as the "glue" (i.e., "common" variables/features) that links disparate sources together, and facilitates variable transfer. This approach may enable the development of richer, more up-to-date datasets that can improve travel demand modeling efforts.

The second group of initiatives entails the use of disaggregate TM data to generate synthetic populations that can yield insights into how individuals in a region travel (Beckman, Baggerly, & McKay, 1996; Birkin, Morris, Birkin, & Lovelace, 2017; Kressner, Macfarlane, Huntsinger, & Donnelly, 2016; Kressner, 2017). The use of disaggregate TM data to provide a nearly-complete enumeration of household and individual-level SED traits may represent an improvement over the 1% or 5% anonymized sample offered by the American Community Survey Public Use Microdata Sample (ACS PUMS), which is currently the primary source of SED inputs for population generation in transportation. Kressner has implemented this idea at a large scale, using TM data to provide disaggregate SED data that is then fused with mobile phone location data to create synthetic travel diary records (2017). This concept has been successfully validated for several cities in the U.S. (Kressner et al., 2016). Along similar lines, researchers in Europe simulated demographics that would match Census data for a city, and then matched travel-related consumer data to these simulated individuals on the basis of age, gender, family status, and social group (Birkin, Morris, Birkin, & Lovelace, 2017).

Thus, while the first research initiative detailed here uses TM data to allow variable transfer across data sources, the second approach uses it to synthesize populations, and study how these synthetic populations travel. Both approaches highlight the importance of integrating passive and active data sources to build and validate disaggregate/aggregate travel demand modeling systems, a tactic that can help take travel demand modeling into the next generation by reducing the reliance on traditional data sources. While the ultimate effectiveness of these approaches, and possible symbiosis of methods, remains to be seen, we believe that there is substantial potential not only in these methods, but also in future approaches that can use TM data to make similarly ambitious attempts to move the field forward.

4. Integrating targeted marketing data with transport datasets

As discussed in Section 2.3, to examine the value and veracity of TM data for use in transport applications, TM data must first be integrated with transportation survey datasets. However, the integration of TM data with other data sources can pose technical and methodological challenges. As a result, in the following subsections, we provide an overview of the datasets used in this study (Section 4.1), followed by a discussion of the process used to integrate TM data with transport survey datasets (Section 4.2, with additional details in online Appendix B).

4.1. Overview of data used in study

For this investigation, we purchased TM data for respondents to: (1) a statewide transportation survey conducted by our research team for the Georgia Department of Transportation (GDOT survey) and (2) the Georgia subsample of the U.S. National Household Travel Survey (NHTS), a nationwide travel behavior-focused survey conducted by the U.S. Department of Transportation. The

following subsections (4.1.1 to 4.1.3) provide details on the data sources in this study: GDOT survey, NHTS, and TM datasets.

4.1.1. Georgia Department of Transportation survey

The GDOT survey (conducted September 2017 to January 2018), is a statewide research-oriented transportation survey that obtained 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. 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 in Fall 2017 from a different TM data provider than the one used for the purchase discussed in Section 4.1.3), and (b) ~5000 individuals who responded to the NHTS and agreed to be contacted for a follow up survey.

Approximately 1800 of the randomly sampled 30,000 respondents returned a completed (usable) GDOT survey (termed the ***GDOT_R*** subset in this paper), and about 1500 of the ~5000 NHTS respondents sampled returned a usable GDOT survey (termed ***NHTS_Agree_R***, for "Agreed to be contacted again, and Responded to the subsequent GDOT survey contact"). Thus, roughly 3300 valid respondents were retained in the GDOT dataset, and TM data enrichment was initiated across all respondents. See Fig. 3 for a visual representation of the GDOT and NHTS sample subsets used in this paper, and see Table A1 in online Appendix A for descriptive statistics on the GDOT sample. For additional details on the survey, please see Kim, Mokhtarian, and Circella (2019).

4.1.2. National Household Travel Survey

The NHTS is a repeated cross-sectional travel behavior survey conducted by the Federal Highway Administration, and deemed the "authoritative source on travel behavior of the American public" (Federal Highway Administration, 2018). 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 (for a chosen travel day) and health. Additional details regarding the NHTS can be accessed at https://nhts.ornl.gov/documentation.

As mentioned previously, approximately 5000 respondents from the Georgia subsample of the NHTS agreed to be contacted again for a follow up survey, and these respondents received a GDOT survey several months after completing the NHTS. Of these, ~1500 usable returns (the ***NHTS_Agree_R*** subsample) represent respondents for whom we have *both* GDOT and NHTS data (i.e. an overlapped sample). The remainder of the 5000 respondents represent individuals who agreed to be contacted again, and thus received a copy of the GDOT survey, but did not respond to it (***NHTS_Agree_DNR***, Agreed but Did Not Respond to the subsequent GDOT survey contact). Of the total NHTS Georgia subsample, ~3500 respondents indicated that they did not want to be contacted again, and as such did not provide shareable name and address information (***NHTS_DNAgree,*** Did Not Agree to be contacted again for a follow-up survey). Thus, these three NHTS subsets along with the GDOT-only subset (***GDOT_R***) represent **four** distinct subsets of respondents that comprise the transportation survey datasets used in this project (see Fig. 3 for a schematic depiction of the subsets and Table A1 for SED characteristics).

TM data were purchased across all respondent subsets; however, due to differences in the available name/address information, the TM enrichment process was necessarily different across subsets. We delineate the subsets in detail here because they allow for the illustration of different approaches to TM integration for analysts with differing types of name/address information present in their datasets. The "common variables" shown in Fig. 3 represent variables that were obtained by both the NHTS and GDOT surveys, and a sample of these selected for TM data validation are shown in Table A3. Note that since TM data enrichment occurred across all possible records in both surveys, TM variables constitute *external* "common variables", greatly expanding the potential for future analyses, a point discussed in Section 3.2.

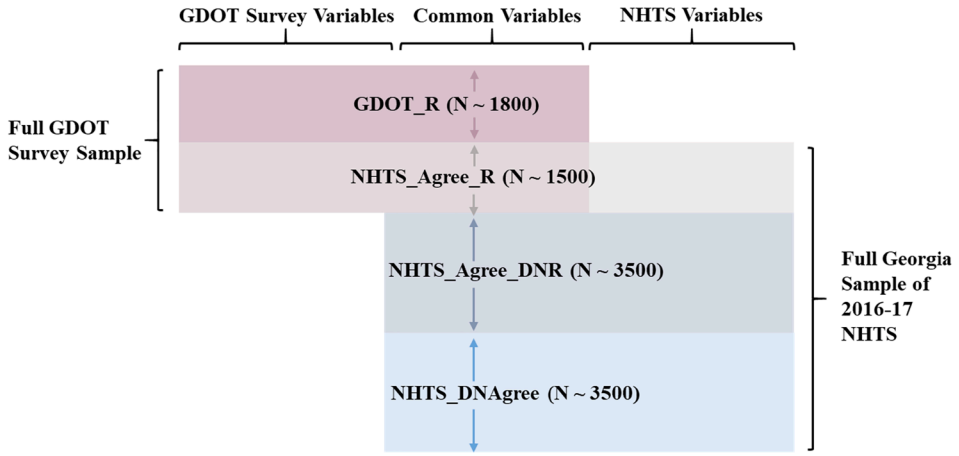

Fig. 3. Schematic Representation of GDOT and NHTS Data Subsets.

4.1.3. Targeted marketing dataset

The TM data purchased for use in this study were obtained from a large U.S.-based TM data provider that is an industry leader in data quality, and which is used by many business entities for their marketing needs. Selection of the provider used in this study hinged on the firm's ability to provide a rich array of variables for the smaller sample size (~10,000 cases) and nontraditional (exploratory, research-based) data needs of this project. In addition to the TM firm's natively collected/derived variables, their database also houses supplementary variables purchased from well-known firms such as Claritas, SEMcasting, etc. At the time of our acquisition, the firm's database contained $p \approx 5500$ variables ('p' is used to represent number of variables throughout this paper), all of which were purchased for this study.

