A Guideline Framework for Using Information and Communication Technology (ICT)-Based Data in Travel Demand Modeling

Katherine E. Asmussen
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: kasmussen29@utexas.edu

Vivek Verma
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: vivek.verma@utexas.edu

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 1-512-471-4535; Email: bhat@mail.utexas.edu

ABSTRACT
The omnipresence of information and communication technologies (ICTs), such as smartphones, GPS, Bluetooth, and tablets, is inevitably influencing travel behaviors. In this study, we develop a guideline framework for how ICT data sources can beneficially augment travel demand modeling (TDM) processes. Though ICT data's sole use is typically not adequate for TDM, the data can be gainfully used in combination with traditional data sources (TDSs) to support modeling. However, careful consideration is required to understand the differences between the traditional and the emerging data sources, and strategies must be devised to integrate these new sources into the TDM process. Such a methodical integration process can benefit the TDM process in terms of both results and resource needs, compared to models based solely on traditional strategies and travel surveys.

Keywords: Transportation, Data Fusion, Information and Communication Technology (ICT) Data, Travel Demand Models, Mobility Patterns.

1. INTRODUCTION

The label "ICT" describes technologies that offer access to information through telecommunications. This includes smartphones, social media platforms (Twitter, Facebook, Tripadvisor), GPS navigators, sensor data (from Bluetooth devices, videos, MAC-address-based sources, drones, and remote sensing/satellite images), tablets, smart cards, travel transactional data (e.g., car-sharing and ride-hailing), and various wireless devices that are becoming ubiquitous. ICTs provide a new and easier way to passively access data at the individual level, allowing analysis to shift from a vehicle-level scale to a much finer human-level mobility pattern scale. A departure from the traditional household travel survey data, emerging data are characterized by voluminous amounts of near-real-time data, and they do not require sampling, as they are able to reach entire sections of the population (e.g., when using telephone and bank services) and entire groups of interest in certain moments or locations (e.g., when using smart cards, toll roads, etc.). At the same time, these emerging "big" data sources are also typically coarser and more sectorial than traditional data, and many challenges remain in the use of these emerging ICT-based data sources. For example, data collected from smartphones could be incomplete, as data could be missing due to various scenarios: (a) when the phone is indoors with no signal connection between the phone's GPS sensor and a satellite, (b) when the phone's battery gets discharged, and (c) at the beginning of trips due to the GPS software's start-up time. In addition, areas with poor cellphone coverage will adversely affect cell-based data collection efforts. Table 1 provides a synopsis of the main differences between traditional and emerging data.

Table 1. Differences between Traditional and Emerging Data

Traditional data	**Emerging data**
Strategic sample	Almost an entire section of a population
Respondent stated (RP/SP)	"True" observed data[1]
Fine/demographic (disaggregate)	Coarse (aggregate)
Multidimensional (many small pieces of information for the same person) on respondents' mobility and demographics	Unidimensional (more sectorial) focused on mobility
Small sample/easy to process	Voluminous/more difficult to process
Clarity in sharing/privacy/access	Fuzziness in sharing/privacy/access (not all data are clearly associated with a specific person)

(Source: Modified from (*1*))

In the context of the above discussion, in this study, we develop a guideline framework for combining ICT data with traditional survey data to answer questions such as: (1) how can available ICT data be used to complement traditional datasets; (2) how should emerging data be integrated into the current travel demand modeling processes used across the country; and (3) how is this fusion with traditional data sources and integration into the TDM process impacted by the types of information (individual-level sociodemographics, *identifiers*, trip paths/purpose information, transportation systems) contained (or lacked) by ICT data? To be sure, the integration of ICT data

[1] The label "true" data is used here simply to highlight the fact that emerging data are not "filtered" by the respondents when interviewed. "True" is not used with the meaning of "free of bias" as all systems have their own limitations, which might be technical, stem from data transformation, or relate to user manipulation (for example, fraud or multiple validation occurrences within smart card data). Providers of these data also do not always give access to the original data, but sometimes only to processed data.

sources into TDM processes depends on factors such as the frequency of data collection, the format of data collection, and the information content captured. Another important consideration for integrating ICT data in TDM processes is the data content, or the range of information available from the ICT source, and whether the ICT source provides information on the movement of individual travelers or on an aggregation of the movement of individual travelers.

This study will focus first on categorizing evolving ICT-based methods for collecting data into six distinct classes. Next, a strategy to deduce the context, content, and unit level of each ICT data source will be presented and discussed. Third, four fusion methods are proposed on how to best integrate ICT data source with TDSs for TDM purposes. Fourth, an integration framework is proposed to help map every ICT data source to the appropriate fusion strategy. The study will conclude with an assessment of the framework and each fusion method.

2. FRAMEWORK DEVELOPMENT AND ASSESSMENT

2.1. Classes of ICT Data Sources

Technological advancements in the past decade have made possible the passive and large-scale collection of travel data using portable devices. ICTs provide a new and easier way to passively access data at the individual, vehicle, and regional levels, allowing analysis to shift from a vehicle-level scale to a much finer human-level mobility pattern scale. In the initial steps of developing the fusion framework, we first had to identify and assess the numerous classes of available ICT data sources that other planning, transportation, and research organizations and agencies across the world have begun to use. We performed a thorough review of nearly 40 ICT sources on metrics such as data content (demographic/geospatial/both), frequency of collection (one time/constant), availability of trip information (such as activity purpose, and trip mode choice), accessibility of data (open source/private collection format/must buy), collection method, completeness of data, accuracy of geo-spatial information, quality of collection, and representation of population. While the review is not presented in the current manuscript, we were able to identify significant commonalities across all data sources, which led us to the identification of six common classes, including:

- Social media data (e.g., online platforms such as Twitter, Facebook, and Instagram)
- Vehicle data (e.g., connected or probe-based vehicle information, such as that used to develop Wejo's datasets, or taxi-only trip records)
- Mobile phone records
- Apps (e.g., user-based apps such as Uber, Strava, Waze, and various travel diary apps)
- Smart card transactions (e.g., bus and other public transportation user passes)
- Origin-destination pre-aggregated data (e.g., from StreetLight or Inrix)

Each of these ICT source classes provide differently formatted geospatial information about an activity, event, or trip and can be used in different ways when converted, translated, and integrated with traditional data sources. ICT Class-specific integration methods will be discussed with the integration framework in later sections of this study.

