UTTRI

Research Report

The State of the Art in Urban Transportation Data Collection

Report 1, iCITY-SOUTH

Eric J. Miller, Khandker Nurul Habib

September 2017

iCITY-SOUTH:
Urban Informatics for Sustainable Metropolitan Growth in Latin America

REPORT 1:
THE STATE OF THE ART IN URBAN TRANSPORTATION DATA COLLECTION

A report to CAF, the Development Bank of Latin America.

CAF BANCO DE DESARROLLO DE AMÉRICA LATINA

Más oportunidades, un mejor futuro.

By:

Eric J. Miller, Ph.D.
Professor, Department of Civil Engineering
Director, UTTRI

Khandker Nurul Habib, Ph.D., P.Eng.
Associate Professor, Department of Civil Engineering

September, 2017

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

UTTRI

EXECUTIVE SUMMARY

In 2015, the University of Toronto Transportation Research Institute (UTTRI) launched the *iCity* research program, which is dedicated to applying modern *urban informatics* (the combination of data collection, data science, modelling, visualization and high-performance computing methods) to the promotion of sustainable metropolitan growth. As one component of CAF's strategy for promoting its urban sustainable mobility objectives, it has partnered with the UTTRI to create the *iCity-South* research program to apply the *iCity* urban informatics vision and capabilities in Latin American cities. One project within the iCity-South research program is investigating new data collection methods in Montevideo, Uruguay. This report is the first in a series of reports documenting the findings of the Montevideo project.

This report presents a summary of the current state of the art in urban transportation data collection methods. It provides the starting point for subsequent reports which investigate specific data collection methods, notably: traditional household surveys, transit smartcard data, cellphone data (cellular data records, CDR); among other methods. These reports, in turn, provide the basis for recommendations concerning a comprehensive data collection program for Montevideo, as a prototype for eventual application in other Latin American cities.

Chapter 2 introduces key design concepts and issues that underlie all data collection methods. Technical issues discussed in include: definition of terms, sampling procedures, sample size determination, sample recruitment, various temporal considerations and questionnaire design. Chapters 3, 4 and 5 discuss the three primary methods for collecting transportation data: household-based surveys, choice-based surveys, and a broad range of technology-based, passive data collection methods, respectively. Household-based survey methods discussed in Chapter 3 include: face-to-face, mail-back, telephone and web-based surveys. Choice-based survey methods discussed in Chapter 4 include roadside intercept surveys, transit ridership surveys, and place of employment or school surveys. ICT-based data collection methods discussed in Chapter 5 include cellular data records, GPS traces, smartphone apps, a variety of sensors (both fixed and portable), transit smartcard data, other mobility service usage data and third-party passive data streams.

Chapter 6 presents an overview of the core-satellite approach to integrating multiple data collection methods which is recommended for designing an overall data collection program. It also introduces the data science-based data fusion methods needed to implement the paradigm. The core-satellite approach to data collection is a very flexible, efficient approach to making best use of a range of data collection methods within a multi-instrument design. It consists of the following components:

- A *core survey*, which is a large-sample survey which gathers primary information concerning the respondents and their key behaviours.
- Any number of *satellite surveys*, which are smaller-sample, more focussed surveys (or other data collection methods) designed to gather more detailed information about specific behaviours of interest.
- Additional, independent, *complementary* surveys/datasets that might be used to augment the core-satellite database, but may not be directly linkable to the core-satellite data.

Chapter 7 concludes the report with a discussion of the special issues involved in the application of the methods discussed in this report in the Latin American context. These include: population socio-economic heterogeneity; informal/privatized transit services; and government structures and capacities. It also presents a short "look ahead" to next steps in the process of developing a data collection process for Montevideo and CAF.

This report, however, does not attempt to construct such a recommended data collection program. Instead its objective is to provide a description and assessment of the current and emerging state of the art in transportation data collection. This assessment, along with the subsequent reports in this series dealing with in more detail with specific datasets and data collection methods used within Montevideo, will provide the basis for the development of a recommended program in the final report of this project report series.

ACKNOWLEDGEMENTS

This study was funded by CAF. The unstinting support and patience of CAF, and, in particular, Nicolas Estupiñan and Andres Alcala has been greatly appreciated. The authors would also like to thank their colleagues and students at the University of Toronto, Catherine Morency (Polytechnique Montréal), and Martin Lee-Gosselin (Université Laval), whose ideas have influenced the findings of this report in many ways.

TABLE OF CONTENTS

TABLE OF CONTENTS, CONT'D

LIST OF FIGURES

CHAPTER 1
STUDY PURPOSE & MOTIVATION

Urban regions with Latin America (and elsewhere) face enormous challenges in terms of the provision of transportation infrastructure and services to meet the travel needs of their growing population in a cost-effective, equitable and sustainable manner. High quality, comprehensive information concerning travel behaviour and transportation system performance is a fundamental prerequisite for successful urban transportation planning and decision-making to address these pressing, first-order needs.

In recognition of this need, CAF established the Urban Mobility Observatory (OMU, *Observatorio de Movilidad Urbana*)[1] to assemble and utilize standardized transportation-related data for Latin American cities. 29 cities are currently members of OMU. Collecting consistent, time-series data for these cities, however, is a difficult and costly task for CAF and its partner cities.

At the same time, exciting, new transportation data collection sources are emerging to complement or even replace the traditional methods used to collect the OMU data. These include:

- The pervasive penetration of cellphone and smartphone technology within urban populations.
- The widespread adoption of smartcard systems by public transit agencies in many cities.
- Extensive deployment of many types of sensors (video, thermal, Bluetooth, etc.) for monitoring travel flows.
- Increasing availability of very large (typically crowd-sourced) datasets collected in a variety of ways by private sector companies (Google, Waze, Inrix, etc.) that can provide travel information.
- Web-based survey methods to complement/replace traditional survey methods such as home-interviews, telephone interviews, etc.

In 2015, the University of Toronto Transportation Research Institute (UTTRI) launched the *iCity* research program, which is dedicated to applying modern *urban informatics* (the combination of data collection, data science, modelling, visualization and high-performance computing methods) to the promotion of sustainable metropolitan growth. As one component of CAF's strategy for promoting its urban sustainable mobility objectives, it has partnered with UTTRI to create the *iCity-South* research program to apply the *iCity* urban informatics vision and capabilities in Latin American cities.

Two initial projects were chosen to launch the *iCity-South* research program. One involves the demonstration of agent-based microsimulation methods for modelling urban travel demand in terms of developing a prototype microsimulation model for Asunción, Paraguay.[2] The second is investigating traditional and new data collection methods in Montevideo, Uruguay. This report

[1] https://www.caf.com/es/temas/o/observatorio-de-movilidad-urbana/

[2] This project was completed in April, 2017. See Miller, et al., (2017a, 2017b) for the results of this project.

is the first in a series of reports documenting the Montevideo project results. This report presents a summary of the current state of the art in urban transportation data collection methods. It provides the starting point for subsequent reports which investigate specific data collection methods, notably: traditional household surveys, transit smartcard data, cellphone data (cellular data records, CDR); among other methods. These reports, in turn, provide the basis for recommendations concerning a comprehensive data collection program for Montevideo, as a prototype for eventual application in other Latin American cities.

This report draws heavily on two other studies. The first is a comprehensive review of urban data collection methods undertaken by the authors for the Transportation Association of Canada (TAC) in 2011-12 (Miller, et al., 2012). The second is a study led by the authors currently underway to develop a new, comprehensive data collection process for the greater Toronto region (Habib, et al., 2017).

In addition to this brief introduction, this report consists of 6 chapters. Chapter 2 introduces key design concepts and issues that underlie all data collection methods. Chapters 3, 4 and 5 discuss the three primary methods for collecting transportation data: household-based surveys, choice-based surveys, and a broad range of technology-based, passive data collection methods, respectively. Chapter 6 presents an overview of the core-satellite approach to integrating multiple data collection methods within an overall data collection program, and the data science-based data fusion methods needed to implement the paradigm. Chapter 7 concludes the report with a discussion of the special issues involved in the application of the methods discussed in this report in the Latin American context, as well as a “look ahead” to next steps in the process of developing a data collection process for Montevideo and CAF.

This report, however, does not attempt to construct such a recommended data collection program. Instead its objective is to provide a description and assessment of the current and emerging state of the art in transportation data collection. This assessment, along with the subsequent reports in this series dealing with in more detail with specific datasets and data collection methods used within Montevideo, will provide the basis for the development of a recommended program in the final report of this project report series.

CHAPTER 2
DEFINITIONS & BASIC CONCEPTS IN SURVEY DESIGN

2.1 INTRODUCTION

The design of any data collection program is a very challenging technical problem, requiring rigorous applications of sound statistical concepts within a myriad of practical implementation problems, typically within strict budget constraints. Transportation data collection is particularly difficult, given the high dimensionality of the problem: we are interested in understanding travel behaviour in both space and time by highly heterogeneous trip-makers travelling for a wide range of purposes and under a variety of constraints and motivations. This chapter presents an overview discussion of key, basic concepts in survey design that apply to any data collection effort. An understanding of these basic principles and issues is fundamental to developing a successful data collection program.

The purpose of any survey (or other data collection method) is to estimate one or more *parameters* of the population of interest, where these parameters might be population attributes (average income, labour force participation rates, auto ownership rates, etc.) or behaviours (travel mode splits, home-based trip rates, etc.). The true values of such parameters can never be known with certainty (without surveying the entire population). An *unbiased* sample estimate is one whose *expected value* is the true population parameter value (i.e., if the survey was run many times, the average of the estimates obtained would equal the true population parameter value). An *efficient* sample estimate is one whose distribution (variance) around its expected value is as small as possible, so that any one estimate (which is all one generally has in practice) can be expected to lie "close" to the true value. Thus, it is critical that any data collection exercise be very carefully designed so as to ensure that a *representative sample* is drawn that adequately and efficiently characterizes the distribution of attributes and behaviours of the population being observed. In other words, great care must be taken to avoid the introduction of *bias* in the survey results.