Of the total variables available, approximately 1500 represent a general variable set from which most marketers (i.e. typical clients for TM firms) select when purchasing data augmentation services. The additional ~ 4000 variables are termed audience propensity variables, and are developed on contract to be sold to certain corporations, and thus might be updated/changed on a monthly basis. The general variables have no name release restrictions, meaning that the full names can be shared publicly, while the audience propensity variables required a legally binding non-disclosure agreement barring disclosure even of these variables' names. Further, to obtain the full set of all variables, we provided an official statement of use followed by the completion of additional legal paperwork on 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 *large* TM variable set is a non-trivial undertaking that can require months of discussion prior to final approval and variable transmission.

The acquired TM dataset comprises continuous, ordinal, and nominal (dichotomous and polytomous) variables. In Fig. 4 and Table A2 of online Appendix A, we classify the initial received variables (after removing variables that were completely missing, as well as *meta*-data variables like precision levels) into the following topical areas: sociodemographic, land use, attitudes, lifestyle, financial, technology, and transportation. Fig. 4 summarizes the overall variable distribution, and Table A2 further summarizes the variable classification distribution across the TM variables. Given the traditional TM sources of credit card and shopping records, it is intuitive that 61% of the received TM variables are consumer-related variables such as purchase behavior, while 18% are financial variables related to investment, income, and insurance, among others. Examples of transportation variables obtained include business and vacation travel behaviors, vehicle ownership (i.e., brands/vehicle type), vehicle payment type, and brand propensities regarding rental car companies and airlines.

4.2. Targeted marketing data integration framework

In Fig. 5, we summarize the process of acquiring, processing, and integrating TM data with existing transportation survey datasets (bolded elements depict the process used in this study). In this section, we provide only a brief overview, but point interested readers to online Appendix B, where we provide an in-depth guide with step-by-step detail on the integration process used for the datasets in this paper.

There are four primary services of interest offered by TM data providers, and transportation professionals may be interested in any of these services for varying applications (see Section 3). In this section, we discuss *data enrichment* only, as it is the service used to append a range of TM variables to existing records (see Section 10.2.1 of online Appendix B for a discussion of all data services). To use this service, analysts should first determine the quantity and types of TM variables intended to be appended to each record. For a small

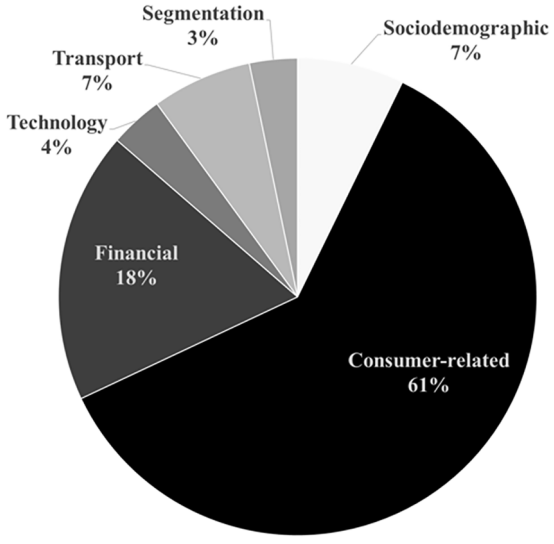

Fig. 4. Overview of Variable Types in TM Dataset (p = 5684).

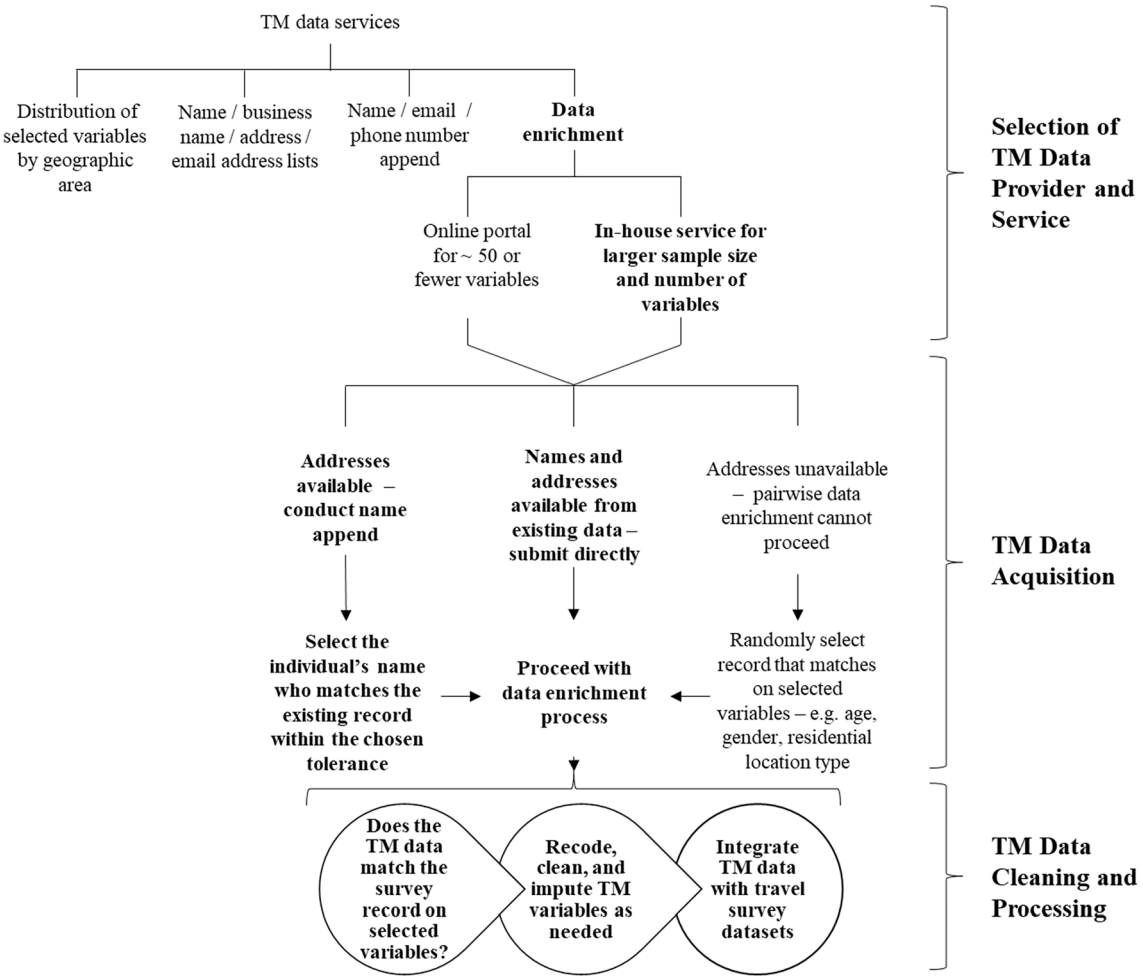

Fig. 5. Simplified Overview of TM Data Integration Process.

number of variables (i.e., 50 – 100), TM providers often have online portals that can be used to quickly and easily append variables. As the number of variables and/or respondents grows, data enrichment must proceed through in-house services that require additional legal paperwork and time.

TM providers typically require names *and* addresses for all cases that are being submitted for TM data enrichment. Submitted lists are matched against names and addresses on file in the TM provider's database, and if the exact first and last name cannot be matched, variable matches degenerate into less precise matches (e.g., address and last name, address only, zip + 4 code, zip code – with each of these successively identifying a larger, less precise area where for example, zip + 4 code may refer to a specific part of a street or a building while zip code may refer to a general area and/or associated mail delivery office). Since transportation practitioners may have varying amounts of name/address information available for their survey datasets, in Section 10.2.2 we demonstrate how we dealt with the four survey data subsets in this study (Fig. 3), as each had differing amounts/types of name/address information available.

Following data acquisition, the resulting TM dataset typically requires substantial cleaning, recoding, and processing before integration with survey datasets. The most critical step entails the individual-level comparison of the TM record for each case to the available survey data. Analysts must first select the variables that will be compared between the TM and survey data, and subsequently should establish the associated tolerance/confidence level for retaining the compared cases given the selected variables. For the dataset in this paper, the variables selected for verification are gender, age, and education level, in order of importance. After processing and retaining cases that are believed to represent the same individual across datasets, TM variables must be recoded (e.g., variable values/levels may need to be made consistent across data sources), cleaned (variables with high levels of missingness or near zero variance may need to be removed or otherwise addressed), and imputed as necessary.