2.2. Types of Information Available in ICT Data

ICT data can be useful to modelers because of the passive nature of data collection, which makes it non-intensive from a human resource perspective. Most ICT data used for TDM contain voluminous human-level activity travel data, covering a wide range of travel dimensions over extended time periods. But the information content provided by different types of ICT data sources

can vary considerably. In the next few subsections, we discuss the content of ICT data sources based on a cascading three-level structure of main context, additional content, and information unit level of data, of which can be visualized in Figure 1.

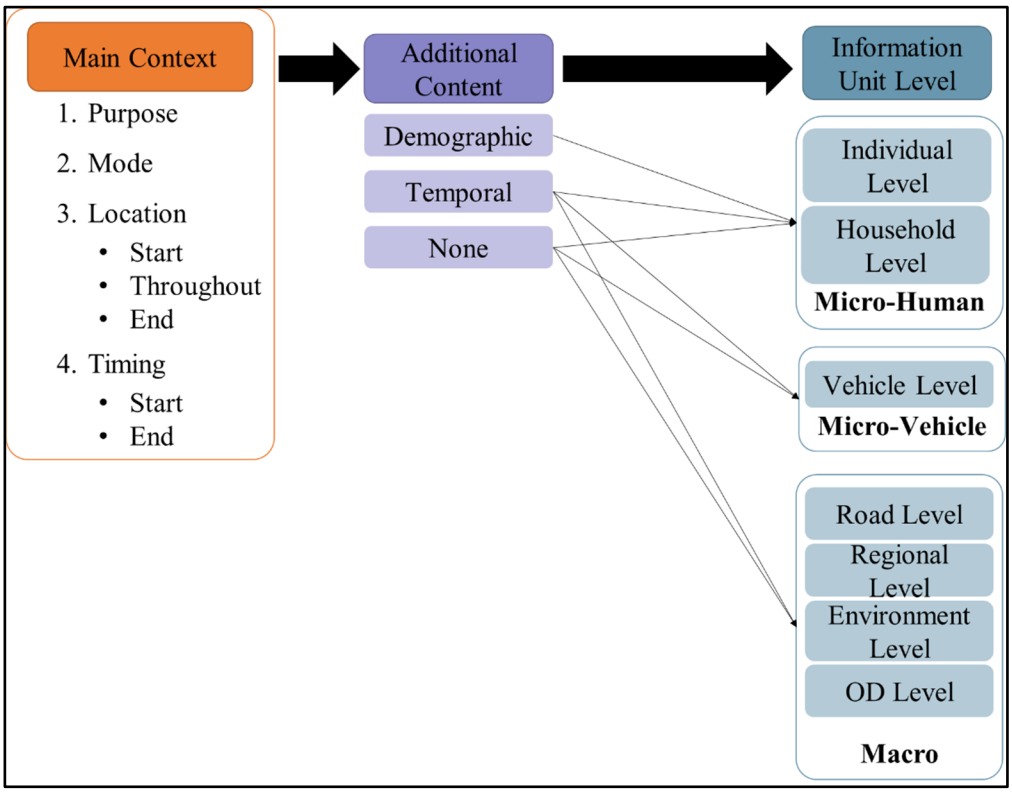

Figure 1 ICT Data–Type Categorization

2.2.1. Main Context

The "Main Context" and its subcategories (see left side of Figure 1) relate to descriptors of the trip and activity data collected in ICT datasets.

- Purpose: The trip purpose, which can be broadly classified as home-based work, home-based non-work, or non-home-based, or more disaggregate by destination activity purpose.
- Mode: The mode of the trip, which can be car, bike, walk, bus, or other public transport.
- Location: The location (generally, the geospatial coordinates) of the origin, destination, or whole trip (tracked continuously).
- Timing: The start time and end time of the trip.

An ICT dataset, in general, will not contain all of these main contexts. But almost all ICT sources include, at a minimum, location information, i.e., the geospatial qualifier of an activity/trip/event. Unfortunately, for many ICT sources, it can be difficult to gather data beyond location and basic timing information. For example, Twitter data (e.g., Tweets about a concert that may be the leading cause for a traffic delay) can be quickly mined for any activity in a specific geographic location at a given time, but determining what the event was and what types of people were there (purpose), or how they got there (mode), and "linking" this with the easily obtained location and timing contextual information is more difficult and time-consuming.

2.2.2. Additional Content

The second column of Figure 1 shows the "Additional Content" that some ICT sources may have, though this is not standard. The two additional dimensions are as follows:

- Demographic: Individual or household information about the user or participant in the activity/trip/event.
- Temporal: Time stamp of when the activity/trip/event identified in the data "response" occurred.

Demographic data can help link activity patterns to groups of individuals. For example, if an ICT source provides the age and gender of a user, then the geospatial data can be linked to the aggregated data on individuals of the same age and gender class that were collected through traditional data sources, such as a household survey. If no demographic content is provided, then the ICT sources can only be aggregated based on the geospatial region that the response occurred in, and potentially only linked to coarse spatial levels, such as block-wide or zonal-level data collected from the census.

Temporal information is usually available in some basic form in ICT data alongside geospatial information (in the context of start and end time of trips, as identified under "Main Context"). But some ICT sources may provide richer temporal information than simply describing when an event occurred. For example, temporal information may include the duration of an event, as is common in ICT data sourced from apps (e.g., Strava or Uber), where the dataset records when a user starts the event and then tracks their geographic location until they finish the event. This event would be recorded as a single entry in the ICT dataset. On the other hand, ICT data collected from mobile phone records will only have the limited temporal information of when a phone call occurred, alongside where it occurred geographically.