Regardless of the application, all survey designs involve the following 11 steps (Cochran, 1977):

1. Establish a clear statement of the survey objectives.
2. Define the population to be sampled and the target groups to be focused on.
3. Identify the specific data that are relevant for the purpose of the survey.
4. Specify the degree of precision required from the survey results (i.e., how much error can be tolerated in the results).
5. Determine the methods to be used in obtaining the survey results.
6. Divide the population into sampling units and list the units from which the sample will be drawn.
7. Select the sampling procedure and the sample size.
8. Pretest the survey and field methods to ensure that the procedures are workable and the survey is understandable.
9. Establish a good supervisory structure for managing the survey.
10. Determine the procedures for analyzing and summarizing the data.
11. Store the data and analysis results for future reference.

Each of these steps is critical to reducing bias (increasing accuracy), increasing efficiency/precision and maximizing the overall effectiveness and usefulness of the survey results. Surveys are inevitably reasonably costly to undertake: it is critical that their utility within the planning and decision-making process is maximized in order to ensure their cost-effectiveness. In particular, it is essential that the survey be designed and executed in a manner that ensures that it does, indeed, provide the data that are needed for the task(s) at hand and these data are as accurate and precise as possible. Survey design elements that need special discussion within this review include the following:

- Population, sampling frame and sampling unit definitions.
- Sampling procedure.
- Sample size determination.
- Interview type.
- Sample recruitment, contact methods and incentives.
- Temporal considerations.
- Questionnaire design.

Each of these issues is defined and discussed in the following sections. Throughout this discussion a key emphasis is on maximizing survey *response rate* (i.e., the percentage of persons selected within the survey sample who actually complete the survey), since the larger the response rate, the less likely it is that the survey data will contain major biases. Keys to high response rates include:

- Minimizing *respondent burden*, i.e., the amount of time and level of effort required for the respondent to complete the survey.
- Maximizing the respondent's motivation to complete the survey.

Building on an understanding of these design issues, Chapters 3, 4 and 5 discuss the wide range of methods for conducting transportation surveys and other data collection procedures. These methods are grouped into three major categories:

- *Household-based* surveys (Chapter 3).
- *Choice-based* surveys (Chapter 4).
- *Technology-based* data collection methods (Chapter 5).

2.2 POPULATION, SAMPLING FRAME & SAMPLING UNIT DEFINITIONS

In survey design a *population* is the set of agents for which inferences are to be drawn from the survey data. Typical populations of interest within urban transportation planning include:

- All individuals living within the study area.
- All households living within the study area.
- All members of a group of special interest (physically disabled persons, the young or the elderly, low-income households, new immigrants within the region, etc.).

The appropriate definition of the population to be surveyed obviously depends on the questions to be addressed by the survey and the information required to answer these questions.

A *sampling frame* is the operational method of (at least implicitly) enumerating/listing the population from which the survey sample is to be drawn. For example, if the population is the set of all households in an urban region, the sampling frame in principle should be a physical list of all these households from which individual households can be selected to interview. In practice, sampling frames are often not literally complete but are assumed to be sufficiently complete and representative (i.e., bias-free) to be useful.

Having cost-effective access to a complete (accurate) sampling frame is an essential component in every survey design. Without a well-defined sampling frame, sample weightings for the construction of population inferences cannot be computed and the representativeness of the sample cannot be guaranteed. In addition to enumerating the population, the sampling frame generally must provide the contact information (telephone number, mailing address, e-mail address, etc.) that is required to contact the selected survey respondents. The inability to construct an appropriate sampling frame is often a major limiting factors in determining transportation survey feasibility.

The *sampling unit* is literally the elementary unit to be surveyed within the sampling frame. If the population consists of households, then the sampling unit is a single household within the sampling frame, if the population consists of employees, the sampling unit is a single employee, and so on. A *weight* (or *expansion factor*) must be applied to each unit sampled that represents the contribution of this unit to the calculation of population attribute values (population totals and averages, etc.). The calculation of these weights depends on the sampling procedure used, discussed below.

By far the most common population of survey interest in urban transportation is the set of households residing within the urban region. Household-based surveys are discussed further in Chapter 3.

2.3 SAMPLING PROCEDURES

For any given sampling frame a number of procedures exist for actually drawing a sample of units (respondents) from this frame. These include (Meyer and Miller, 2001):

- Simple random sampling.
- Sequential sampling.
- Stratified random sampling.
- Cluster sampling.

In all cases, the objective is to achieve an unbiased sample in a cost-effective manner. The "standard" or "base" method is *simple random sampling*, in which the sampling units are literally randomly selected from the sampling frame. Simple random sampling ensures that no bias is introduced into the sampled set through the sampling process. Any of the other sampling procedures listed above are only used when they provide a more cost-effective (efficient) method for achieving a similarly unbiased result. In particular, they may provide a means to reduce sample size (while maintaining precision levels) or to select sampling units in cases in which an explicit sampling frame is difficult to enumerate.

Sequential sampling is a simple variant of simple random sampling in which the sampling frame can be safely assumed to be randomly ordered. In such cases the list can be sequentially, rather than randomly, sampled, thereby generally increasing the efficiency (and simplicity) of the sampling process. Examples of sequential sampling include selecting every n^{th} person arriving at a transit stop to interview or selecting every n^{th} listing in a telephone directory. In all such cases, "n" is the inverse of the sampling rate. E.g., if a 10% sample is required, then every 10th person would be selected.

Stratified sampling permits the controlled and cost-effective over-sampling of important target sub-populations that may be difficult to observe in statistically useful numbers through a simple random sample without requiring excessive sample sizes. In this approach the population is stratified into a set of mutually exclusive and collectively exhaustive categories or strata. Each category is assigned its own sampling rate so as to optimize the number of observations per category required to achieve accuracy and precision targets within each category. Examples of stratified sampling include oversampling transit users in heavily auto-dominated travel markets and the oversampling of small sub-populations of particular policy interest (e.g., physically disabled persons, senior citizens, etc.). In all cases, care must be taken to ensure that sub-population sizes and sampling rates are known for each category so that proper weighting across the categories can be maintained so that correct population totals can be constructed.

Cluster sampling provides a means for cost-effectively surveying sampling units for which a sampling frame is difficult to construct explicitly. For example, place of work surveys generally use cluster sampling, in which a sampling frame of business establishments (BEs) is constructed and a sample of BEs is then randomly drawn. These BEs are clusters of workers (the actual sampling unit of interest), which can be enumerated and sampled (interviewed) within each selected BE. It is also important to note that households are clusters of individuals (each and every individual belongs to exactly one household), and so household-based surveys are implicitly cluster samples of individual persons.

2.4 SAMPLE SIZE DETERMINATION

In general, the precision of population estimates derived from survey samples improves as sample sizes increase. Diminishing returns, however, exist, and there is always a point at which increased sample size is not cost-effective. Standard methods for computing sample sizes exist (see, for example, Appendix B in Meyer and Miller, 2001), but in practice sample size determination is a complicated process, for several reasons, including:

- Surveys generally consist of many questions, each one of which typically will have its "optimal" sample size. Sample size selection is therefore almost always a balance between the competing requirements of different components of the survey.
- Sample sizes are often driven by available budgets. In such cases it must be recognized that the survey precision is being driven by the budget rather than by desired precision targets.

A fundamental three-way trade-off exists in every survey among sample size (which determines survey precision), survey complexity (number of questions, etc., which determines the survey applicability) and survey cost. For a fixed budget, size-complexity trade-offs must be made; for a fixed level of survey complexity, size-cost trade-offs need to be examined; etc. Given that cost is inevitably a concern, these trade-offs play a critical role in every survey design.

Travel surveys are more complex than many other surveys in terms of sample size determination, since two more "dimensions" exist in these trade-off calculations: spatial and temporal precision. A few hundred observations may be sufficient to determine region-wide mode shares. But in order to understand mode shares at the fine-grained spatial scale of traffic zones (as is typically required for transportation analysis and modelling), then much larger sample sizes are required. Similarly, accurate representation of trip rates by time period requires larger sample sizes than if 24-hour totals are only required.

2.5 Sample Recruitment, Contact Methods & Incentives

Once a sample has been drawn from the sampling frame, contacting and recruiting each respondent is the next critical step in the process. In most surveys the majority of non-respondents arise from failure to recruit them to the survey. Failure to recruit selected respondents results in extra survey costs if these respondents are to be replaced, and, more worrying, can introduce significant non-respondent bias into the survey if non-respondents are atypical in their travel behaviour relative to respondents. A classic example of this sort of bias is the case in which successful contacts are never achieved with very frequent trip-makers while very low frequency trip-makers are much more likely to be contacted and willing to complete the survey.

Contact methods often depend on the type of survey being conducted but generally include letters mailed to the respondents and telephone or e-mail contacts. Contact and interview media can be mixed. It is common for example, for telephone surveys to use a letter mailed to the respondent a week prior to the interview as the initial contact. This forewarns the respondent that they have been selected for the interview, helps differentiate the travel survey interview from other, routine "market research" calls, and provides an opportunity to impress upon the respondent the importance of the survey and the value of their participation in it.

Incentives (lottery tickets, cash, small gifts, etc.) are often used as additional inducement for respondents to participate in the survey. These are generally not used for large-scale travel surveys but are often used in smaller-scale surveys, especially those involving considerable respondent burden. Mixed results are reported in the literature with respect to the cost-effectiveness of incentives on response rates (Tooley, 1996; Zimowski et al., 1997; Kurth, et al., 2001; Singer, 2002), with results clearly varying with the specifics of the survey and the population being surveyed.