As before noted, please reference online Appendix B (Section 10.2) for expanded guidance on selecting a TM provider and service, successfully acquiring TM data, and cleaning and processing the obtained dataset.

5. Internally validating targeted marketing data

At this point, we have now integrated the TM variables purchased for this study (summarized in Table A2) with two transportation survey datasets (Section 4.1), using the framework and approach outlined in Section 4.2 and further detailed in online Appendix B. Following this data integration step, we develop comparable variable categories for fundamental SED variables present across the TM data and survey datasets (see Table A3). We now turn to internally validating these key TM variables relative to the GDOT survey and NHTS datasets discussed previously (Section 4.1).

Statewide/regional surveys (which include research-oriented surveys like the GDOT survey, as well as regional household travel surveys), in tandem with nationwide data sources like the Census, American Community Survey, and NHTS, represent the core sources

Table 1
Variable Accuracy Rates across TM and Survey Datasets before and after Processing.

Variable	Match	Before Data Processing				After Data Processing			
		TM vs. GDOT N = 3288 [a]		TM vs. NHTS N = 5148 [a, c]		TM vs. GDOT N = 2699 [b]		TM vs. NHTS N = 4027 [b,c]	
		N	%	N	%	N	%	N	%
Gender [f]	Accurate matches [d]	2864	90.86	4455	95.58	2686	100.00	4019	100.00
	Inaccurate matches [d]	288	9.14	206	4.42	0	0.00	0	0.00
	Not comparable [e]	136	–	487	–	13	–	8	–
Age [f]	Accurate matches	2806	90.75	4023	88.87	2610	99.35	3710	95.10
	Inaccurate matches	286	9.25	504	11.13	17	0.65	191	4.90
	Not comparable	196	–	621	–	72	–	126	–
Tenure [g]	Accurate matches	–	–	4168	87.31	–	–	3519	88.06
	Inaccurate matches	–	–	606	12.69	–	–	477	11.94
	Not comparable	–	–	374	–	–	–	31	–
Race	Accurate matches	2441	84.82	3708	84.56	2020	85.59	2967	85.53
	Inaccurate matches	437	15.18	677	15.44	340	14.41	502	14.47
	Not comparable	410	–	763	–	339	–	558	–
Marital status [g]	Accurate matches	2111	72.22	–	–	1807	73.85	–	–
	Inaccurate matches	812	27.78	–	–	640	26.15	–	–
	Not comparable	365	–	–	–	252	–	–	–
Dwelling type [g]	Accurate matches	1635	63.05	–	–	1348	61.89	–	–
	Inaccurate matches	958	36.95	–	–	830	38.11	–	–
	Not comparable	695	–	–	–	521	–	–	–
Occupation	Accurate matches	498	59.29	701	55.11	455	61.57	641	56.08
	Inaccurate matches	342	40.71	571	44.89	284	38.43	502	43.92
	Not comparable	2448	–	3876	–	1960	–	2884	–
Annual household income	Accurate matches	1686	53.37	2852	56.23	1418	54.62	2215	55.82
	Inaccurate matches	1473	46.63	2220	43.77	1178	45.38	1753	44.18
	Not comparable	129	–	76	–	103	–	59	–
Education [f]	Accurate matches	1167	43.13	1560	40.13	1092	47.44	1456	43.41
	Inaccurate matches	1539	56.87	2327	59.87	1210	52.56	1898	56.59
	Not comparable	582	–	1261	–	397	–	673	–
Household size [h]	Accurate matches	1049	31.90	1790	34.77	879	32.59	1396	34.67
	Inaccurate matches	2235	68.10	3358	65.23	1818	67.41	2631	65.33
	Not comparable	4	–	0	–	2	–	0	–

[a] An overlap sample of 1495 respondents exists in the NHTS and GDOT survey datasets before processing.

[b] An overlap sample of 1245 respondents exists in the NHTS and GDOT survey datasets after processing.

[c] Respondents who *did not want to be contacted again* are removed from the NHTS samples, as this subset had TM pre-processing prior to data enrichment. See Section 4.2 and online Appendix B for more information.

[d] Match percentages exclude "Not comparable" segments and should be interpreted as the percentage of respondents who could be compared with an equivalent category between data sources that are accurately matched (or inaccurately matched). Table A3 in online Appendix A summarizes the variable values that are compared to each other.

[e] The "Not comparable" value includes respondents in "Other/Could not be classified/Not applicable/Prefer not to answer/Missing" categories. These categories were not separated, because they are often confounded across sources. For example, in the TM data sources, "Missing" and "Not applicable" were not distinguishable from each other, although they were distinguishable for some of the questions in the survey data sources.

[f] Gender, age (tolerance +/- 4 years), and education (tolerance: +/- 2 levels) are used in post-processing to ensure that the TM records obtained are appended to the correct individuals. As such, the accuracy for these numbers in the post-processed sample are higher than would be typically expected (or unrealistically perfect, as in the case of gender). Note that even when instituting these matching criteria, we were able to retain 82.09% of the GDOT respondents and 78.22% of the NHTS respondents (i.e. we are relatively confident of having the correct TM records for ~ 80% of survey respondents). There remain "Not comparable" cases for gender, age, and education in the post-processing sample because we retained cases for which gender/age/education are missing in either the TM or survey datasets, as these could not be definitively ruled out based on inaccurate matches.

[g] NHTS did not obtain marital status and home dwelling type of survey respondents, and thus these variables could not be compared between TM and NHTS data. Similarly, GDOT did not obtain tenure, and thus this variable could not be compared between TM and GDOT survey data.

[h] When a tolerance of +/-1 was instituted for the household size variable, the percentage of accurate matches increases substantially, to: 71.92%, 72.18%, 72.38%, 72.29%, in respective order of the four percentages listed in the table.

of data used in transportation planning and forecasting. Thus, examining TM data relative to these transportation surveys, and further, being able to compare the NHTS and GDOT surveys relative to each other, represent unique contributions of this paper.

5.1. Investigating differences between TM and travel survey variables

The first step in assessing the quality of TM data lies in verifying the "accuracy" of its values for critical variables, such as SED variables, in the TM database. An ideal approach would entail the validation of TM variables with values from official records or reports (or alternatively, in-person verification). Given the absence of reliable SED data from publicly available disaggregate personal records, as well as the focus of this paper on examining TM data within a transport context, here we validate selected TM variables based on corresponding variables obtained/derived from the GDOT survey and the NHTS. Comparing TM data to federal and statewide transportation survey data can help transportation practitioners to better understand whether it is possible to replace, augment, and/or model specific travel behavior survey data with TM data, and further, can provide guidance for addressing identified discrepancies.

To date, the only TM validation studies for SED variables of which the authors are aware include an aggregate level validation for TM data at the block group level (Kressner and Garrow, 2014), as well as a small-scale household-level validation between TM and travel survey data (Kressner et al., 2014). As findings from these prior studies will be compared to results from the analysis in this paper, it is pertinent to note that the household-level validation in Kressner at al. (2014) used survey data from hard-to-reach populations, thus indicating a bias in the survey data toward individuals living below the poverty line.