2.2.3. Information Unit Level

The Information Unit Level in an ICT dataset may correspond to one of the three categories discussed below, which are listed in the right "column" of Figure 1:

- Micro-Human
  - Individual Level: The dataset contains sociodemographic information about an individual, such as their age, gender, income, and education.
  - Household Level: The dataset contains household-level sociodemographic information, such as household income, number of household members, and the number of trips made by the household.
- Micro-Vehicle
  - Vehicle Level: The dataset contains vehicular-level information, i.e., speed and location of each vehicle.
- Macro (aggregate)
  - Road Level: The dataset contains information about the number of vehicles present on the road at a particular time or the number of vehicles passing a point on the roadway during a given time interval (aggregated vehicular counts). Capacity/flow metrics can be derived using these datasets.
  - Regional Level: The dataset contains information characterizing an entire region, for example, the number of trips made, of inflowing and outflowing vehicles, or of households in a region.

- Environment Level: The aggregated number of trips made in a neighborhood or surrounding. This dataset contains aggregated vehicular counts, which can provide information on trips attracted to and produced in that environment.
- Origin-Destination (OD) Level: The number of trips (or vehicular counts) made across an OD pair.

Each of the above unit levels can be integrated with traditional data sources. ICT data that is at a Micro-Human Level is disaggregate data. Each "response" categorizes what a single individual (or a group of individuals, if they are all performing an activity/trip/event together) has done. Micro-Human-Level ICT data can be collected from an app, mobile phone records, smart card data, and other similar ICT sources. Some Micro-Human-Level ICT data includes *identifiers* (such as an email address, residential address or zip code, or name). When accompanied by these *identifiers*, ICT data is easier to link with other traditional data sources (such as a household travel survey) from the same (or a similarly located) individual.

Next, Micro-Vehicle-Level data is typically collected through a connected smart vehicle (such as a Tesla, which records all geospatial or temporal routes and trips the vehicle takes on a user's and the company's app, so the vehicle is passively recording data, not the user), a public bus, an Uber ride, or a taxi. Typical traditional datasets that are used in certain phases of TDM are either at the Micro-Human or Macro unit level (such as physical and operation features of a road that are traditionally provided for OD matrices); therefore ICT data that is at a Micro-Vehicle Level is much less useful as a complementary and supplementary input alongside traditional datasets because it is more difficult to merge with traditional and regional demographic data. Micro-Vehicle ICT data is still valuable, especially when it can be aggregated to identify or calibrate popular routes and paths taken in an OD matrix.

Finally, Macro-Level data is typically collected at a road, regional, or OD level. Much of the large-quantity Macro-Level data is collected by traffic counters (which are not typically categorized as an ICT data source, unless their data is processed with other ICT information, as is done by StreetLight and INRIX) or by other passive collection methods. In this case, data originally collected at the Macro Level may provide more generalized information about a region, such as the number of cars passing through the region over a specific time period or the count of the number of events occurring at a given location. Because the data is collected at the aggregate level, Macro-Level ICT sources usually do not provide demographic data, though a general demographic description of the geospatial region can be acquired by linking to traditional census data. It is also common for Micro-Human and Micro-Vehicle data to be aggregated by geospatial region or a road segment to create a Macro-Level data source.

2.3. Converting and Translating ICT Data Sources: Four Fusion Methods

The data sources input in the different TDS steps reviewed in the previous section are valuable to the TDM process. Indeed, the travel and behavior information and patterns collected through a traditional household travel survey are irreplaceable, and similar data collected through ICT cannot entirely supplant these data sources. This is because many ICT data sources lack "complete" information. A complete source contains demographic and geospatial information along with trip-related information (e.g., trip purpose, mode choice). ICT sources that have some but not all of these types of information can be used to enrich traditional survey data and improve modeling results. For example, social media can be used to extract demographic data via personal information that is publicly available on users' profiles (such as email address, age, gender, and even names). This type of data represents a reasonable percentage of the population, but it raises

privacy issues. Some other sources, such as smart card data, implicitly provide mode information. Trip purpose, on the other hand, generally must be self-reported, though it may be possible to overlay contextual information related to the trip, such as location, to a land-use geographic information file to impute this. In some cases, the timing, duration, and frequency of a trip, coupled with location information, may be used to identify it as a work-purpose trip.

Overall, adding demographic and other contextual data to aggregate geospatial and temporal information tends to be time-intensive and less accurate than current TDS collection methods. While there are limitations to integrating ICT data that do not contain demographic or other trip data, systematic methodological and database techniques may facilitate such a fusion process. In order to begin the process of integrating ICT sources alongside TDSs, a discussion of how to exactly do so is essential. Based on the identification and assessment of the context, content and unit level of the data included in each ICT source was performed, four clear fusion methods arose to convert, translate, and finally fuse any ICT data sources with TDSs and used within the TDM process.

After much consideration, a full assortment of conversion, translation, and integration strategies may be grouped into four main strategies, each of which is assessed along a rating scale[2], which rates the level of intrusiveness and cost of integrating an ICT data source with TDS within a TDM context. The four fusion strategies include:

- **Aggregated Merge** (also referred to as Merge 1): An aggregated dataset (or aggregated traffic count data – for example, the total number of vehicles on a road across the period of an hour, rather than details about each individual vehicle) at the Regional Level is merged with regional census data to relate region-wide demographics to trip attributes (attractions and productions) or travel demand. The aggregated trip attributes can also be used to validate and calibrate the models already produced using TDS.
    - Because this ICT data is at the Macro or Micro-Vehicle Level or lacks identifying demographics, it must be aggregated before it can be integrated with TDS and used in TDM. The ICT data responses or events/activities/trips can be grouped by their geospatial qualifiers—either following census blocks or different zonal areas, depending on the grouping of the TDS they are being merged with—and aggregated across the chosen area. Conversion (through aggregation) and integration can be done through GIS or similar software, generating spatial and information links between large volumes of data. A visual of the Merge 1 process can be found in Figure 2.
    - Investment rating
        - Time: Aggregating disaggregate data and linking it to TDS can be accomplished **quickly** through a database management software (DBMS) such as SQL or geospatial software such as GIS. Therefore, the time investment is relatively low.
        - Capital: Capital investment is **low**, though it depends on the cost of the ICT data source (more data and more data context typically cost more).