For mail-back and web surveys, which depend upon the respondent self-completing the questionnaire, *follow-up* contacts with persons who have not yet responded to the survey can often significantly increase response rates by reminding the respondents about the survey and reinforcing the importance of their participation. These follow-ups can, again, be by mail, telephone or e-mail, depending upon the survey, the original contact medium, etc.

Dillman (1978, 2009) emphasizes the need for a “total design” approach to recruitment, incentives, follow-up, questionnaire design and all other aspects of Cochran’s survey design process in order to maximize response rates. Researchers who closely follow Dillman’s methods routinely report much higher response rates than in cases where much less attention is paid to design details (e.g., among many others, Miller and Crowley, 1989 and Hoddinott and Bass, 1986).

2.6 TEMPORAL CONSIDERATIONS

Numerous temporal considerations exist in survey design. These include:

- Survey observation period.
- Day(s) of the week to be surveyed.
- Seasonality effects.
- Observation method.
- Cross-sectional versus time-series surveys.

Each of these issues is discussed briefly below.

***Survey observation period*:** The most common travel survey observation period is one (24-hour) day, but both shorter time periods (morning peak period, etc.) and longer time periods (two or more days, one week, multi-week) are also used, depending on the survey purpose. Longer observation periods obviously provide additional information concerning a given respondent’s travel behaviour but with associated increases in survey cost and respondent burden (Pendyala and Pas, 2000; Chalansani and Axhausen, 2004).

***Day(s) of the week*:** Most travel surveys focus on weekday travel, given the emphasis in transportation planning on weekday peak-period travel as the dominant concern in facility and service design. Weekend travel, however, has very different travel patterns than weekday travel and therefore stresses the transportation system in different and important ways. Furthermore, there is considerable day-to-day variation in travel between weekdays and between weekend days, as well as important interactions between weekday and weekend travel decisions. In particular, some aspects of seven-day patterns, notably shopping trips, have evolved considerably in the past two decades. Understanding travel behaviour and needs over the entire week is also of importance in designing transit services that provide “real” alternatives to the private, car, particularly in suburban neighbourhoods. Activity-based travel models also are drawing increasing attention to the interplay between weekday and weekend travel decision-making.

As noted above, the most common type of household survey collects travel behaviour information for a single 24-hour day, usually a weekday. In large surveys, respondents are interviewed over a period of several weeks or even months, with each respondent being interviewed on a randomly selected day. The result is observations of travel distributed approximately uniformly across different days of the week (e.g., Monday through Friday) and, typically, across multiple weeks. These data are then generally averaged to yield a description of travel behaviour for a "typical" or "average" day.[3] While universally adopted, this approach does represent a form of temporal aggregation. Each day of the week has its own characteristic travel behaviour (trip rates, O-D patterns, etc.) and the "typical" day we construct from our survey data in a literal sense never truly occurs. Similarly, variations in weather and other factors will introduce week-to-week variations in travel which, it is assumed, "average out" when the survey data are combined to compute "typical" travel behaviour.

Seasonality effects: Over and above day-to-day variations, travel behaviour obviously varies by season. In particular, summer travel patterns are different from the rest of the year due to schools generally not being in session and due to prevalence of summer vacations. If seasonality is explicitly of interest for one reason or another, then some form of time-series[4] survey is required to observe the season-to-season variations of interest. Most travel behaviour surveys, however, are cross-sectional[5] in nature, in which case the season in which the survey is to be conducted must be explicitly considered. A common practice to conduct surveys during the fall (or possibly spring) season. A number of factors underlie this decision, including:

- Fall and spring are usually assumed to represent "typical" travel conditions, in that they avoid both "atypical" summer conditions and winter weather effects.
- Weather conditions are hopefully moderate and have little overall impact on the observed travel behaviour.

Regardless of season chosen, if travel surveys are to be repeated over, then the same season should be used in each repetition of the survey so as to maintain comparability of the data gathered over time.

Also, weather effects should always be considered in the design of any data collection effort, as well as in the analysis of the data collected. This may include ensuring that any instrumentation used can withstand the range of weather conditions under which it is expected to operate, possibly suspending surveying (or other data collection efforts) on extreme weather days, and, at a minimum, always recording the actual day(s) on which the survey/data collection occurs so that weather information can be attached to the data records for possible use in analyzing the data. In particular, increasing interest in weather effects on travel behaviour exists (e.g., Saneinejad et al., 2011), while weather effects are very important to account for when modelling transportation air pollution dispersion and exposure (Hatzopoulou and Miller, 2010).

***Observation method*:** Travel behaviour data can be collected either by asking respondents to *actively self-report* their travel, or by using technology to observe this behaviour, with the later being commonly referred to as *passive data collection*, since the observed trip-maker is not directly involved in the data recording (and, in some cases, may not even be aware that data are being collected). Active self-reporting can occur in two main ways (discussed further in Chapters 3 and 4):

- *Retrospectively*, in which they are asked to recall their travel behaviour over some prior time period. For travel behaviour, the prior time period is typically the previous day ("what did you do yesterday?"), although longer historical time periods may be useful for

[3] Or, equivalently, when building travel models, data from all days are used to construct models of travel behaviour for a "typical" day.

[4] See below for definition and discussion of time-series versus cross-sectional surveys.

some purposes ("in your previous job, what mode did you use to travel to work?). When questioning retrospectively, two modes of questioning are possible: *"typical behaviour",* in which the respondent is asked to self-report their typical or "most likely" behaviour ("what mode to work do you typically use?"), and "*actual behaviour*" or "*revealed preference*" in which respondents are asked what they actually did in a given situation. Most household travel surveys employ the "actual behaviour" approach ("what trips did you make yesterday"). A strong argument exists for requesting actual behaviour whenever it is feasible to do so, since (a) it avoids people "integrating" over their past behaviour to come up with what they think is their typical behaviour, which may generate reporting biases of various types, and (b) it allows for the observation of "random events" (someone who usually drives to work taking transit because the car is in the repair shop, etc.) that, with sufficient sample size, will yield a statistically more valid representation of a "typical day's" trip-making (i.e., a typical day includes a certain number of "untypical" events).

- Using *diaries* to record travel behaviour as it occurs, i.e., a trip (or activity) diary is filled out by the respondent as the behaviour occurs (or very shortly afterwards). Diaries are usually more expensive to process and may impose additional respondent burden but they generally produce more detailed and accurate data (Golob et al., 1997).

Passive data collection can also be undertaken in two primary ways (both of which are discussed further in Chapter 5):

- *Dynamically in real time*, using methods such as smartphone apps or stand-along GPS devices.
- *Indirectly*, by analyzing data collected for other purposes from which travel behaviour can be inferred (e.g., transit smartcard records, cellphone data, credit card transaction data, etc.).

***Cross-sectional versus time-series surveys*:** A cross-sectional survey is one that is executed at effectively one point in time.[5] A time-series survey is one that it is repeated over time, typically at regular intervals. Four basic types of time-series surveys exist:

- *Repeated cross-sections*. This involves repeating the same survey at multiple points in time with independent samples of respondents being drawn each time. This approach allows behaviour within the system to be tracked over time, but the behaviour (and changes in this behaviour) of individual persons, households, etc. cannot be tracked over time. The Transportation Tomorrow Survey (TTS) program in Toronto (DMG, 2007) and the series of Montreal surveys dating back to 1970[6] are both examples of on-going repeated cross-section surveys in Canada.
- *Panel surveys*. In a panel survey, a set of respondents is recruited and then repeatedly interviewed over time. This allows changes in behaviour among these respondents to be tracked over time, which is extremely useful for building dynamic models of travel behaviour. Panels, however, are expensive to construct and maintain, respondent burden tends to be high (with corresponding attrition in panel membership often occurring), and

[5] A large travel survey for a large urban region will typically take several weeks or months to complete. Such surveys are generally still considered as being cross-sectional in nature, since the observations are typically pooled to describe the average behaviour that occurred during the survey interval.

[6] http://www.transport.polymtl.ca/eodmtl/.

agencies may be unwilling to wait for the results of a long-term panel to materialize. As a result, panels are relatively rare in transportation planning (Golob, et al., 1997), although a few notable examples exist such as the Dutch Mobility Panel (Wissen and Meurs, 1989), the Puget Sound Transportation Panel[7], the Swedish Panel Study[8], the Toronto Travel Activity Panel Survey (Doherty et al., 2004; Roorda and Miller, 2004), and the Quebec City Travel and Activity Panel Survey (Lee-Gosselin, 2005; Roorda et al., 2005). Multi-day/week surveys can be thought of as short panels, among the most notable of these is the German six-week Mobidrive survey (Chalasani and Axhausen. 2004).

- *Retrospective surveys.* Retrospective surveys ask respondents to recall past activities. Such surveys have been shown to produce useful data for "longer-term" decisions such as auto ownership (Roorda et al., 2000), residential location processes (Hollingworth and Miller, 1996; Haroun and Miller, 2004) and employment careers, providing a cost-effective, attractive alternative to panel surveys for such long-term processes. They are not generally useful for short-term, day-to-day travel decisions, except, as already discussed, for the "what did you do yesterday" type of questions typical of many travel surveys.
- *Continuous surveys.* An alternative to repeated cross-sectional surveys is a continuous survey approach in which smaller samples of respondents are surveyed on an on-going basis. Potential advantages of the continuous approach include avoidance of massive swings in budget and staffing and the constant availability of "current" data. The statistical challenges of "integrating" over this continuous smaller-sample data stream to construct "current snapshots" of system behaviour, however, are non-trivial. Numerous examples of continuous surveys exist in both Europe and Australia (Zmud et al., 2011). The growing use of technology-based data collection methods (cellphones, smartphones, web-based surveys, etc. – see Chapter 3) may provide new opportunities for large-scale continuous surveys (El-Assi, et al., 2017).