Accordingly, the data validation presented in this section extends the preceding investigation by: (1) expanding the household-level validation to significantly larger (from $N = 116$ to $N \approx 5000$) and more representative samples; (2) allowing for the simultaneous pairwise comparison of TM data with two different types of transportation surveys; and (3) illustrating the effect of TM data processing on variable match rates. To facilitate comparison of the validation process with Kressner et al. (2014), we define a match (on a given variable) between the same case in two different datasets as being accurate if the case has the same value (within a tolerance band, if applicable) for that variable in both samples, and inaccurate otherwise. For example, if a given individual is in the 18–24 age category in the NHTS survey, but in the 25–34 age category for the TM data, then that case is considered an inaccurate match on age. The shares (or, if expressed as percentages, rates) are calculated only on comparable cases, as follows:

$$Accurate\ match\ share = \frac{Number\ of\ comparable\ cases\ with\ same\ variable\ value}{Number\ of\ comparable\ cases}, \text{and}$$

$$Inaccurate\ match\ share = \frac{Number\ of\ comparable\ cases\ with\ different\ variable\ values}{Number\ of\ comparable\ cases},$$

where "comparable cases" refers to cases that were able to be assigned a value that was able to be developed across all data sources (Table A3). Noncomparable cases include those with missing, not applicable, not able to be classified, other, "I don't know", and "prefer not to answer" responses to the variable in question in one or both datasets being compared. Using these definitions, Table 1 and Fig. 6 summarize match rates across the entire TM and survey datasets used, while Table A4 and Fig. A1 in online Appendix A summarize these rates for the *same* respondents across all three datasets (i.e. for the overlapped sample). Table A4 also includes GDOT/NHTS variable match rates to allow for insight into differences between the surveys. Prior to comparing the variables selected for validation, it was necessary to recode several variables into directly comparable categories; Table A3 in Appendix A summarizes this process, and details final variable values used. For consistency, in this paper we did not ourselves impute values for the NHTS, GDOT, or TM variables; however, some of the TM variables were imputed/infilled prior to our receipt of those variables. The TM variables that were specified as imputed in the TM database include household income and household size variables, which had missing values filled in with zip code and/or zip + 4 code data, and the marital status variable. which was filled in with undisclosed imputations. We note that other TM variables may have also been imputed in some way, but those listed here are the ones that were transparently listed as having been imputed in the TM database documentation.

As illustrated in Fig. 6 and Table 1, the match rates when comparing the TM and survey data are generally consistent for both the NHTS and the GDOT survey, with the highest accuracy rates occuring for gender, age, tenure, race, marital status, and dwelling type, and the lowest accuracy rates occuring for occupation, income, education, and household size. We see that gender has the overall highest percentage of accurate matches for both NHTS and GDOT data (90.9% and 95.6%, respectively), followed by age with match rates of around 89–91%. We posit that gender and age may have the highest match rates between TM and survey data due to the ease of obtaining these variables from publicly available records (e.g. birth records), although gender identification is also believed to be derived based on typical male and female names in the Caucasian population. This latter proposition is based on the observation that foreign names (e.g. names of Asian or Native American origin) are often listed as unidentifiable with regard to gender. Race had accuracy rates of $\sim$ 85% for both surveys, representing the fourth highest match rate among SED variables examined.

Housing tenure was comparable between NHTS and TM data only, and had the third highest accuracy rate of 87.31%, while marital status and dwelling type were only comparable between GDOT and TM data, and had the next highest accuracy rates of 72.2% and 63.05%, respectively. Occupation had lower accuracy rates of $\sim$ 59% between GDOT and TM data and $\sim$ 55% between NHTS and TM data; however this is likely because $\sim$ 75% of the cases could not be compared. While dwelling type and occupation were not studied in prior literature, we note that for gender, tenure, and marital status, our findings are consistent with those of Kressner et al. (2014). Regarding age and race, we found significantly higher accuracy rates than the prior work, potentially suggesting either a bias in TM data in reporting these variables for under-represented populations, or that the TM database used in this study had more accurate data on age and race.

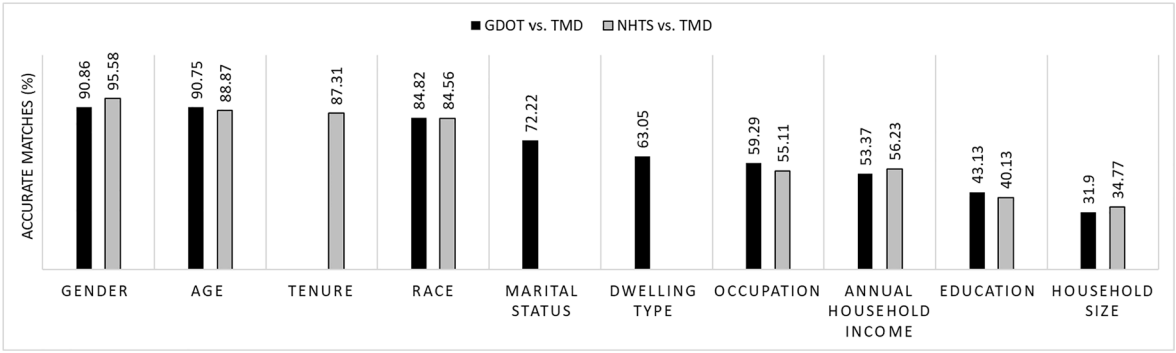

Fig. 6. Variable Accuracy Rates across TM and Survey Datasets *before Processing.*

Rounding out the rest of the variables, income, education, and household size all had accuracy rates below 55%, findings also shown by Kressner et al. (2014). This indicates the robustness of the finding that these individual-level variables have low accuracy rates (i.e., almost 50% or lower) in TM data, since they continue to do so five years after an initial validation study. Such performance may be attributable to the relative transience of these variables; for example, income, education, and occupation can all change several times over an individual's lifetime (we note as well that these variables do not change *consistently* over time relative to a transient variable like age). Similarly, household size is a constantly in-flux variable, as individuals marry/divorce/die and give birth to children, and as children move out of/into the household. When we allow a tolerance of $\pm$ 1 when calculating the household size accuracy rates, we see that the match rates more than double (from $\sim$ 30% to $\sim$ 70%, for both categorical and continuous versions of the variable), supporting the conjecture that for dynamic variables, TM may take several months to years to receive updated information, which at least partially accounts for the low accuracy rates observed. Thus, it is worth noting that for a low performing variable like household size, TM *can* provide more accurate estimates *within* a certain tolerance.

As discussed before (Section 4.2), gender, age, and education were used to process the TM data to retain records that were believed to correspond to the correct indvidual in the survey data sources. As Table 1 and Table A4 show, even after data processing (i.e., the sample is filtered to include only individuals who are considered to be definite matches between the TM and survey data), all of the variables with the exception of age and gender saw only small improvements in accuracy, suggesting that the accuracy rates observed for race, marital status, dwelling type, occupation, income, education, and household size are largely representative of the rates that could be typically expected for such variables in TM databases.

In Sections 5.2 and 5.3 we explore additional validation approaches, first examining distributional differences in the accurate and inaccurate matches, followed by a modeling effort that examines the factors influencing individuals' propensities to be matched correctly in the TM database.

5.2. *Distributions of accurate and inaccurate match rates among TM and survey sample pairs*

In this section, we examine whether the distributions of variable accuracy and inaccuracy are associated (i.e. correlated) with the (typically categorical) values the variable can take on. If the accurate and inaccurate match rates are similar across the values a given variable can take on (i.e. no association), then we can say that for that variable, there is no specific value category that is performing significantly better/worse than the others. This facilitates the assessment of which demographic values are reported with higher accuracy by the TM data. To achieve this goal, we report results from the chi-squared test of independence; however, due to the limitation that the chi-squared statistic is strongly influenced by sample size, we also report Cramer's V, a statistic that adjusts the chi-squared statistic using both sample size and number of cells in the contingency table. This adjustment allows Cramer's V to be comparable across contingency tables with different sample sizes and numbers of cells. Cramer's V ranges from 0 to 1, with higher values indicating high association.

Table 2 summarizes variable frequencies as well as measures of association across all variables and samples studied; note that the data used in this section is before matching on gender, age, and education had occurred so as to ensure that the results reported here are applicable to TM data in general (i.e. not biased by data processing). The final two columns of the table also present a direct comparison of the two survey datasets to each other to provide some context for comparing the other distributions (i.e. how much congruence exists even between the same questions asked on two surveys of the same sample). As shown in the table, the chi-squared test of independence is significant for almost all variables, even in cases where Cramer's V is relatively small (see for example, household size). This is likely due to sample size effects, and accordingly, here we primarily discuss Cramer's V statistic. We use Cohen's effect sizes (which are dependent on degrees of freedom) for Cramer's V to select which effects are large enough to merit discussion (Cohen, 1988).