[2] Of course, any mapping of these three strategies (conversion, translation, and integration) to a general fusion method will involve generalizations and qualitative judgments. An accurate rating would depend just as much on the quality of the specific data in an ICT source as on the integration method. But our framework should provide a general sense of the potential value of any ICT data source's integration into any TDM process. The assessment criteria employed here include (1) the relative time frame for accurately translating, converting, and integrating the ICT data, (2) the capital cost to acquire, convert, and integrate the ICT data, and (3) the amount of additional effort, thought, strategizing, and labor required to complete the translation, conversion, and integration.

- Effort: Aggregating data and linking it to TDS requires **similar effort** as would be required for linking two TDS with one another. Therefore, the effort investment rating depends on the common techniques already performed on the TDM inputs and if they are compatible with the available ICT data source formats.

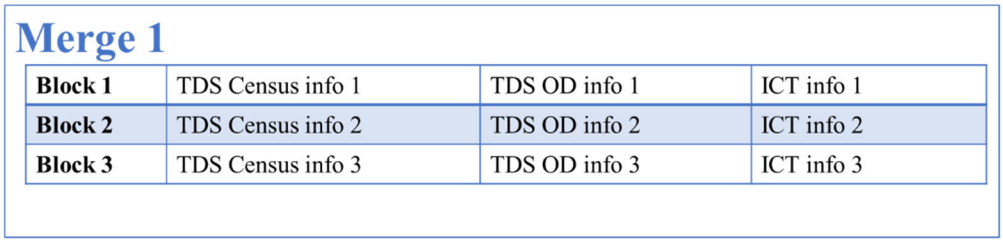

Figure 2 Integration Strategy: Merge 1

- **Demographic Merge** (also referred to as Merge 2): The ICT datasets are merged (or joined) with Individual- or Household-Level data (collected through household surveys) using *identifiers* (e.g., phone number, email address, last name, or street address). Trip data passively collected through ICT sources is more accurate than (and can supplant) stated trip details, potentially improving the accuracy of resulting TDMs. A visual of the Merge 2 process can be found in Figure 3.
  - When accompanied by *identifiers* or detailed demographic data (such as age, gender, or other detailed individual or household data that can categorize a "type" of respondent), each response or activity/trip/event in the ICT data can be merged with the respective respondent or respondent "type" in the TDS. The ICT data response is added to the same storage row as the related TDS response in order to augment that individual's response with supplementary information. An example in shown in Figure 3, which also illustrates the difference between **Merge 2** and **Add**. This process can be performed with any DBMS such SQL, Excel, or Microsoft Access, depending on data volume.
  - Investment rating
    - Time: **If the same *identifiers* are present** in both the TDS and ICT data, then the link is easy and takes **minimal time**. **If generalizations need to be made** about sociodemographic groups (e.g., if only gender or age are available) in the ICT data, then **more time** may be required to integrate the source.
    - Capital: Capital investment is **low**, though it depends on the cost of the ICT data source (more data and more data context will typically cost more).
    - Effort: Similar to the time investment, the effort required for this Integration Strategy depends on the format of the ICT data, as well as the best approach to convert the data (to strategically aggregate or disaggregate the data so that it can be linked with the TDS through a demographic or identifier link). This effort **may be minimal or weighty**, depending on the format and conversion method required. If the analyst believes it will require significant effort and time, then they may determine that the investment is not worth it.

Merge 2

Person 1	TDS info 1	ICT info 1
Person 2	TDS info 2	ICT info 2
Person 3	TDS info 3	ICT info 3

Figure 3 Integration Strategy: Merge 2

- **Demographic Add** (also referred to as Add): If the ICT data source cannot be merged with TDS using *identifiers*, but it already includes sufficient demographic information, it can be used independently to model travel demand. However, trip data collected by the ICT source may be incomplete, or it may reflect only a specific type of mode or trip, in which case modelers must develop separate models. A visual of the Add process can be found in Figure 4.
    - If sufficient demographic data is available in an ICT source, then the Integration Strategy may be to simply add the ICT data to the TDS as a new response, instead of merging it with the existing responses. For example, if the ICT is an app such as Strava that collects detailed information on every bike ride, walk, or run users in an area take, then the data may just be added to the responses that individuals in that area reported in a TDS such as a household travel survey. See Figure 4 for a visual of the **Add** Integration Strategy. However, the analyst must be careful when performing this strategy to avoid introducing errors. The ICT dataset may include a trip/activity/event that the TDS has already recorded or an individual who has already been included in the TDS. This may cause the dataset to be overbalanced toward groups that participate more in crowdsourcing data collection efforts.
    - Investment rating
        - Time: If the ICT data source is comprehensive, then it can simply be added to a TDS, an **immediate** process. If the ICT data source is missing a considerable amount of responses or demographic information or upon evaluation is not compatible with the TDS (e.g., it only measures bike routes, while the TDM is of car travel), then a **significant amount of time** may be required to reformat and convert the data, or the analyst may avoid this Integration Strategy altogether.
        - Capital: Acquiring data sources that include both demographic and detailed event data is usually **more costly** than other ICT data sources because it typically costs more to collect and store more data.
        - Effort: Similar to the time investment, the effort required for this Integration Strategy depends on the format of the ICT data when collected and how many conversions the DBMS must perform to make the data compatible with the TDS. This effort **may be minimal or may be very time consuming**.