2.7 Questionnaire Design

Questionnaire design is critical to the quality of data that are obtained through the survey and to minimize respondent burden and understanding of the questions being asked. A detailed discussion of good questionnaire design is well beyond the scope of this review. Excellent manuals on questionnaire design include (Dillman, 1978, 2009).

[7] http://psrc.org/assets/1484/PSTP_summary.pdf.

[8] The Swedish Panel Study gathered data on a variety of household market and nonmarket activities, including two detailed time-use surveys. http://www.nek.uu.se/faculty/klevmark/hus.htm.

Chapter 3
Household Travel Surveys

3.1 Introduction

Undoubtedly, the "workhorse" used world-wide for gathering person-based travel data is household surveys. Montevideo has recently completed a new household travel survey, the *Montevideo Household Mobility Survey* (MHMS), which is discussed in detail in the second report in this project's report series. The purpose of this chapter is to present a more generic discussion of the current state of the art/practice in household travel surveys and of current issues in their use. Section 3.2 provides an overview of household travel surveys, including definition of key terms, while Section 3.3 briefly discusses typical interview methods. Section 3.4 focuses on increasingly-used web-based surveys.

3.2 Overview of Household Travel Survey Methods

A *household* is usually defined as "a person or group of persons who occupy the same dwelling and do not have a usual place of residence elsewhere …. The dwelling may be either a collective dwelling or a private dwelling. The household may consist of a family group such as a census family, of two or more families sharing a dwelling, of a group of unrelated persons or of a person living alone."[9]

The household is an extremely useful sampling unit for general purpose travel surveys for a number of reasons, including:

- Although travel is executed by individuals, travel decisions are made to a large degree within a household context, both in terms of the resources and constraints available to individuals for their trip-making (income, car availability, opportunities to ride-share, etc.), and in terms of household-based activities (joint shopping and recreation trips, serving dependents, household chores, etc.) (Miller, 2005).
- Individual-based sampling frames are not generally used in transport surveys,[10] whereas household-based sampling frames are commonly constructed. As noted in Chapter 2, the household represents a "cluster" structure for individuals, permitting cluster-bases sampling of individuals to be achieved through the household survey. That is, by enumerating and randomly sampling households we can implicitly enumerate individuals and construct a statistically valid sample of individual trip-makers. Traditionally, household sampling frames have included:
    - Assessment rolls.
    - Utility billing lists.
    - Telephone directories.

[9] http://www.statcan.gc.ca/concepts/definitions/household-logement-eng.htm.

[10] Although individual sampling frames are used for special-purpose surveys, such as surveys of driver licence holders.

Thus, in a household travel survey, a representative sample of households in the survey target area is selected from an appropriate sampling frame. The sampling procedure used is usually a simple random sample or else a spatially stratified sample, in which a target sampling rate is applied within each of a set of districts within the study area to ensure a geographically balanced sample (see, for example, DMG, 2007). Census data are often used to provide population totals that are used to establish sample weights for population level estimates. Detailed data are gathered on each household member aged above a pre-defined threshold. The data collected vary from one survey to another but they can broadly be classified into three types:

- Household data, including location, type, number of persons, number of vehicles, etc.
- Personal data, including age, gender, employment status, occupation, possession of valid driver's license, etc.
- Trip data, including origin and destination of trip, trip purpose, trip mode(s), start time of the trip, etc. In some surveys, further details are collected on transit trips (e.g., access stop, route, number of transfers, etc.).

Depending somewhat on the interview type (discussed further below) and other design considerations, the trip data can be collected either retrospectively ("what trips did you make yesterday?") or by means of a trip diary that is filled out by the respondent for the survey day(s). Increasingly, particularly in the US and Europe, trip diaries are being replaced by *activity* diaries, in which emphasis is placed on recording the attributes of out-of-home activities and the trips associated with participation in these activities rather than just on the trips per se. This activity-based approach facilitates the development to activity-based travel models, but also, it is argued, results in higher reported trip rates than those obtained in traditional trip-based surveys, in which non-work/school and/or non-home-based trips tend to be more readily forgotten (DMG, 1991, 1993; Ashley et al., 2009).

Activity-based surveys also facilitate the collection of *in-home* activities in addition to the *out-of-home* activities (trips) which have been the traditional focus of travel surveys (Doherty and Miller, 2000; City of Calgary, 2002). In-home – out-of-home activity trade-offs are becoming increasingly important for understanding:

- Trip generation rates for entertainment, shopping and social activities (among others).
- The impact of ICT (information communications technology) on work (and school) tele-commuting.
- The increasing propensity for people to work at home rather than an out-of-home location.

The collection of in-home activities is still not common, even in activity-based surveys, and the practical modelling of in-home activities is in its infancy, but the need to consider in-home activity effects in our analyses and modelling is growing. At a minimum, information concerning "work-at-home" and tele-commuting is required and is generally collected even in conventional travel surveys.

A critical design issue is the determination of the respondents within each selected household who undertake the interview. Two options exist:

- A single respondent completes the interview on behalf of all household members. This single respondent is typically arbitrarily self-selected on the basis of who answers the

door or the telephone (FTF or telephone interviews) or volunteers to complete the questionnaire (mail-back and web surveys), although targeting a particular household member is also an option.
- Every household member is interviewed.

Single-respondent surveys are much easier to administer, generally cheaper to execute and typically will have higher response rates than all-member surveys, which are difficult and expensive to coordinate and impose more burden collectively on the household. They do, however, have serious potential to generate significant non-respondent bias in that the respondent often will not have complete information concerning the non-respondents' travel behaviour, unless special care is taken to minimize this risk.[11] Options for reducing non-respondent biases include:

- Having the respondent prepare in advance for the interview by collecting the required travel data from the other household members. It is generally difficult to monitor the effectiveness of this strategy.
- Adopting a mixed strategy, in which all household members fill out a trip diary, but only one household member provides the diary information to the survey staff.
- In telephone interviews have each household member come to the phone in turn to complete his/her portion of the interview. This increases response accuracy but obviously can pose interview scheduling issues (a randomly placed call to the household has a good chance of not finding all household members available). Multiple household member responses do occur informally in current telephone surveys, and, in some cases, are systematically encouraged.

3.3 HOUSEHOLD SURVEY INTERVIEW METHODS

Interviews of respondents can be undertaken in a variety of ways, which can be categorized broadly into four classes:

- *Face-to-face (personal) interview*. The interviewer meets in person with the respondent, directly puts the questions to the respondent and records the respondent's answers. When the interviewer uses a laptop to ask questions and input data, this is commonly referred to as Computer Assisted Personal Interviewing, or CAPI for short (Market Research World, 2011). This method allows for real-time verification of data, which further enhances the data quality and minimizes data gaps.
- *Mail-back survey*. The respondent is given (or sent via the mail) the survey questionnaire along with instructions on how to fill it out. The respondent completes the questionnaire and mails it back to the survey team using a stamped, addressed envelope that had been supplied to the respondent along with the questionnaire.
- *Telephone interview*. The interviewer conducts the interview of the respondent over the telephone. This type of interview is typically conducted *via* CATI (Computer Assisted Telephone Interview) systems, which enable the interviewer to enter data directly into the computer and verify the validity of the input information.

[11] This potential bias may also increase with household size, which increases the burden on the respondent to report the actions of more household members.

- *Web survey*. The respondent is given access to a web site that presents the survey questionnaire in an interactive format to the respondent. The respondent completes the survey on-line in response to questions and other prompts provided to the respondent from the computer program.

The first great round of household travel surveys occurred in the 1950's and 1960's throughout North America as part of the general emergence of comprehensive long-range urban transportation planning and the associated development of the first generation of computerized four-step travel demand forecasting model systems that occurred during this time period (Meyer and Miller, 2001). These surveys universally were face-to-face, reflecting both the resources available for surveys in this era as well as the relative lack of alternative technologies. These surveys were generally retrospective (previous day trips collected) from a single respondent, although FTF interviewing certainly does not preclude either the use of trip diaries or the interviewing of multiple household members. While the response rate and data quality of household FTF interviews are the highest relative to other methods of household surveys, it is the most expensive on a per-completed survey basis (Sharp and Murakami, 2005).

Mail-back surveys are rarely used in North America for large-scale household travel surveys, but they are commonly used in Australia, with self-completion questionnaires typically being personally delivered to respondent households (FTF travel diary interviews are also used in Australia) (Inbakaran and Kroen, 2011). The classic mail-out/mail-back survey involves mailing a questionnaire to each household in the sample, with the household then completing the survey and mailing it back. Variations, however, exist, including hand delivering the questionnaire (perhaps as part of an up-front, preliminary FTF interview) with the respondent mailing the completed questionnaire back to the survey team, or mailing the questionnaire to the household but then retrieving the completed survey via a telephone interview, among, undoubtedly other permutations. The survey itself can be either retrospective or a diary.

Mail-back surveys are generally considered to yield very low response rates with lower quality data but with the compensation of being cheap to execute. The costs of printing, envelopes and postage, however are in fact non-trivial (especially on a per-response basis if low response rates are, indeed, experienced and/or if multiple follow-up mailings are employed in an effort to boost the response rate), as are the data processing costs to get the completed questionnaires into machine-readable form. Also, as Dillman (1978, 2009) and others have consistently demonstrated, very adequate response rates are, indeed, achievable if proper care and effort is taken in survey design and execution.

The third method, telephone interviews, generally strikes a balance between the two previous methods in terms of data quality, response rate and cost. The random sample is typically drawn from a comprehensive list of household landline numbers or via random digit dialing (RDD). Each household in the sample is contacted by an interviewer who collects the survey data over the phone. The respondent is asked questions concerning household, personal and travel characteristics, the latter typically on a retrospective basis, although retrieval of a previously completed trip diary over the telephone is also possible. A CATI system is generally used, which greatly enhances the efficiency and accuracy of the data recording process. Household recruitment methods can vary, but Canadian practice, for example, is typically to mail a letter to the respondent household approximately a week before the planned telephone interview explaining the purpose of the survey and indicating the time they will be contacted for the interview. Telephone-based surveys, however, are increasingly challenged by low response rates (call-screening with answering machines), sampling biases (fewer and fewer households have land lines), and cost (RDD is very expensive on a per-completion basis).