The Cramer's V statistic for age has a large effect size, and a closer look at the frequencies indicates that the TM dataset is doing a better job at reporting ages for individuals in higher age categories, which is intuitive given that these individuals likely have more established transactional histories, and accordingly their ages are likely to be better represented in TM databases. Tenure also has a large Cramer's V effect size, with the frequencies showing that TM data are doing a poor job in identifying renters, an intuitive finding

Table 2
Identifying Patterns in Accurate and Inaccurate Match Rate Distributions for Sample Pairs.

Variable	SED characteristics	Frequency [a]					
		TM vs. GDOT [b] N = 3288[c]		TM vs. NHTS [b] N = 5148[c]		GDOT vs. NHTS [b] N = 1495[c]	
		Accurate Matches	Inaccurate Matches	Accurate Matches	Inaccurate Matches	Accurate Matches	Inaccurate Matches
Gender	Male	1553	125	1861	275	661	10
	Female	1311	285	2594	416	803	15
	χ^2 *statistic (df)*	79.937 (1) ***		0.882 (1)		0.096 (1)	
	Cramer's V	0.156 (small [d])		0.013 (small)		0.008 (small)	
Age	18–24 years	13	20	35	68	8	2
	25–34 years	168	88	370	260	95	4
	35–44 years	265	65	524	213	127	11
	45–54 years	452	87	695	200	209	11
	55–64 years	667	115	952	228	346	21
	65 + years	1241	85	1447	152	626	29
	χ^2 *statistic (df)*	221.380 (NA [e]) ***		426.310 (5) ***		7.818 (NA [e])	
	Cramer's V	0.260 (large)		0.288 (large)		0.072 (small)	
Tenure	Owner	–	–	3384	224	–	–
	Renter	–	–	784	708	–	–
	χ^2 *statistic (df)*	–		1199.5 (1) ***		–	
	Cramer's V	–		0.485 (med. to large)		–	
Race	Asian/Pacific Islander	8	50	5	75	12	4
	Black/African American	362	197	801	418	243	0
	Native American	0	25	2	10	2	9
	White/Caucasian	2071	444	2900	637	1135	40
	χ^2 *statistic (df)*	305.240 (NA [e]) ***		381.600 (NA [e]) ***		220.230 (NA [e]) ***	
	Cramer's V	0.311 (large)		0.281 (med. to large)		0.390 (large)	
Marital status	Married	1446	459	–	–	–	–
	Single	665	389	–	–	–	–
	χ^2 *statistic (df)*	53.859 (1) ***		–		–	
	Cramer's V	0.135 (small)		–		–	
Dwelling type	Stand-alone house	1574	331	–	–	–	–
	Apartment/condo	61	993	–	–	–	–
	Mobile home	_ [f]	_ [f]	–	–	–	–
	Attached home/duplex/ townhouse	0	318	–	–	–	–
	χ^2 *statistic (df)*	1953.200 (NA [e]) ***		–		–	
	Cramer's V	0.772 (large)		–		–	
Occupation	Professional, managerial, or technical	432	593	587	968	316	108
	Sales/service	24	278	31	601	54	44
	Manufacturing, construction, maintenance, or farming	21	57	42	228	22	4
	Clerical or administrative support	21	100	41	265	36	24
	χ^2 *statistic (df)*	139.910 (3) ***		302.550 (3) ***		20.107 (3) ***	
	Cramer's V	0.303 (large)		0.331 (large)		0.182 (medium)	
Annual household income	Less than US $50,000	628	367	1604	833	485	45
	US $50–99,999	556	595	738	786	358	158
	More than US $100,000	502	511	510	601	327	75
	χ^2 *statistic (df)*	55.760 (2) ***		176.880 (2) ***		82.642 (2) ***	
	Cramer's V	0.133 (small)		0.187 (small)		0.239 (medium)	
Education	Some grade school/high school	0	74	0	177	24	11
	Completed high school or GED	155	199	293	592	155	12
	Some college/technical school	213	764	333	1224	407	71
	Bachelor's degree	443	546	500	751	368	58
	Completed graduate degree (s)	356	531	434	843	370	15
	χ^2 *statistic (df)*	176.870 (4) ***		202.360 (4) ***		46.669 (NA [e]) ***	
	Cramer's V	0.232 (med. to large)		0.198 (medium)		0.177 (medium)	
Household size	Single-person HH	331	592	713	1177	455	38
	Two-person HH	430	999	627	1296	550	83
	Three-person HH	90	344	187	446	98	75
	Four-person or larger HH	198	300	263	439	148	46
	χ^2 *statistic (df)*	47.836 (3) ***		21.124 (3) ***		132.59 (3) ***	
	Cramer's V	0.121 (small)		0.064 (small)		0.298 (large)	

***, **, * = significant at 1%, 5%,10%, respectively.

[a] Distributions examined before matching on gender, age, and education (i.e. before data processing) as described in Section 5.1.

[b] For the GDOT vs. TM and GDOT vs. NHTS distributional comparisons, the GDOT survey is used to inform the SED characteristics of the accurate and inaccurate matches for the contingency table. Similarly, for the NHTS vs. TM distributional comparison, the NHTS is used to inform the SED characteristics for the contingency table. This assumes that the survey data are "correct" relative to the TM data, which is not necessarily always true. Nevertheless, we have reason to believe that for most of the cases, survey data are likely to be more reliable relative to TM data. Furthermore, as the goal of the study is to study TM data relative to transport survey data, we believe that using the survey data sources to inform the SED tabulations for the contingency tables is appropriate.

[c] Counts do not add up to 100% or the total N because of noncomparable categories; the number of such cases for each variable can be found in Table 1.

[d] Cohen's effect size classifications for Cramer's V are represented in parentheses following the Cramer's V value (Cohen, 1988).

[e] When the number of cases in a cell is small, a Monte Carlo procedure is used to calculate the test's p-value (Hope, 1968).

[f] The GDOT data did not have any individuals who reported living in mobile homes, so this category is not included in the distributional comparisons. Note however that the TM data *did* have 23 individuals who reported living in mobile homes.

given that: (1) renters are more likely to be lower-income individuals with fewer TM data records (and thus less accurate information); (2) renters tend to move more frequently than home owners, thus making it more difficult to maintain appropriate address information; (3) the apartment or unit number may be unavailable or incorrect for renters living in a multifamily dwelling at a given street address; and (4) renters may be living at rental properties that are single-family homes, making it difficult for TM data to accurately identify the tenure arrangement. Race also has large Cramer's V effect sizes, with the results showing that across TM and NHTS data, Asians and Native Americans are the most likely to be inaccurately represented, followed closely by African Americans. Thus, both TM data and NHTS more accurately represent individuals who identify as White, a finding that may be attributable to Whites being more integrated into the financial/transactional fabric of U.S. society, and thus TM having more accurate records/sources of information for these individuals. It is likely also partially due to missing ethnicity being infilled by the TM data provider using aggregate data, with the dominant race at aggregate levels more likely to be White.

Dwelling type had the largest effect size across all variables studied, with the results showing that the TM dataset is much more likely to correctly identify individuals living in single-family homes. This is likely due to the same reasons discussed earlier for the tenure findings, and suggests that TM databases do not have reliable/accurate sources of information for individuals' living arrangements, particularly in cases where address details are less precise. Occupation also has a high effect size, with TM data being significantly more likely to inaccurately identify occupation type for those who are not in the professional, managerial, or technical category, although there are more inaccurate than accurate matches across all categories. NHTS is also more likely to differentially represent occupation type relative to GDOT survey responses for these categories.

Education is seen to have a medium to high effect size, with TM data being more likely to inaccurately identify individuals who have not completed high school and individuals with some college/technical qualifications. We note that education does present some difficult-to-interpret findings here, with individuals who have a completed high school degree or bachelor's degree being more likely to have correct matches, while individuals with some college/technical qualifications and those who have completed a graduate degree being less likely to have correct education records in the TM data. In general, we would have expected that individuals with higher levels of education would have more sources of personal information (e.g. employment records) from which the education level can be gleaned, since in line with previous reasoning, they may have more established footprints in the TM database. We discuss this finding further in Section 5.3.