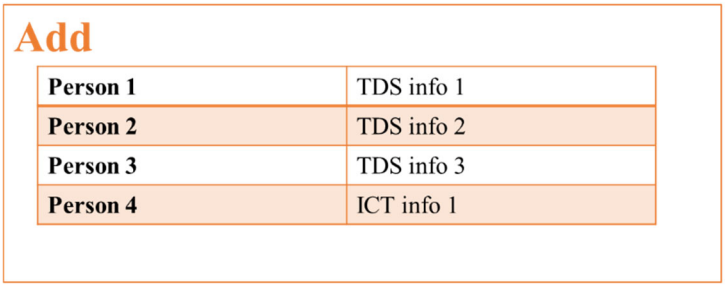

Figure 4 Integration Strategy: Add

- **Independent Use:** Some very efficient ICT data sources might include significant demographic data in addition to trip details. These ICT data sources can be used "independently" (i.e., without integration with TDS) for TDM.
    - In an ideal scenario, ICT data sources contain complete demographic and trip information, which allows them to be used independently (i.e., without any TDS). These data sources, while rare, provide a very large sample and hence very accurate models. In some cases, where the data focuses on a certain Mode or trip type (e.g., Uber rides), we can derive models for such trips by Mode (vehicle type) or Purpose, perhaps distinguishing between commute and non-commute trips by sorting through origins and destinations. In practice, these data sources can be used very similarly to TDS. They can be used to build parallel models, compare with and validate traditional processes, and better calibrate already-existing models.
    - Investment Rating:
        - Time: These datasets, if acquired in raw format, might **require significant time** to process, clean, and filter the useful data. If the data provided is already clean, the process should be **quicker** and similar to traditional processes. The time required will then depend on the amount of data the analyst desires to use.
        - Capital: This type of data can be **very expensive** to collect, as it contains many variables for many respondents. Such collection requires sophisticated technology, which can involve many challenges. Storage of such large-scale data is also expensive.
        - Effort: If the data is acquired in raw format, this Integration Strategy will **require significant effort** for pre-processing and cleaning and might require specialized data engineers. However, if acquired in a usable format, this data can be **used immediately** for TDM, without being integrated with any other data sources.

2.4. Integration Framework

Based on 1) the content of ICT data sources (the cascading three-level structure of main context, additional content, and information unit level of data) and 2) the four fusion strategies, an integration framework was developed in order to map the characteristics of a given ICT data source to one (or more) of the methods of fusion. The framework flowchart, shown in Figure 5, provides guidelines of the decision making process to project which fusion strategy or strategies would work best with each data source.

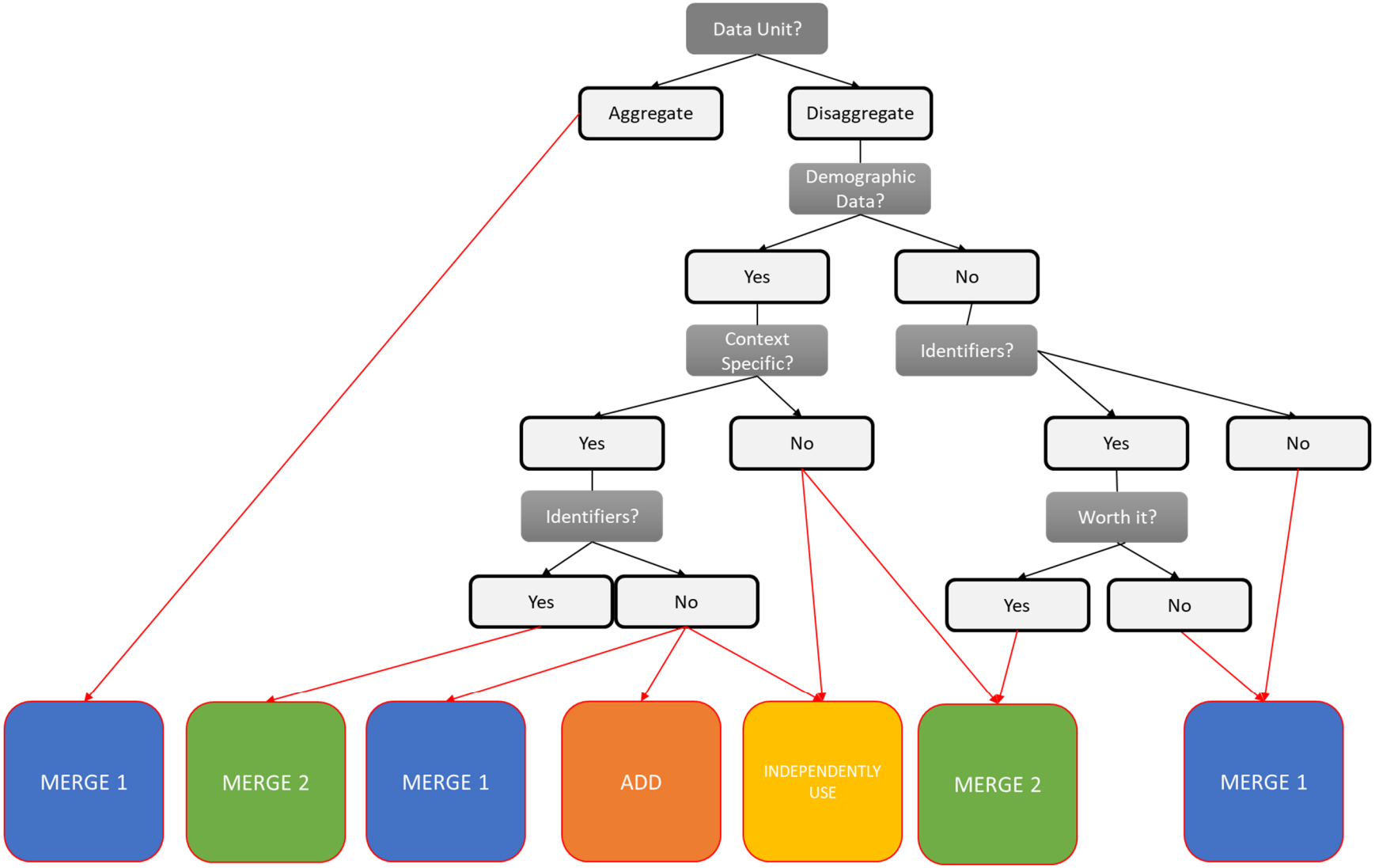

Figure 5 Mapping ICT Data Source to Fusion Method

To use the Integration Framework, select a specific ICT data source and work from the top to the bottom, following the arrow to the appropriate response to each question. The proposed framework broadly categorizes ICT data sources and recommends for each category a strategy to integrate the ICT and traditionally sourced data. In all cases, integration requires in-depth consideration of the data formats; thus we have proposed a broad, higher-level framework that should encapsulate all possible types of data sources, including both ICT data sources and TDS.

2.5. Examples of Framework Use

Next, we review two ICT sources from different ICT Classes and show how the mapping framework may be applied to determine the "best" fusion method.