Further, telephone interviews have traditionally relied on sampling households with land-lines. Land-line usage, however, is declining dramatically in most developed countries, while in developing countries many have to a large extent by-passed land-line technology in favour of mobile technology as the primary telecommunications medium. Mobile phones generally do not provide good sampling frame definitions for several reasons, including:

- Directories of mobile phone numbers often do not exist or are very expensive to assemble.
- Since a mobile phone user may well be contacted outside the home while travelling or engaging in an out-of-home activity, people contacted may be often be very reluctant to participate in the survey.
- Mobile phones attach to a person, not a household. This surveying the travel behaviour of entire households becomes very problematic.
- Conversely, a random sample of mobile phone can well lead to multiple contacts with a given household, which is also problematic to deal with.

Detailed information on the design, testing, organization and implementation of the above survey methods is well documented in the literature (TRB Travel Survey Methods Committee, 2012; Transport Canada, 2009; NCHRP 2007, 2008; Miller, et al., 2012). Recent reviews of these methods are also available in a series of papers developed by the TTS 2.0 research team (Srikukenthiran, et al., 2017a,b; Harding, et al., 2017a; Chen, et al., 2016; Pulikanti, et al. 2015).

3.4 WEB-BASED SURVEYS

Increasingly, various implementations of web-based surveys have gathered the sorts of trip or activity data from people and/or households that are collected when conducting travel surveys using other methods. Sharp and Murakami (2005) compare data collection methodologies and define elements of internet surveys:

- Respondents complete the survey on the web (self-reporting survey).
- Only households with access to internet are covered.
- Response rates are usually lower than other methods.[12]
- Quality of data varies and depends on the validation functions, interface design and question formulation.
- Data are rapidly available.
- Overall cost is low, but start-up costs can be high compared with data-collection cost.

Usually, respondents fill the survey on their own, to their best knowledge. The clarity and user-friendliness is, in these circumstances, a key element. As for other technology-assisted tools, the

[12] Although, as with most survey instruments, response rates may possibly be increased through appropriate use of incentives.

web allows for real-time validation and can automatically provide feedback on the validity of the answer. A certain level of quality control at the time of the survey participation is then possible.

Web-based surveys have similar applications as other personal or household travel surveys, including the use of all transportation networks or only the use of a specific network (transit or carsharing for instance). They can assist in collecting information on households, people, trips and preferences.

While possessing many strengths, web-based surveys share many of the same limitations as other survey modes:

- **Survey design**. Designing a web interface that is clear for every respondent is not a simple task. It cannot be a simple transposition of a paper and pencil survey or an interviewer interface. The interface has to be immediately understood and attractive. Hence, the initial investment in the interface design is one of the important costs of web-based surveys. Also, technology is changing fast in the web industry and it is difficult to keep a design operational for a long time (Bourbonnais and Morency, 2011). Using a web-based survey means that resources must be applied on a regular basis to update the application and redesign the interface.
- **Low response rates**. This is a general issue with travel surveys, notwithstanding the survey mode. Efficient tools to recruit participants are required. For web-based surveys, social networks combined with interfaces for computers, tablets and smart phone will facilitate wider distribution of the call for respondents.
- **Sample representativeness**. This issue depends on various elements:
    - The quality of the sampling frame (i.e., the quality and representativeness of the list of people or households from which a random sample is drawn). In the case of web-based surveys, sampling can be performed by sending invitation letters or emails to selected households or people or by sending invitations at large. In the former case, there is still a certain control of the sample composition but not in the latter case. Also, gaining access to a representative list of emails is not trivial and emails are usually linked to people and not households.
    - Only people with access to the internet can complete the questionnaire. The scale of this issue will decrease with time but there will probably always be people without access and they probably will not have similar attributes and behaviour to those who have access. Also, skills in the use of the internet are not uniformly distributed over the population. However, the general level of familiarity will increase over time.
- **Data quality**. The quality of data gathered using self-reported surveys highly depends on the quality of the interface. The web allows using various types of validation and graphical representation and it can help in ensuring good quality data. Post-processing procedures to further validate and impute, when necessary, are still required. As with many survey instruments, “call-backs” (by phone or email) can be used to verify responses if need be.

For recent discussions of web-based survey design with the TTS 2.0 project, see Loa, et al. (2015) and Chung, et al., (2017).

CHAPTER 4
CHOICE-BASED SAMPLE SURVEYS

4.1 INTRODUCTION

A choice-based sample survey is one in which the sampling frame consists of a set of people who have all made the same choice (mode, route, etc.). The most typical choice-based sample surveys used in urban transportation are roadside surveys (in which the sampling frame consists of trip-makers who have chosen to travel by car along a given route) and transit onboard surveys (in which the sampling frame consists of people who have chosen to take transit using a given transit line). These are discussed in some detail in Sections 4.2 and 4.3, respectively.

Many other examples of choice-based sample surveys, however, exist, including air traveller surveys at airports, parking lot surveys, shopping mall surveys (or other "special generator" locations), etc. These various types of intercept surveys generally involve the same design issues as transit intercept surveys, and so are not explicitly discussed in this report. Another important type of survey, however, which loosely falls into this category are place-of-employment and place-of-school surveys, in which respondents are chosen on the basis of where they work or attend school. This survey approach is briefly discussed in Section 4.4.

Choice-based sample surveys are a very efficient data collection approach in cases in which information is only required for the targeted population (car drivers, transit riders, etc.). This is particularly the case for small or specialized populations that might be difficult and/or expensive to sample effectively through general-population-based sampling frames (such as are used for household surveys). The increasing cost of conventional household travel surveys, as well as the increasing need for a variety of reasons to study specific segments of trip-makers in detail, makes choice-based sample surveys an attractive option (Pendyala et al., 1993).

On the other hand, choice-based sample surveys obviously do not provide sufficient data for the analysis or modelling of travel choices that lie outside the survey context. For example, a roadside interview survey can provide information about the origin-destination pattern of trip-makers using the road at the survey point, but it cannot provide the basis for modelling the decision to drive relative to taking another mode of travel for the observed trips. Choice-based sample survey data, however, can be combined with other, more general data sets (e.g., a household travel survey) to provide needed details concerning the given process that may be missing or inadequately observed in the more general survey. A common example of this occurs in many US mode choice modelling exercises in which the household travel survey does not contain sufficient transit trips to adequately estimate the mode choice model parameters.[13] In such cases, it is common to use onboard transit ridership survey data to augment the household survey data (Cambridge Systematics 1996). In such applications, care must be taken to correctly combine the choice-based data within the model parameter estimation process, but standard procedures for doing so exist (Manski and Lerman, 1977).

[13] This situation arises through the combination of the low transit mode splits (5% or less) experienced in many US cities combined with the small sample sizes typically employed in US household travel surveys.

Choice-based sample survey design generally involves the same steps and concerns as population-based surveys. The biggest differences between the two types of surveys include:

- Sampling frame definition is theoretically straightforward: it is the population of users of the facility being surveyed. Care, however, must be taken to accurately count this population, and cases exist in practice in which this is not a trivial task.
- Sample selection and contact is accomplished through some form of *intercept* method, in which trip-makers are identified while in the process of making their trips. Different intercept/contact methods are used, depending on the specific type, constraints and needs of a given survey.
- The sampling methods most commonly used are generally either simple random sampling or sequential sampling (e.g., stop every 10th car), although stratification of trip-makers is also conceivable (e.g., over-sample families in an airport intercept survey).
- The most common interview methods include:
    - *Face-to-face* (FTF) is very commonly used, typically involving a quick interview at the point of interception. Paper-and-pencil or a computer-aided laptop system can both be used for data recording. In situations which permit the interviewer and respondent to sit down together and spend a few more minutes together (such as in an airport lounge) more detailed computer-aided questionnaires (including stated-preference/response type experiments) can be undertaken.
    - *Mail-back* surveys can be handed out to the respondent at point of contact, leaving the respondent to subsequently fill out and return the completed survey form. These questionnaires are typically very short and succinct in nature.
    - Having identified respondents at the intercept point, the respondents are later contacted and recruited for the survey. This subsequent contact is done generally by either mail or telephone[14], with the survey being either a mail-back or telephone interview. A typical example of this approach is the gathering of car license plates at various intercept points followed by contacting the car owners corresponding to the set of observed license plates.

4.2 Roadside Intercept Surveys

Roadside intercept surveys are one of the oldest and best established ways of collecting road-based travel data by intercepting and questioning people during the course of their travel (Lestina et al., 1999, Beirness and Beasley 2010, Cambridge Systematics 1996). Traditionally, such surveys have been used to collect auto origin-destination (O-D) trip patterns, and so they are frequently referred to by practitioners as “O-D surveys”.[15]

In cases in which transit and non-motorized travel is not of interest, roadside O-D surveys possibly can be used in place of household travel surveys to develop models of road-based trip generation, distribution and assignment. In all cases, such survey data can be used to supplement more general travel surveys, especially in situations in which data concerning types of trips is

[14] E-mail addresses typically are not known but could, of course, be used if they are known.

[15] Roadside intercepts can also be used to count and/or interview persons travelling by transit (which is one form of transit user intercept surveys, discussed in the next section) or even pedestrians and cyclists. The focus in this section, however, is on roadside intercepts of auto users, by far the most common application of this survey type.

sparse within the general survey (e.g., O-D patterns in small sample general surveys), as well as to help in travel model calibration/validation.