Marital status, household income, and household size have small effect sizes for the TM data comparisons in this study, and so deviations on these variables may be due to random fluctuations. However, it is interesting to note that three-person households are much more likely to have differences between the GDOT and NHTS surveys relative to the other household size categories, a finding that may be attributable to the one-year difference in survey administration for the GDOT and NHTS surveys. Further examination indicated that most of the incorrectly classified households in this category were two-person households in the NHTS survey that became three-person households in the GDOT survey, suggesting possible life stage changes like marriage or the birth of a child occurring in the (average) one-year gap between surveys.

To compare the findings from this study to the literature, we see that Kressner et al. (2014) used chi-squared tests of independence to examine patterns of association, and found no significant associations, with the primary exception of marital status. There was a higher occurrence of single individuals who had a correct match for marital status relative to married individuals, which Kressner et al. (2014) suggested may be because the TM database assumes that an individual is single until information is obtained that proves otherwise. However, the frequencies for marital status in the study presented here tell the inverse story, with TM doing a better job of identifying marital status for those who are married. This difference in finding may be attributable to the particular population that was sampled in the prior study or to differences in how the marital status variable was developed in the two separate TM databases. Kressner et al. (2014) also found that there were more households than expected whose targeted marketing data matched for the African American category, with fewer individuals who matched for the White category, but similarly we believe that this may be due to the distinctive population sampled for that study, a proposition also suggested by the authors.

5.3. *Exploring biases for survey respondents more likely to be matched in TM databases*

This section follows closely from the preceding sections, but refocuses the examination at the individual level as opposed to the variable level, examining the factors influencing individuals' propensities to be matched correctly in the TM database. Individuals are considered to have a correct match in the TM database if the survey record reflected the same gender, age within a +/- 4-year age

tolerance, and education within a +/- 2-level tolerance with the returned TM record. Understanding which individuals may be better represented in TM databases facilitates an understanding of biases that can result when using TM data for transport applications. To assess these biases, we develop a binary logit model (Table 3) to predict whether a given respondent obtains a correct match in the TM database in terms of the gender, age, and education thresholds instituted during the matching process (unmatched respondents also include individuals whose TM records were missing gender, age, and education, as these TM records could therefore not be checked relative to the respective survey record). For simplicity, we limit this model to the GDOT survey dataset (N = 3288; reduced to 3121 through the removal of missing values for this model); and the exogenous variables tested in the model include gender, age, race, education, occupation, household size, household income, marital status, and a measure of population density.

Gender, age, all levels of education, and race identification as African American or Caucasian are all significant predictors of the probability of receiving a correct match in the TM database used for this analysis. Women are less likely to be among those who have a correct match, an intuitive finding given that TM databases are largely derived from financial records and transactions which are often still dominated by males. Older individuals are also more likely to have a correct match, which may point to the increased probability of older individuals to have more established financial/transactional footprints. In the case of age, the inherent *survey* bias of the GDOT dataset toward older individuals is likely reflected in the small disparity between mean ages for the matched and unmatched records, and accordingly we suggest that there may be a greater difference between these means in a survey dataset that is more representative of all ages.

With regard to race, with Asian/Pacific Islander as the reference group, we see that Blacks/African Americans and Whites /Caucasians are significantly more likely to be among those who receive a correct match. However, as the incidences show, Blacks have a greater proportion of unmatched records than matched records (whereas the opposite is true for Whites), suggesting that while Blacks are more likely to be included in matched records relative to Asians/Pacific Islanders, they are on the whole likely to be underrepresented in the TM database.

The model also indicates that, relative to individuals who have not completed high school, those with higher levels of education are more likely to have correct matches in the TM database, although those with graduate degrees are less likely to be matched relative to those who have completed some or all of their undergraduate education. This latter nuance, also shown in Section 5.2, may point to a higher proportion of foreigners among those with graduate degrees, relative to those with undergraduate degrees (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2007). Foreigners may be more likely to have incorrect TM records on several accounts; for example, individuals who have recently moved to a country are likely to have fewer records from both administrative and transactional sources. In addition, as before mentioned, gender misidentification may be higher for foreign names. Nonetheless, overall, the education findings suggest that TM databases may overrepresent more highly educated individuals, which is in line with the conceptual understanding that TM databases have more robust records for individuals with more financial assets and transactions associated with their names.

The findings in this section[1] support conventional intuition about the nature of TM databases, and along with the model findings in Table 3, serve to remind analysts interested in using TM data that at least currently, there are certain demographics, notably women and certain ethnicities, that are disproportionately affected by underrepresentation in TM data.

6. Discussion

Using various validation methods, we have now shown that TM data are able to provide accurate information (relative to self-reported data) for some variables and populations, while underrepresenting others. This is not unexpected, given that all data sources, active and passive, will inevitably suffer from unique biases and shortcomings. In fact, this serves to reinforce the earlier suggestions that it is critical for researchers working with new data sources to first internally validate novel data using an array of methods, and preferably, to also have these data validated by differing teams of researchers. Without undertaking thorough internal validation investigations, biases present in various datasets may be unknowingly integrated into decision-making processes and affect key transport outcomes like equity and wellbeing. While it is outside the scope of this paper to provide an extensive discussion on approaches that can be used to address dataset biases (see for example: Cahan, Hernandex-Boussard, Thadaney-Israni, & Rubin, 2019), our aim here is to provide a brief recap of the internal validation exercises, and to provide recommendations for methods that may be useful in the specific context of the TM data being examined in this study.

In Section 5.1, we saw that TM data are able to provide accurate data on several key variables (gender, age, tenure, and race) for 75% or more of individuals in the two survey samples studied. We wish to emphasize a point first made by Kressner et al. (2014), that even the variables that were found to have the lowest accuracy rates (~31–34%), indicate that with TM data, we may be able to accurately predict these variables for at least a third of the population at a significantly lower cost than it would take to acquire these variables using surveys. In Section 5.2, we examined distributions of accurate and inaccurate matches across all variable values to provide an understanding of how specific categories of each variable are performing. This investigation showed that age, race, dwelling type, occupation, and education perform differently across categories. This means that it may be especially important to

[1] For exploratory purposes, we also tested several TM variables in the model; however, since the TM variables for the non-matched individuals may not be correct at individual and/or household levels, we did not include these in the final model, but only mention them here. Two TM variables of interest that are significant are consumer prominence and technology adoption, with higher levels of both indicating increased likelihoods of having a correct record in the TM database. The consumer prominence indicator is a measure of how large the consumer footprint of the individual might be, while the tech adoption indicator is a measure of how likely a household may be to purchase new technologies at premium prices.

Table 3
Binary logit model of whether a GDOT survey record is correctly matched to TM database.

Variables [a]	Coefficients	Variable Incidence (%) [b]	
		2568 Matched Records	553 Unmatched Records
Constant	−1.593***	–	–
Gender (female)	−0.230*	47.08	56.67
Age	0.021***	60.11[c]	54.97[c]
Race			
Reference group: Asian/Pacific Islander	–	1.32	4.34
Black/African American	0.889**	16.90	20.43
Native American	0.252	0.66	1.27
White/Caucasian	1.081***	81.11	73.96
Education			
Reference group: Some grade /high school	–	1.79	4.52
Completed high school or equivalent	0.983***	11.06	10.67
Some college/technical school	1.204***	30.84	24.95
Bachelor's degree	1.233***	31.00	25.68
Completed graduate degree (s)	0.752**	25.31	34.18
Model attributes			
Number of observations	3121		
$\mathscr{L}(0)$	−2163.312		
$\mathscr{L}(\boldsymbol{c})$	−1457.822		
$\mathscr{L}(\widehat{\boldsymbol{\beta}})$	−1401.567		
$\rho^2(\mathscr{L}(0)$ base)	0.350		
Adjusted ρ^2 ($\mathscr{L}(0)$ base)	0.352		
$\rho^2(\mathscr{L}(\boldsymbol{c})$ base)	0.039		

***, **, * = significant at 1%, 5%,10%, respectively.

[a] The variables in this model are derived from the GDOT survey records for these respondents. As with all data sources, the GDOT survey may have its own implicit survey/nonresponse biases that may influence these numbers.