The first example is for app data. Figure 6 visualizes the process of applying the mapping framework, with the purple checkered boxes indicating the flow of questions and responses followed by the app data source, concluding with the Integration Strategy **Merge 2.** App data is disaggregated and collects information on all users. This usually includes demographic data. Apps are deployed to provide a service for a specific group of users, and therefore they have a specific context. When signing up for an app, a user typically has to create a profile, where they supply some sort of identifier, such as an email address or a name.

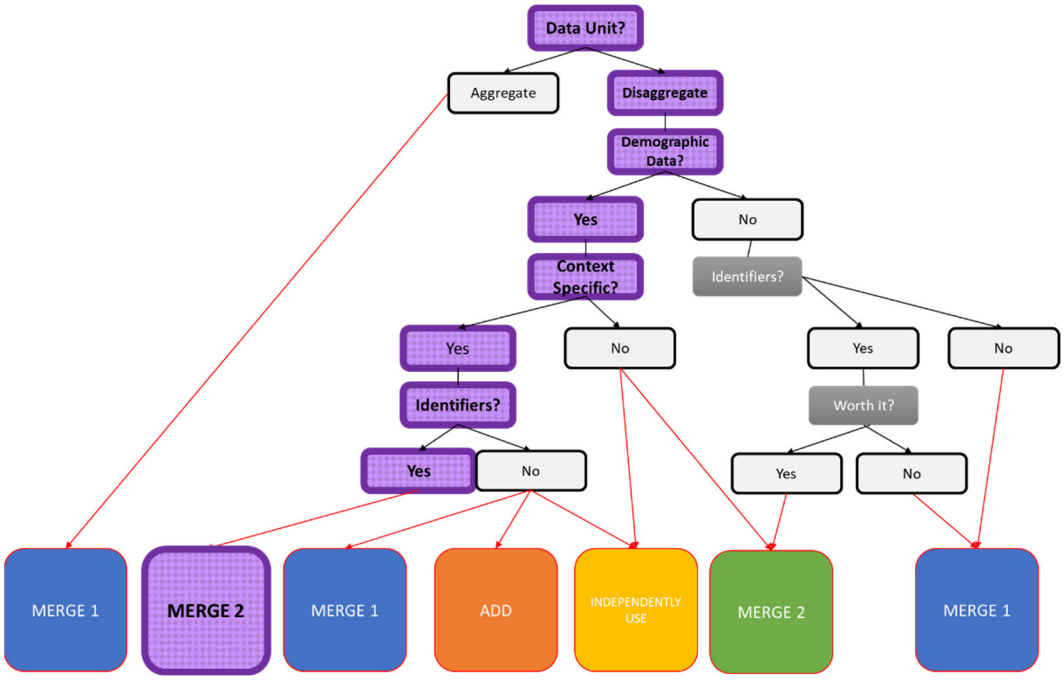

Figure 6 Example of ICT App Data Source Integration Flow

The second example uses social media. Social media data is a special case. If large amounts of time and effort are invested in data mining and linking profiles, it can be disaggregated and a user and their demographics can be identified and linked to a given Tweet or Instagram post, which already contain geolocation. If an analyst is willing to undertake this investment, then the social media data source will follow a flow similar to the app data source, concluding with Merge 2. If the analyst will not be performing this disaggregation, as is more common with social media sources, then the data is aggregated over a geospatial region. This allows for identification and categorization of the type or cause (such as a car crash causing Twitter users to tweet about being stuck in traffic) of traffic or travel on a given road or in a given area. In this case, the Framework leads to the Integration Strategy **Merge 1** (see the dark blue checkered path in Figure 7),

recommending that the analyst integrate the aggregated temporal and geospatial data that has been gathered regarding a specific event or travel/traffic purpose with TDS that describe the demographics or normal travel patterns of that geospatial area.

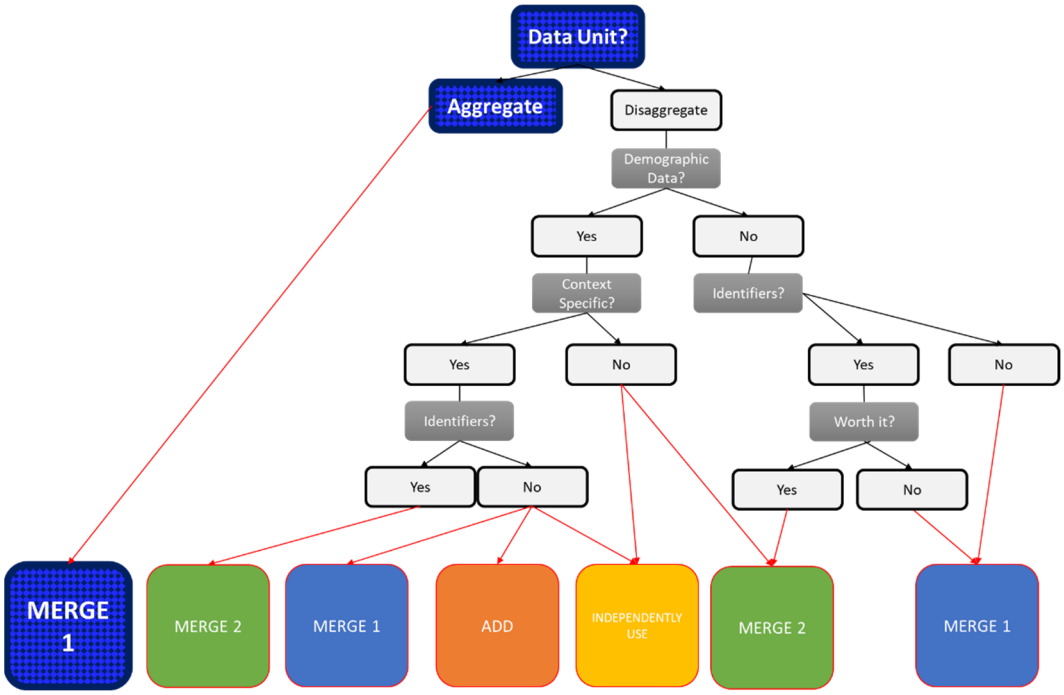

Figure 7 Example of ICT Social Media Data Source Integration Flow

2.6. Performance Assessment of Proposed Framework Regarding Prediction and Application Capabilities

Each ICT fusion strategies can be assessed along the following four performance assessment indices:

- **Cost and Time Efficiency Index (CTI):** This index rates the efficiency of cost (capital) investments necessary for the integration strategy, such as those for obtaining, storing, and managing the data, as well as the time and effort required. It can give travel demand modelers insight about the practical feasibility of the strategy given their budget and time constraints. A higher CTI indicates higher cost efficiency, while a lower CTI indicates that greater capital and effort investment is required, therefore less efficiency.
- **Completeness and Coverage Index (CCI):** This index indicates to what extent the integration strategy results in more complete data in terms of spatial and temporal coverage. This helps researchers determine if a particular integration strategy is more beneficial for TDM analysis purposes and if it can be used for large-scale planning purposes. A higher CCI indicates large-scale, generalizable data, while a lower CCI means the data has limited coverage and/or accuracy.
- **Resourcefulness and Information Index (RII):** This index represents the extent of the "additional information" (such as trip purposes or details about the driver) obtained by combining ICT and traditional data sources as compared to simply using the latter. This index also connotes the usefulness of the particular ICT source in the context of policy making. A higher RII indicates that a fusion strategy offers significantly more information than traditional methods alone, while a lower RII indicates that the addition of the ICT source does not make a significant difference.

- **Closeness to Reality Index (CRI):** This index represents how close the combined available information in the ICT dataset is to reality. Additionally, it indicates whether the modeling approach using the ICT data is better suited for short- or for long-term planning solutions. A higher CRI indicates that the strategy yields a more realistic model, which could be used for long-term planning purposes, while a lower CRI indicates that the fusion strategy is better for short-term or immediate planning.

It is important to note that using these metrics to assess a general fusion framework will involve generalizations and qualitative insights. The actual rating will depend as much on the ICT source and its data quality as on the integration strategy. However, we here aim to provide a broad and general sense of the four main fusion strategies. Each index will be rated with *low*, *medium*, or *high*. The reader is referred to Table 2 for a visualization of the final performance assessment for each fusion strategy.

Table 2. Review of Integration Framework Assessment

Framework Performance Assessment Criteria	Fusion Strategies			
	Aggregated Merge	**Demographic Merge**	**Demographic Add**	**Independent Use**
Cost and Time Efficiency Index (CTI)	High	Low	Low	Low
Completeness and Coverage Index (CCI)	Medium	High	Medium	Medium
Resourcefulness and Information Index (RII)	Medium	Medium	High	High
Closeness to Reality Index (CRI)	Medium	Medium	High	High

2.6.1. Assessment of Aggregated Merge (Merge 1)

In this approach, micro- or macro-level traffic count data is first aggregated and then merged with the TDS and in the TDM process. The data are aggregated spatially (according to census block groups) as well as temporally (to hourly demand or daily demand) and then merged with census data and/or travel surveys to obtain broader, more accurate results. The efficiency of this approach can be represented by the following metrics:

- ***CTI**:* The fusion of TDS with aggregated and merged ICT data can be performed relatively quickly with minimal capital investment (depending on the ICT source). The effort required is also not significantly higher than for traditional methods. The resulting dataset does not require significant storage once aggregated and can be efficiently used for TDM. Hence, the CTI for this strategy: **HIGH**.
- ***CCI**:* Because the data lacks demographics and relies on aggregation of traffic counts, the completeness of this approach could be limited. The coverage depends on the depth and density of the data collected by the ICT source. The CCI for this strategy: **MEDIUM**.
- ***RII:*** The "additional information" provided by this strategy depends on the density and accuracy of the data collected through the ICT source. Moderate resourcefulness can be anticipated from Merge 1 in the context of policy making, as it offers a more accurate reflection of actual travel, and hence could be relied upon to make better decisions. Therefore, this approach is an improvement over traditional methods for TDM. The RII for this strategy: **MEDIUM**.

- ***CRI:*** The closeness to reality of the developed model using the ICT data source will depend on the quality of the census or travel survey data and the amount of data collected by the ICT source. As the collected demographic data will not be representative enough for a very large area, adding the more accurate observed travel demand for calibration could balance that. This strategy could be used for medium-term planning purposes, as it represents moderate closeness to reality. The CRI for this strategy: **MEDIUM**.

2.6.2. Demographic Merge (Merge 2)

This approach involves merging an ICT source to a TDS with the help of *identifiers*. It could provide additional information about a particular "type" of individual by fusing some additional sociodemographic variables with trip details. This strategy could be evaluated as follows:

- ***CTI:*** This approach would require more time and effort, as it involves matching *identifiers* with individuals, as well as generalizing individuals on the basis of sociodemographics. The cost depends on the amount of data captured by the ICT source, but generally obtaining disaggregated data is expensive and will also lead to higher storage costs (more data volume means larger storage size requirements). The CTI for Merge 2: **LOW**.
- ***CCI:*** The kinds of ICT sources appropriate for Merge 2 contain demographic *identifiers* and thus will provide more "complete" information. However, the coverage will depend on the depth and quality of data collected. The CCI for Merge 2: **HIGH**.
- ***RII:*** The amount of additional information provided by the ICT source in this case is clearly higher than in Merge 1, as it enables modelers to relate the trip details and attributes (such as trip purpose) at an individual level. This would allow them to build a model with a larger sample size (as compared to generalized census-level data). However, not all data points collected by ICT will be useful, as demographics won't necessarily correspond to all *identifiers*. The RII for Merge 2: **MEDIUM**.
- ***CRI:*** This approach is expected to result in more accurate travel demand models, as it covers a larger sample that could explain more variance in travel demand and travel behavior in general. The presence of *identifiers* would help modelers analyze travel behavior at a disaggregated level. The level of reality represented will of course depend on the quality of the ICT data source, but the merged dataset is expected to yield better results than TDS alone in most scenarios. Keeping all that in mind, the CRI for Merge 2: **MEDIUM**.