Cambridge Systematics (1996) presents a complete manual for roadside intercept surveys. It classifies intercept surveys into four main categories:

- *License plate surveys*, in which license plates are recorded (either manually or, more typically, by being automatically photographed) as vehicles pass by intercept points. The recorded license plate numbers are then matched with vehicle registry information to identify the home locations of the drivers, and either a telephone or mail-back survey is conducted to obtain detailed information concerning the observed trip.[16]
- *Roadside handout surveys*, in which vehicles are stopped at intercept points and a mail-back questionnaire is given to the driver to complete and return by mail at some subsequent point.
- *Roadside interview surveys*, in which vehicle drivers are interviewed "on the spot" once they have been stopped at an intercept point.
- *Combined roadside handout and interview surveys*, in which a quick roadside interview is combined with a more detailed mail-back handout.

In addition to the usual issues of sampling rate, questionnaire design, etc., important sample design issues for roadside surveys include:

- Choice of survey method (license plate survey, roadside handout, etc.).
- Selecting what road sections/links to sample in the survey.[17] Budget and other resource issues often limit the number of intercept points that can be feasibly used. These must, therefore, be chosen with care so as to maximize the information obtained about overall system performance.
- In the case of stopping vehicles, feasible locations for stopping the vehicles must be found that introduce minimal disruption to the vehicle stream and that are safe for drivers and interviewers alike. Police involvement is generally required to ensure safe and orderly implementation of the survey sites and compliance of drivers to stop for the interviews.
- Selecting the type(s) of road users to survey. If certain user types are to be over-sampled or excluded, then care must be taken to develop an appropriate stratified sampling procedure.

Typical roadside intercept survey data elements include:

- Travel data: trip purpose, arrival and departure time, travel time, vehicle type, origin and destination addresses, number of persons in the vehicle, travel routes and frequency of trips

[16] Note that if licence plate numbers are observed both entering and exiting a cordon around a study area then computerized number matching algorithms can be used to link together vehicle entry and exit points, thereby generating estimates of origin-destination flows passing through the cordon. The successfulness of this process depends on the having a large number of observation points and high licence recording rates, so that a large number of vehicles are successfully observed at both entry and exit.

[17] Roadside survey intercept points are often located at selected locations along a cordon or screenline so as to maximize the observation of flows into and out of the cordoned area.

- Demographic data: residential location, household size, occupation, household income, age and gender.
- Attitudinal data, such as perception about congestion, potential use of alternative routes, alternative means of travel etc.

Roadside interviews have the advantages that they provide immediate collection of the required data and they generate very high response rates – it is very difficult for a driver, once stopped, to refuse to do a quick interview. Disadvantages of the approach include:

- The method is clearly intrusive and disruptive.
- Setting up intercept stations is costly and requires extensive coordination with police, etc.
- The method generally is not practical for high-volume roadways.
- Relatively few interview stations usually are feasible to maintain.
- The interview generally needs to be very brief and so very limited information is typically collected.
- The legal authority to stop people to question them is required.

Roadside handouts share the disadvantages of roadside interviews in terms of needing to stop vehicles on the roadway. The mail-back handout questionnaire potentially can be somewhat more complex than the roadside interview survey since it is not completed on the spot. Response rates relative to the roadside interview, however, inevitably will be lower, especially since little generally can be done with respect to incentives and/or follow-up contacts to encourage completion of the questionnaire. The handout approach is less disruptive to traffic flow than roadside interviews, since the cars generally are stopped for shorter periods of time (just long enough to explain the survey and hand out the mail-back questionnaire) and may be less costly to run since road side survey staff are able to process a higher number of vehicles per unit time. This may also lead to higher sampling rates, which may at least partially compensate for lower response rates.

The combined roadside interview/handout method in some circumstances may combine the strengths of both individual approaches in that it ensures a good base of critical information collected during the short interview, complemented by more detailed information gathered through the mail-back handout. Motivation for completing the mail-back questionnaire may be higher, given that the respondent has already completed the short roadside interview. The challenges of stopping vehicles along the road, of course, remain with this method.

License plate surveys possess several attractive features, including:

- They do not require stopping vehicles and can be used in even very high-volume situations.
- They are relatively cheap and straightforward to implement.
- The telephone or mail-back survey can be reasonably complex in nature.

The license plate approach, however, is not without its own difficulties. These include:

- Privacy concerns. Many people object to their movements being monitored without their consent.
- Cooperation of the motor vehicle registry office to release vehicle owner addresses and/or telephone numbers is required.

- The driver of the vehicle during the recorded trip may not be the owner of the vehicle.
- The follow-up contact (by either mail or telephone) must occur as shortly as possible after the trip is recorded so that the respondent's recall of the trip is as fresh and accurate as possible.
- Response rates may be lower.

Regardless of the survey method employed, common limitations of roadside intercept surveys include:

- Only data concerning motor vehicle trips are obtained. Travel behaviour by other modes is usually ignored.[18]
- Given the survey methods employed, generally fairly limited information concerning the respondent's trip-making behaviour, personal attributes, etc. can be obtained, with this often being limited to just the observed trip.
- Data collection is limited to the (typically few) sites at which intercepts occur. Thus, information concerning O-D patterns (and road-based travel behaviour in general) within the urban area is inevitably limited to those trips which happened to choice a route passing by one of the intercept points. The approach thus is most effective when studying a specific corridor or when travel patterns are otherwise constrained in a way that they can be usefully observed using a relative handful of key observation points.

Cordon/screenline-based O-D surveys (for both road and transit, see the next section) can be combined with household travel survey data to improve the estimation of O-D matrices. For maximum effectiveness this obviously requires coordinating the timing of the roadside and household interview surveys. In addition to the use of household survey data, roadside data can also be combined with traffic count data and traffic route assignment models to generate improved O-D matrix estimates (Guy and Fricker 2005). The availability of WiFi technology and mobile phone network positioning capabilities also open possibilities for improved monitoring and modelling of road-based travel (see Section 4).

4.3 TRANSIT USER INTERCEPT SURVEYS

Transit users can be intercepted at the transit stations or stops while they are waiting for their bus or train, or onboard while they are travelling on the vehicle. Such surveys can provide origin-destination information for transit riders together with personal and household characteristics. Such data are typically used for estimating transit ridership, understanding behavioural patterns of transit users and characteristics of transit riders. Onboard survey data are used for scheduling and operations planning, long-range planning and design, performance analysis, preparation of statistics and reports, and market evaluations (Cambridge Systematics 1996, TAC 2008). Surveying passengers onboard vehicles can be easier to implement compared to intercepting respondents at stations or stops since the respondent is "captive" while onboard the vehicle, although interviewing passengers onboard very crowded vehicles can be challenging. Regardless of intercept location, transit user intercept surveys are typically generically referred to as "onboard" surveys or "transit rider" surveys.

[18] In principle pedestrian and bicycle trips can be surveyed using the same methods that are described for herein for motor vehicles, although this is rarely done in practice.

Onboard surveys are standard tools for many transit agencies (Seskin and Stopher 1998). Hartgen (1992) explains the evolution of the concept of choice-based sample survey conducted by the transit agencies as opposed to relying upon large scale travel surveys. Stopher (1992) presents an example approach for developing a transit ridership forecasting method at the route level by using transit onboard survey data. Onboard surveys are also used for monitoring route-level ridership trips, transit rider travel patterns and attitudes, and before-and-after assessment of route service changes on transit ridership.

Typical approaches to conducting onboard surveys:

- The driver of the transit vehicle hands out the survey questionnaire to the passengers and passengers either return the completed form while leaving or mail it back to the designated address. This technique is suitable for buses or other vehicles without all-door boarding.[19]
- Surveyors aboard the transit vehicle distribute the survey and collect the completed responses. They may interview the passengers onboard as well as count the boarding passenger. This technique is suitable for trains, large buses, BRT or LRT.[20]
- Surveyors aboard the transit vehicle distribute the survey and passengers mail the completed responses later to the designated address.

Cambridge Systematics (1996) presents a complete manual for transit onboard survey. Typical data elements that can be collected through transit onboard surveys include:

- Travel data: boarding and alighting location/stop/station, trip purpose, arrival and departure time, travel time, origin and destination addresses, access and egress mode, transit routes, fare payment type, auto ownership and auto availability for the trip
- Demographic data: household size, occupation, household income, age and gender
- Attitudinal data, such as perception about transit service, customer satisfaction, etc.[21]

Baltes (2002, 2003) presents a best-practice manual for transit onboard surveys. It describes the necessary steps for conducting a successful onboard survey of public transit customers, and provides a clear understanding of the total customer surveying process and its importance in planning and transit service design. Specific items discussed include various methods of onboard data collection, questionnaire design, sample size determination, data entry procedures, data reporting procedures and data archiving procedures. Blash et al. (2002) argue that although standard manuals are available for designing onboard surveys, significant challenges typically face agencies in their design and execution of practical surveys. These are mostly sampling related, relating to the selection of the specific routes and time periods to survey.

TCRP (2005) presents a comprehensive discussion on transit survey techniques. It provides a summary of transit agencies' experiences in planning and implementing onboard and intercept surveys. It reports a survey of 52 agencies, 96 percent of which had experience in conducting

[19] Also providing union agreements permit such activities, there are not safety issues involved in the method and boarding volumes are low enough for this to be practical.

[20] This assumes that trips are long enough to complete the survey while the passenger is onboard.

[21] For example, Habib et al. (2009) used onboard survey data of Calgary transit to investigate riders' attitudes towards transit services.

onboard surveys. Survey results reveal that large agencies typically conduct five or more onboard/intercept surveys per year and small agencies typically conduct surveys every 1 to 3 years. Onboard surveys conducted by large agencies are primarily focused on specific routes or geographic areas, but surveys conducted by small agencies often involve the entire transit system. This report points out that proper guidance for the questionnaire design and incentives is often missing in existing practice.