[b] Variable incidence represents the percentage of matched and unmatched records falling into the respective variable categories; for example, 46.85% of the matched records are females, while 56.67% of the unmatched records are females. Again, the GDOT survey was used to obtain the values for these variable incidences.

[c] As age is a continuous variable in the model, the mean ages for the matched and unmatched records are reported here in place of the incidence. Thus, note the sample bias toward older ages, even among unmatched records but especially among matched records.

realize that TM data may be providing incorrect information at a higher rate for certain individuals; for example: younger individuals, renters, minorities, etc. In Section 5.3 we explored the biases present for those who were considered to have a correct record match in the TM database, finding that at an individual level, women, minorities, younger individuals, and those with lower levels of education are less likely to have a correct record in the database.

Practitioners seeking to address biases such as those described here may: (1) seek to augment the data source in question with additional records/cases from other data streams that may be more representative of specific populations (i.e., data fusion); (2) develop algorithms/models to impute variable values for segments of the population that have increased probability of having incorrect values; (3) develop weights that can adjust the sample for the variables on which biases have been identified (Solon, Haider, & Wooldridge, 2015); and (4) interpret results within the lens of the biases that may exist, ensuring that the proper caveats are applied when making policy recommendations. These approaches represent some of the possible solutions that we believe could be applied to address the TM data biases identified in the preceding section. However, there are certainly other approaches, and we believe that all transportation researchers and practitioners who work with user-centered data should make it a priority to explore the methods and approaches that can be used to address dataset biases.

A final point of discussion entails placing the findings shown here in the context of transportation applications such as those described in Section 3. Throughout the paper, it has been emphasized that examining the veracity and value of TM data (and any novel data source) is critical for guiding the integration of select TM variables in transportation applications. Based on the results from the internal validation, which aimed to assess the veracity of the data, we see that while TM data are able to provide representative data for core SED variables relative to transportation surveys, that performance decreases as the SED variables become more complex and/or transient (i.e., variables that change over time). Accordingly, it is likely that by their nature, the transportation-related variables (such as business-related travel and vehicle ownership) present in the TM dataset may be subject to much lower rates of accuracy, not to mention the host of other challenges discussed in Section 2.3. These findings underscore the approach emphasized throughout this paper in which we recommend that, at least at present, TM data be seen as a source of augmentation and supplementation, rather than replacement, for traditional transportation variables and datasets.

7. Summary and conclusions

Given the "growing resistance among U.S. householders to surveys in general" (PTV NuStats, 2011, p. 43), it is increasingly important to examine additional sources of data that can be used to supplement transport modeling needs. In this paper, we make the case that targeted marketing (TM) data are ripe for integration into transportation applications, beginning with a detailed look at the benefits and challenges of using TM data (Section 2). We then develop a typology illustrating that TM data can be useful to a range of transportation applications and research, allowing for improved transportation models and innovative approaches that could reduce our reliance on traditional transportation data sources (Section 3). Using our experience integrating TM data with two transportation surveys (NHTS and a GDOT-funded survey), we present a framework of the TM data enrichment process (Section 4), providing a detailed case study of the process for analysts who may wish to pursue similar TM data integration and enrichment (Appendix B).

We use the resultant integrated datasets to assess the veracity of the TM data purchased, and demonstrate that TM data match gender, age, tenure, race, marital status, and household size at rates of 70% or greater relative to self-reported survey data (Section 5.1). However, we see that TM data exhibit differential accuracy across some variable categories; for example, the database does a poor job correctly identifying tenure and dwelling type for renters and those not living in single-family homes (Section 5.2). This may suggest that transportation professionals who use TM data in the future may need to impute or otherwise supplement data for demographic categories that tend to be inaccurately reported in TM databases. Additionally, an examination of TM biases reveals that men, older and better-educated individuals, African Americans, and Caucasians are more likely to have correct records in TM databases (Section 5.3). These are comparable though not identical to typical HHTS respondent biases, suggesting that similar approaches taken to address biases in transportation survey data may need to be applied here (Section 6).

In addition to the forthcoming companion study that provides a thorough external validation of the value (worth/usefulness) of TM data (Shaw, 2021; Shaw & Mokhtarian, in-progress), there are numerous avenues of future work that can be pursued in the aim to better understand the potential benefits of TM data in transportation. Notably, practitioners may be interested in better understanding the veracity of *travel behavior* variables that are present in TM databases, and thereafter, to investigate the value of such variables within modeling and forecasting efforts. To date, the authors are aware of only one paper that has sought to examine the veracity of travel behavior variables present in TM data, and that work has yielded promising results that certainly call for the further investigation of such variables by all transportation professionals (Lovelace et al., 2016). In addition, it will certainly be critical for the transportation community to have various teams of researchers investigate the applications and research directions proposed in the typology described in Section 3, as currently only a handful of studies thus far have tested the veracity and value of TM data in similar applications or contexts. Of special interest will be methods for integrating and fusing TM data with other active and passive data sources, as this approach will aid in overcoming biases present across the various data sources while creating an enriched dataset that can facilitate novel analyses and insights.

However, while we see significant potential in the use of TM data in transport applications, there remain challenges hindering the wide-scale application and integration of these data for modeling purposes in the transport domain – challenges that could intensify as we move through a period of increasing privacy regulations. Both as engineers and as private citizens, it is in our best interest to pursue TM data research and practice opportunities that will protect individuals' privacy while allowing for societal gains. It will be increasingly important for professionals to work with policymakers to strike such a balance, particularly in light of the growing need to supplement traditional data sources with various passively collected data sources, all of which are subject to the same privacy regulations discussed in this paper. In closing, we hope that this resource will encourage transportation professionals to further explore the benefits of targeted marketing data for moving transportation research and practice forward, while encouraging the contribution of new perspectives on approaches and methods that can be used to address some of the many challenges inherent not only in TM data, but third-party, passive, big data sources at large.

CRediT authorship contribution statement

F. Atiyya Shaw: Conceptualization, Methodology, Formal analysis, Validation, Data curation. **Xinyi Wang:** Data curation. **Patricia L. Mokhtarian:** Supervision, Writing - review & editing. **Kari E. Watkins:** Writing - review & editing.

Acknowledgements

This work was funded under the Teaching Old Models New Tricks (TOMNET) Center, a University Transportation Center sponsored by the U.S. Department of Transportation through Grant No. 69A3551747116. The authors are especially grateful to Ali Etezady who first identified and encouraged the potential contribution of this paper, and to Sung Hoo Kim who played an integral role in the development of the GDOT survey. This study has also profited from discussions with Drs. F. Alemi, G. Circella, Y. Lee, A. Malokin, and other members/visitors of the TOMNET team. Any opinions and findings expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor organizations.

Appendices A and B:. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tra.2021.04.021.

References

Acxiom. (2020). General Data Protection Regulation Privacy Notice. Retrieved from: https://www.acxiom.com/about-us/privacy/gdpr/.

Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice 30 (6), 415–429. https://doi.org/10.1016/0965-8564(96)00004-3.

Binder, S., Macfarlane, G.S., Garrow, L.A., Bierlaire, M., 2014. Associations among household characteristics, vehicle characteristics and emissions failures: An application of targeted marketing data. Transportation Research Part A: Policy and Practice 59, 122–133. https://doi.org/10.1016/j.tra.2013.11.005.

Birkin, M., Morris, M., Birkin, T., Lovelace, R., 2017. Using census data in microsimulation. In: Stillwell, J. (Ed.), Census Users Handbook 2011. Ashgate, London.

Birkin, M., 2019. Spatial data analytics of mobility with consumer data. Journal of Transport Geography 76, 245–253. https://doi.org/10.1016/j.jtrangeo.2018.04.012.

Cahan, E.M., Hernandex-Boussard, T., Thadaney-Israni, S., Rubin, D.L., 2019. Putting the data before the algorithm in big data addressing personalized healthcare. npj Digital Medicine 2 (78). https://doi.org/10.1038/s41746-019-0157-2.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. L. Erlbaum Associates, Hillsdale, N.J.