2.6.3. Demographic Add (Add)

If an ICT dataset includes demographic data but cannot be used independently, as it covers only a part of the area's population, it can be rendered usable by appending it to traditional data as additional data points. This could help enlarge the sample and better represent elements of the population that could be missed by traditional survey methods. The efficiency of this approach is assessed by the metrics below:

- ***CTI:*** If the ICT source is in the same format as the TDS, this approach does not require significant time. The cost is more likely to be on the high end, as modelers need more variables from the ICT source. The CTI for Add: **LOW**.
- ***CCI:*** This approach provides additional data points to already-existing traditional household survey data. The completeness depends on the number of variables obtained by

the ICT source, but in general this method enlarges the sample and can potentially capture unreported trips. Hence, the CCI for this strategy: **MEDIUM**.

- ***RII:*** The ICT sources that are appropriate for this strategy can greatly enrich traditional datasets. They significantly increase the sample's overall information level, which leads to better-informed travel demand models. The additional information these data sources can provide to policy makers also increases, as the data points obtained from ICT cover a vast geospatial area, and sources can be aggregated or disaggregated as needed to address the issues of interest to policy makers. Hence, the RII for Add: **HIGH**.
- ***CRI:*** The CRI of this method can reasonably be assumed to be higher than for Merge 1 and Merge 2, as the data points obtained by ICT are "complete" and have accurate recordings of sociodemographic variables and trip attributes, with minimal discrepancies. The models built through this method will be based on a mix of traditional and ICT data points and will provide a broader overall picture of travel behavior. This strategy can be safely used for medium- to long-term planning purposes. Overall, the CRI for Add: **HIGH**.

2.6.4. Independent Use:

In an ideal case, an ICT source can be used independently, i.e., without merging it with TDS. This can only be done if the ICT source contains complete demographic information along with the trip attributes and covers a reasonable sample size. More often than not, ICT data sources used independently have a "constant collection" or real-time aspect to them, allowing use of the data for more than just TDM purposes. This approach can be useful to make travel demand models for specific mode types. For example, Uber can efficiently build a travel demand model for ride hailing trips using its own data, as it has both demographic and trip attributes. This method often leads to other challenges, though, which are analyzed below, along with the metrics ratings:

- ***CTI:*** The cost of this type of data is typically very high, and dealing with such a volume of data and making it usable also requires expertise. If obtained in raw format, this type of data will require significant cleaning and pre-processing before TDM use. The overall CTI for this approach: **LOW**.
- ***CCI:*** The kind of source necessary for this method typically must be complete. However, coverage varies from source to source. Although sources eligible for Independent Use can be good tools to model particular types of trips (typically mode-based; e.g., bike or ride-hailing), they often do not provide data that covers overall trips for a large part of the population. This strategy's CCI rating: **MEDIUM**.
- ***RII:*** ICT datasets that contain complete demographic and trip characteristics are expected to provide far better insights compared to TDS, due to their greater volume and accuracy. When ICT sources that continuously collect and pass on data are used, it requires planners to quickly adjust and improve their models. This can be a great tool for increasing the dynamism and responsiveness of transportation policy and planning. The RII for this strategy: **HIGH**.
- ***CRI:*** Data collected through ICT sources are very close to reality as they are directly collected from the field, which minimizes errors. Also, if the acquired ICT data is continuously updated, it will maintain this closeness to reality throughout the modeling and calibration process. If historical data is also acquired, it can be used for comparison and to study changes in demographics and travel patterns; when aggregated appropriately, this

can provide important insights. This strategy can be used very efficiently for both short- and long-term planning purposes. Hence, the CRI for this strategy: **HIGH**.

3. CONCLUSIONS

In this paper, we have proposed a flexible guideline framework that is adaptable and customizable to the amount of information in the source. The proposed framework broadly categorizes ICT data sources and recommends strategies to integrate ICT-based and traditionally sourced data. Of course, we should emphasize here that, to understand urban activities, it is insufficient to identify statistical travel patterns of human beings (the emphasis of emerging big-data sources), while ignoring the purpose and content of their activities (more the emphasis of traditional household travel survey data collection programs). Data collected on "where" and "when" individuals travel must be supplemented with the "whys" and "with whoms". Besides, the majority of ICT data sources require large amounts of time, capital and effort to collect, process and organize the raw data. And, in cases where data has already been processed and organized, as is data obtained from third-parties, such as StreetLight or Inrix, there is limited information on how specific travel-related parameters have been estimated (such as daily VMT or population numbers). This leads to much third-party-processed ICT data to essentially be a "black box" to modelers. But, overall, the shift in data collection is not so much a traditional-versus-emerging data debate as much as it is a means to best harness the relative strengths of the different data sources to provide the most complete picture of human and vehicular movements, and in as efficient, comprehensive, and effective manner as possible. If integrated effectively, ICT data can prove valuable for urban and rural transport planning and design, informed safety assessments and crash countermeasure measures, infrastructure design and maintenance, and travel demand management.

ACKNOWLEDGEMENTS

This research was partially supported by the U.S. Department of Transportation through the Center for Understanding Future Travel Behavior and Demand (TBD) (Grant No. 69A3552344815 and No. 69A3552348320), and by the Texas Department of Transportation (TxDOT) under project 0-7134. The authors would like to thank Wade Odell, David Freidenfeld, Farideh Dassi, James Kuhr, Jeremy Rogers, Cleo Williams, and Catherine Wolff of TxDOT for their support and constructive advice throughout this project. The authors are grateful to Lisa Macias, Sarah McGavick, Gina Blazanin, and Natalia Ruiz Juri for help in framework development, and Lisa Macias for additional help in formatting this document.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: K.E. Asmussen, V. Verma, C.R. Bhat; data collection: K.E. Asmussen, V. Verma, C.R. Bhat; analysis and interpretation of results: K.E. Asmussen, V. Verma, C.R. Bhat; draft manuscript preparation: K.E. Asmussen, V. Verma, C.R. Bhat. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Cherchi, E. and Bhat, C.R. (2018). Data analytics and fusion in a world of multiple sensing and information capture mechanisms, 11th International Conference on Transport Survey Methods. *Transportation Research Procedia*, *32*, 416–420.