Chu (2006) examines the feasibility of using local transit onboard surveys for the measurement of state-level transit performance levels. Their definition of performance level is defined by the proportion of choice and captive transit users among all transit riders. The report suggests that if such estimates of the local transit agencies are available and can be synchronized in time from all individual agencies, then the state-level measurement can be developed by using a stratified sampling technique by using the local agency-level weights. However, it seems that many transit agencies overlook the issue of defining captive and choice riders in their onboard surveys. Moreover, in many cases, onboard surveys do not cover the whole transit system of the corresponding agencies.

Transit user intercept surveys are less disruptive than roadside intercept surveys as the respondents are targeted while they are waiting at the station/stop or in-vehicle. Surveys that are filled out while the passenger is onboard (or while waiting at a stop/station) do not impose any extra time on the passengers (i.e., the response burden is very minimal) and so these surveys tend to have higher response rates than mail-back approaches.

As with road intercept surveys, the major limitations of onboard surveys are (a) they do not provide information on non-transit users and (b) they are limited to the transit routes/stations that are selected to be surveyed. For this reason some transit agencies conduct household telephone interviews to supplement their transit user intercept surveys (TCRP Synthesis 63, 2005).

Transit onboard surveys can potentially be combined with Automatic Passenger Count (APC) systems on transit vehicles and/or smartcard data (see Chapter 5) through various data fusion techniques (see Chapter 6) to enrich the overall database characterizing transit ridership.

4.4 PLACE OF EMPLOYMENT/SCHOOL SURVEYS

Given the importance of work- and school-based trip-making in urban regions, especially during the morning and afternoon peak travel periods, it can be useful to survey workers and students directly by recruiting them at their place of employment or their school locations. Such a survey may focus only the travel of the workers/students surveyed, or it may collect information on all members of the worker's or student's household. In the later case, the survey can be referred to as an "inverted sampling" approach, in that households are not sampled at their place of residence (the standard "household survey" approach discussed in Chapter 3), but at the workplaces and schools of their workers and students.

Employment- (or school-) based surveys often employ a cluster sampling approach in which business establishments (schools) are first randomly sampled, and then the employees (students) at the selected sites are then surveyed (often 100% of the workers/students are contacted). This cluster sampling approach can be very efficient in cases where lists of business establishments (schools) are available or relatively easy to construct.

Such surveys may be used as a "satellite" survey to provide more in-depth information about sub-populations, such as post-secondary students, that may be difficult to reach within a conventional, general-purpose "core" survey (see Chapter 6 for discussion of "core-satellite" survey designs). A full "inverted sample" survey in which entire households are accessed via their workers and/or students may help in such cases reach otherwise hard-to-find households.

A very successful school-based survey was recently conducted of post-secondary students attending universities within the City of Toronto. The StudentMoveTO survey was a web-based survey that achieved a 8.3% overall response rate (15,220 completed responses) from all students attending the four universities located within the City.[22] Post-secondary students are a very difficult sub-population to reach within conventional surveys given their variety of on-campus and off-campus living arrangements and other factors. It is expected that some variant of this initial survey will be incorporated within Toronto's overall travel data collection program in the future.

Toronto's current TTS 2.0 survey methods project is also investigating the feasibility of large-scale employment-based surveys as part of an overall data collection program. Field tests are still underway using a web-based survey instrument, but to date, considerably difficulty has been encountered in terms of, first, recruiting firms who are willing to permit their employees to be contacted, and, second, getting these employees to complete the survey (Srikukenthiran, et al., 2017c). Clearly, this approach will work best in cases in which employers are strongly incented to support the survey and in which employees can be efficiently and effectively contacted. It also needs to be noted that in cities with significant "informal" employment (as is the case in some Latin American cities), it may be very difficult to adequately enumerate business establishments and their employees.

[22] For more information concerning the StudentMoveTO survey, results and associated research reports, see http://www.studentmoveto.ca/about/.

CHAPTER 5
ICT-BASED DATA COLLECTION METHODS

5.1 INTRODUCTION

We are in the midst of a revolution in travel-related data collection methods based on various types of Information and Communication Technology (ICT) systems and applications. While not without their own technical challenges, these new methods generally hold the promise of being able to provide very large, unprecedented quantities of high-quality travel information very cost-effectively. While varying considerably in technical details, these ICT-based methods and data generally share several key attributes and issues that need to be understood if these methods are to be successfully designed and applied.

First, these methods can generate massive streams of data, typically in real time, more or less continuously, day after day, week after week. The potential thus exists to gather time-series travel behaviour (rather than the usual cross-sectional snapshot obtained through a conventional travel survey) for very large samples of trip-makers, thus dramatically changing the nature of the data available for transportation system monitoring and control, travel demand modelling, and other operational and planning applications. At the same time, these massive streams of data require advanced computer storage and processing software to gather, maintain and analyze the data, and may well imply new methods for modelling travel behaviour.

Second, the data are often collected by third-parties, often in the private sector, and often for primary purposes that may not be directly transportation-related. This creates at least two potential challenges with respect to the use of such data. First, access to / cost of private sector data may be an issue in some cases. Second, in cases where the data are primarily being collected for other purposes, then considerable manipulation of the data and/or fusion with other data sets may be required in order to make the data usable for transportation-related purposes. This issue can exist even in cases where a transportation agency is collecting the data for one purpose, but fails to consider other applications in designing the data collection procedure (e.g., many smartcard systems, see below, or transit onboard surveys, discussed in Chapter 4).
The methods generally are labelled as being *passive*, in that they require no (or, in some cases, limited) *active* recording of information on the part of the trip-maker being observed. Indeed, in many cases, the people are not even aware that they are being observed/tracked. This is in contrast to all forms of surveys discussed in Chapters 3 and 4, in which respondents must actively respond to questions about themselves and their travel behaviour. Passive data collection has the potential to eliminate many problems of conventional surveys, such as respondent fatigue (in particular creating the possibility for observing a person over extended periods of time) and respondent reporting errors (e.g., forgetting to record short trips).

Third, for a variety of reasons, the data are generally only available in an anonymized form. That is, while we may observed trips being made, we generally do not know the attributes of the trip-makers. Thus, for travel demand modelling purposes (among other applications), one may need to *fuse* these data with other datasets which provide information that allows the analyst to impute socio-economic attributes of these trip-makers (see Chapter 6). Or, alternatively, some active survey component may need to be added to the passive trip-making data collection component in order to gather the socio-economic attributes of the trip-makers. This may, involve, for example, an "up-front" survey of respondents who have agreed to load a smartphone app to track their travel to gather their relevant socio-economic characteristics.

Similarly, in many cases the route (path through the network), mode and purpose of a given trip captured within the data collection process are not directly observed and so must be imputed through post-processing. Travel mode may be inferred from observed travel speeds and perhaps other attributes of the observed trip, although challenges generally remain in dealing with non-auto modes (transit, walking, biking). Inferring trip purpose usually requires fusing trip records with GIS land use and "points of interest" (POI) datasets. Identifying paths taken generally requires attaching the spatial traces for a given trip to GIS-based networks. The accuracy of this process depends critically on the spatial precision of the data gathered, which can vary considerably, as discussed immediately below.

Fifth, the amount of spatial information concerning a given trip, and the spatial precision of these data, can vary considerably from technology to another and from one urban area to another, depending on how a given technology is implemented locally. Not all methods gather sufficient information to identify all trip attributes (origin, destination, purpose, mode, route). The precision of any information gathered (e.g., stop locations) generally varies considerably from one method to another, and even within a given method within a given urban area can vary significantly from one area to another within the region. In cases where the spatial data are wither not available at all or are sufficiently spatially imprecise to permit imputing with reasonable accuracy, for example, trip mode or purpose, one may need to add an active survey component to the data collection method to query the trip-maker concerning these attributes. This might occur dynamically, as the trip is being executed, or post-facto (e.g., asking respondents to provide the needed information at the end of each day of the data collection period), depending on the method and its design.

Sixth, virtually all ICT-based data collection methods gather data about individual trip-makers, not households. If the travel behaviour of all members within a given household (along with attributes of the household as well as of the individuals), then special effort must be made (i.e., explicitly recruiting all household members to participate) to do so. Not all ICT-based methods, however will support household-level analysis/modelling. An open question at the moment is whether widespread adoption of individual-based ICT data collection methods will require significant redesign of our currently household-based travel models or not.

Seventh, as alluded to above, ICT-based data collection methods can be divided into two broad categories: those that observe travel behaviour in some way which does not require permission of the trip-maker or the trip-maker's knowledge that the data are being collected (e.g., roadside Bluetooth sensors capturing smartphone signals of trip-makers passing the sensor) and those that require the trip-maker to agree to participate in the data collection exercise (e.g., agree to load a trip-tracking app on their smartphone), even if the subsequent data collection is completely passive. In the latter case, many of the usual challenges of survey design and implementation remain, notably sample frame definition, sample selection, recruitment (and incentives to participate) and respondent retention.

Thus, the twin issues of sample recruitment and representativeness, generally remain with ICT-based methods, albeit perhaps in new ways. Even in the case of a totally passive, third-party dataset (say, cellphone-based CDR data, see Section 5.2) obtained from a cellphone service provider the issue of data representativeness exists. Who are the customers of this service provider and are they representative of the general trip-making population (both socio-economically and in terms of their trip-making behaviour)? Is the way in which trips being recorded introducing any significant biases in the data being collected? And so on.

Eighth, technology-based data collection obviously depends on the people that are to be tracked to have access to the given technology. Cellphone- and smartphone-based methods require trip-makers to have cell/smartphones (and for them to keep them turned on while travelling). Web-based survey methods require respondents to have some form of computing device (laptop, smartphone, etc.) and internet access. And so on. While most of these technologies are becoming increasingly ubiquitous within many urban regions, they certainly are not yet (and may never be universal). In particular, important travelling sub-populations, such as the poor, the elderly, etc. may have very limited access to many of these technologies. Obviously, a data collection method based on a technology that provides biased coverage of the travelling public will yield biased results. This does not necessarily rule out the use of the method for certain purposes, but it will need be integrated within a larger data collection program that gathers data for the missing / under-represented sub-populations using other methods (see Chapter 6).