Connelly, R., Playford, C.J., Gayle, V., Dibben, C., 2016. The role of administrative data in the big data revolution in social science research. Social Science Research 59, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015.

Erevelles, S., Fukawa, N., Swayne, L., 2016. Big Data consumer analytics and the transformation of marketing. Journal of Business Research 69 (2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001.

Federal Highway Administration. (2018). National Household Travel Survey. Retrieved July 23, 2019 from https://nhts.ornl.gov.

Handy, S., Cao, X., Mokhtarian, P., 2005. Correlation or causality between the built environment and travel behavior? Evidence from Northern California. Transportation Research Part D: Transport and Environment 10 (6), 427–444. https://doi.org/10.1016/j.trd.2005.05.002.

He, S.Y., Miller, E.J., Scott, D.M., 2018. Big data and travel behaviour. Travel Behaviour and Society 11, 119–120. https://doi.org/10.1016/j.tbs.2017.12.003.

Hope, A.C.A., 1968. A Simplified Monte Carlo Significance Test Procedure. Journal of the Royal Statistical Society. Series B (Methodological) 30 (3), 582–598. https://doi.org/10.1111/j.2517-6161.1968.tb00759.x.

Khan, S.M., Ngo, L.B., Morris, E.A., Dey, K., Zhou, Y., 2017. Social media data in transportation. In: Chowdhury, M., Apon, A., Dey, K. (Eds.), Data Analytics for Intelligent Transportation Systems. Elsevier, pp. 263–281.

Kitamura, R., 2009. Life-style and travel demand. Transportation 36 (6), 679–710. https://doi.org/10.1007/s11116-009-9244-6.

Kim, S., Mokhtarian, P. L., & Circella, G. (2019). The Impact of Emerging Technologies and Trends on Travel Demand in Georgia. Final Report, Georgia Department of Transportation Research Project 16-31, available from the authors and at http://g92018.eos-intl.net/G92018/OPAC/Index.aspx.

Kressner, J. D., Macfarlane, G. S., Huntsinger, L., & Donnelly, R. (2016). Using passive data to build an agile tour-based model: a case study in Asheville. Paper presented at the International Conference on Innovations in Travel Modeling, Denver, CO. Retrieved from: https://pdfs.semanticscholar.org/27ae/db2df8e8709ece22d2042aea75d403df9285.pdf.

Kressner, J. D. (2017). Synthetic Household Travel Data Using Consumer and Mobile Phone Data (Report No. 184). Washington D.C.: National Cooperative Highway Research Program (NCHRP) Innovations Deserving Exploratory Analysis (IDEA) Program, Transportation Research Board. Retrieved from: http://www.trb.org/Research/Blurbs/176216.aspx.

Kressner, J.D., Garrow, L.A., 2012. Lifestyle segmentation variables as predictors of home-based trips for Atlanta, Georgia, airport. Transportation Research Record: Journal of the Transportation Research Board 2266, 20–30.

Kressner, J.D., Garrow, L.A., 2014. Using third-party data for travel demand modeling: comparison of targeted marketing, census, and household travel survey data. Transportation Research Record: Journal of the Transportation Research Board 2442, 8–19.

Kressner, J. D., Carragher, M. F., & Watkins, K. E. (2014). A household-level pairwise comparison of targeted marketing data and self-reported survey data. Paper presented at the 93rd Annual Meeting of the Transportation Research Board, Washington D.C. Retrieved from: https://www.researchgate.net/publication/341909525_A_Household-Level_Pairwise_Comparison_of_Targeted_Marketing_Data_and_Self-Reported_Survey_Data.

Laney, D., 2001. 3D data management: controlling data volume, velocity, and variety. META Group Research Note 6.

Lavalle, S., Lesser, E., Shockley, R., S. Hopkins, M., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT Sloan Management Review, 52(2), 21-31.

Lovelace, R., Birkin, M., Cross, P., Clarke, M., 2016. From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows. Geographical Analysis 48 (1), 59–81. https://doi.org/10.1111/gean.12081.

Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? Advances in Classification Research Online; 24th ASIS SIG/CR Classification Research Workshop. Retrieved from https://journals.lib.washington.edu/index.php/acro/article/view/14671/12311.

Ma, J., Li, H., Yuan, F., Bauer, T., 2013. Deriving operational origin-destination matrices from large scale mobile phone data. International Journal of Transportation Science and Technology 2 (3), 183–204. https://doi.org/10.1260/2046-0430.2.3.183.

Macfarlane, G. S. (2014). Using Big Data to Model Travel Behavior: Applications to Vehicle Ownership and Willingness-to-Pay for Transit Accessibility. (Doctor of Philosophy Dissertation). Georgia Institute of Technology, Atlanta, GA. Retrieved from: https://smartech.gatech.edu/handle/1853/51804.

Macfarlane, G.S., Garrow, L.A., Mokhtarian, P.L., 2015a. The influences of past and present residential locations on vehicle ownership decisions. Transportation Research Part A: Policy and Practice 74, 186–200. https://doi.org/10.1016/j.tra.2015.01.005.

Macfarlane, G.S., Garrow, L.A., Moreno-Cruz, J., 2015b. Do Atlanta residents value MARTA? Selecting an autoregressive model to recover willingness to pay. Transportation Research Part A: Policy and Practice 78, 214–230. https://doi.org/10.1016/j.tra.2015.05.010.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future. Washington, DC: The National Academies Press. Retrieved from: https://www.nap.edu/catalog/11463/rising-above-the-gathering-storm-energizing-and-employing-america-for.

National Research Council. (2013). Nonresponse in Social Science Surveys : A Research Agenda. Washington, DC: The National Academies Press. Retrieved from: https://www.nap.edu/catalog/18293/nonresponse-in-social-science-surveys-a-research-agenda.

PTV NuStats. (2011). Regional Travel Survey: Final Report. Atlanta, Georgia. Retrieved from: https://cdn.atlantaregional.org/wp-content/uploads/tp-2011regionaltravelsurvey-030712.pdf.

Ruiz, T., Mars, L., Arroyo, R., Serna, A., 2016. Social Networks, Big Data and Transport Planning. Transportation Research Procedia 18, 446–452. https://doi.org/10.1016/j.trpro.2017.01.122.

Salomon, I., Ben-Akiva, M., 1983. The Use of the Life-Style Concept in Travel Demand Models. Environment and Planning A: Economy and Space 15 (5), 623–638. https://doi.org/10.1068/a150623.

Solon, G., Haider, S.J., Woolridge, J.M., 2015. What are we weighting for? Journal of Human Resources 50 (2), 301–316. https://doi.org/10.3368/jhr.50.2.301.

Shaw, F.A., 2021. Methods for Enriching Transportation Survey Datasets: With Sample Applications Using Psychometric Variables. Doctor of Philosophy Dissertation). Georgia Institute of Technology, Atlanta, GA.

Shaw, F. A & Mokhtarian, P. L. (in progress). Assessing the value of targeted marketing data for modeling travel behavior. Available upon request from authors.

Shaw, F. A., Wang, X., Mokhtarian, P. & Watkins, K. (in progress). A framework for enriching survey datasets using big data and machine learning, with an application for transferring attitudinal variables across transport surveys. Available upon request from authors.

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., 2017. Critical analysis of big data challenges and analytical methods. Journal of Business Research 70, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C: Emerging Technologies 58, 162–177. https://doi.org/10.1016/j.trc.2015.04.022.

Van Acker, V., Goodwin, P., Witlox, F., 2016. Key research themes on travel behavior, lifestyle, and sustainable urban mobility. International Journal of Sustainable Transportation 10 (1), 25–32. https://doi.org/10.1080/15568318.2013.821003.

Welch, T.F., Widita, A., 2019. Big data in public transportation: a review of sources and methods. Transport Reviews 1–24. https://doi.org/10.1080/01441647.2019.1616849.

Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies 87, 58–74. https://doi.org/10.1016/j.trc.2017.12.003.

Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society 11, 141–155. https://doi.org/10.1016/j.tbs.2017.02.005.