Finally, these technologies are generally changing quite rapidly, both generically (e.g., smartphone technology) and with respect to transportation-specific applications (e.g., smartphone apps for tracking trip-making). As a result, it is difficult to "pick winners and losers" among competing technological options, since their capabilities, costs, strengths and weaknesses. Nevertheless, some general trends in the field are arguably beginning to emerge, as discussed in the following sections.

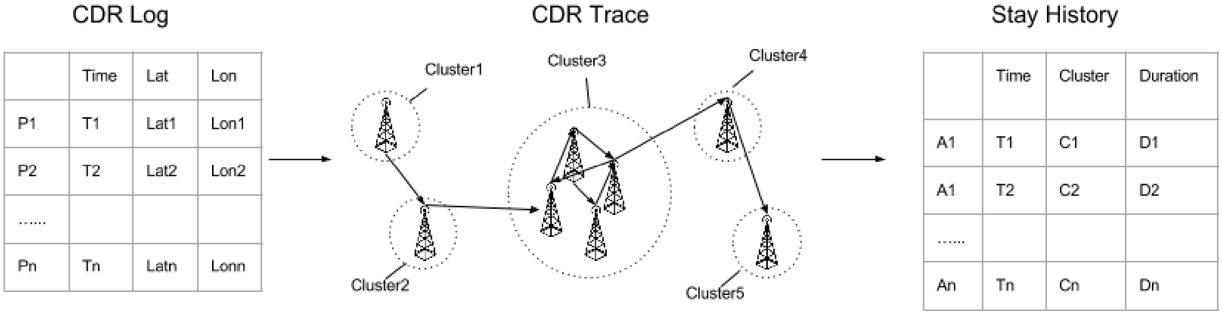
At least six major categories of ICT-based data collection methods are currently available (and are being used to varying degrees in urban regions globally). These are:

1. Cellphone based *cellular data records (CDR).*
2. *Geographic Positioning System (GPS)* tracking devices.
3. *Smartphone-based apps*.
4. A wide variety of roadside *sensor* devices and systems.
5. Transit *smartcard* and other mobility service usage data.
6. Other *third-party passive datasets.*

Each of these categories of methods is discussed briefly in turn in the following sections. More detailed discussion of many of these methods can be found in Miller, et al. (2012).

5.2 Cellular Data Records (CDR)

Every cellphone that is turned on and that is within range of one or more cellular transmission units is in constant communication with the telecom network providing the cellular transmission service through these units and is generating a continuous time stream of transmission events, called Call Detail Records (CDRs)[23] that are recorded and stored by the telecom company. The location of the phone (and hence the trip-maker carrying the phone) can be estimated based on the location of the cell tower with which it is communicating at each point in time. The spatial accuracy of these imputed locations varies considerably depending on the spatial distribution and density of the network of transmission towers, among other factors. In general, however, the spatial precision is not adequate to provide precise definitions of route or speed (and, hence, travel mode), but trip stop (or "stay") locations (including home and work locations providing records for multiple days for the same phone are available) can be determined with at least modest precision by identifying "clusters" of CDRs for the same phone that are grouped closely together in both time and space (see Figure 5.1).



	Time	Lat	Lon
P1	T1	Lat1	Lon1
P2	T2	Lat2	Lon2
……			
Pn	Tn	Latn	Lonn

	Time	Cluster	Duration
A1	T1	C1	D1
A1	T2	C2	D2
……			
An	Tn	Cn	Dn

Figure 5.1: Converting raw CDR data into trip stop/stay locations
Source: Yin, et al. (2016)

The process of converting raw CDR data into usable information for travel demand analysis purposes is complex. CDR data, however, can potentially provide information on thousands, if not millions, of trips within an urban continuously over time. The sheer size of such "big data" make this type of data worthwhile to investigate, despite the issues with its spatial precision and processing challenges. Applications of CDR data to travel demand analysis and modelling have been tested by many researchers (see, among others, Isaacman, et al., 2011 and Jiang, et al., 2013). Yin, et al. (2016) provide a good summary review of this work, describe in detail the data processing problem, and provide a very interesting example of the use of such data to construct an activity-based travel model for the San Francisco Bay area. The work of Gonzalez is also illustrative of the state of the art in this field (Gonzalez, 2008, Colak, et al., 2016, Jiang, et al., 2016a,b, among others).

The Montevideo iCity-South project has recently received a sample of Antel CDR data for Montevideo. At the time of this report's preparation the study team is just beginning to analyze these data to explore their usefulness for describing travel behaviour in Montevideo. The results of this investigation and recommendations for further study in this area will be the subject of a future report within the project's report series.

[23] Note that, despite the name, a phone call does not have to be made to trigger such records: the phone is always in communication with its network.

5.3 GPS Tracking Systems

GPS units can be installed in vehicles or transported by individuals, either as a dedicated GPS device or within GPS-equipped smartphones (the latter are discussed in Section 5.4). GPS applications mainly differ with respect to the precision of the spatial location and the rhythm of data collection. GPS data can be collected continuously, every second, every minute or when a specific event occurs. Typical devices will provide: longitude/latitude of position, UTC (coordinated universal time) time/data, speed of travel, direction of travel (heading), altitude/elevation, and indicators of the quality of the estimated position (e.g., number of satellites). Devices have become smaller and more efficient with time, a trend that is likely to continue. A review of the evolving application of GPS and other movement-aware technologies in travel surveys will be found in Lee-Gosselin, Shalaby and Doherty (2010). They point to parallel streams of development: *active* applications that interact with respondents to complete and/or validate automatically-logged data, and *passive* applications that give priority to extended observation periods using intelligent post-processing software, possibly calibrated during an "active" start-up phase.

GPS can be used in real-time applications (e.g., for operational control of highway facilities) or in planning applications (e.g., to analyze, *a posteri*, spatial-temporal traces that reflect the travel behaviour of people). They can also be used on floating-car studies to gather real-time travel time data for selected roads and paths within the transportation network.[24] GPS devices fall into two primary categories: in-Vehicle and portable systems.

In-Vehicle GPS Systems: Depending on their temporal resolution (i.e., the temporal frequency of recorded GPS data points), data from in-vehicle GPS systems can provide travel times and/or speed data on segments of transportation networks. Data can be collected using predetermined routes (probe cars) or using random routes (equipped fleet of vehicles such as taxis, minibus, shared cars for instance). Example applications include: Saunier and Morency (2011), MTO (2009), Greaves and Ellison (2011), Sharman and Roorda (2011), Sharman, et al. (2014), and de Fabritiis, et al. (2008). Auto manufacturers are increasing marketing data collected from their proprietary GPS-based onboard navigation systems.

Portable GPS Systems: In-vehicle GPS systems obviously can only provide data on vehicle-based (auto and truck[25]) movements. Portable GPS units are also available and can be used to collect continuous spatial-temporal data on the movements of people by all travel modes. These units are often used in a survey setting, usually in combination with a conventional household survey questionnaire. Many examples can be found in the literature, including: Abt SRBI (2011), Bohte and Maat (2009), Greaves, et al. (2010), Wolf et al. (2011) Frignani, et al. (2010), Chung and Shalaby (2005), Yang, et al. (2010), Kestens (2011), Geostats (2011), and Stopher, et al. (2011).

[24] Or travel times for other modes of travel (transit, walking, etc.).

[25] GPS units are very widespread in truck fleets for fleet management purposes and thereby potentially provide considerable data for analyzing truck movements. These data, however, are almost always proprietary and rarely are made available for public sector or academic analysis.

5.4 Smartphone Apps

While smartphone penetration within most urban areas has not yet reached the near-ubiquitous levels of "ordinary" cellphones (especially among the lowest income strata), these penetration rates are continuously increasing. Smartphones are an extremely interesting potential tool for travel behaviour data collection for many reasons. Most notably, the data collection device/technology is "self-supplied" by the respondent[26] and comes "pre-packaged" with all of the hardware (and much of the software) needed to gather travel behaviour data, as discussed further below.

Unique identifier of the unit: Data collected from a smartphone are anonymized, but it is usually possible to follow a specific unit (and, hence, a specific person, since smartphones almost always are used by a single individual) over space and time. Thus, the travel behaviour for a given (anonymous) individual can be continuously traced over extended periods of time. In such cases, home, work, school and other major "anchor" trip locations that are regularly visited can generally be identified from the smartphone traces alone, without active questioning of the phone owner. The ability to track an individual continuously over days and even weeks (possibly even months) could well be transformative in terms of our ability to understand and model travel behaviour.

Gathering trip time-space traces: Most importantly, the time-space traces of a smartphone user can be collected using a battery of "onboard" sensors which are standard in all smartphones. In addition to GPS, these most notably include algorithms for establishing the phone's position at each point in time, based on triangulation of signals from all GSM tower locations and Wifi networks within reach of the phone, combined with the GPS data. While results can still vary from one app to another and one geographic location to another, the optimized combination of GPS, GSM and Wifi data results in much better spatial precision in observed trip traces than is generally achievable by either GPS- or GSM-based calculations alone. Other onboard devices such as accelerometers can also be used to improve estimates of instantaneous speed, etc. As noted in Section 5.1, the ability to infer travel mode, route and purpose fundamentally depends on the accuracy of the time-space traces that the measuring device can achieve. The technology is also, of course, travel mode independent: as long as the smartphone is on, data will be recorded, regardless of the mode being used.[27]

Data Storage & Communications: Data storage and retrieval is always a technical challenge with any portable data recording device. Smartphones, again, natively provide much of the solution to this technical problem in that they can store in their memory a reasonable amount of data for subsequent transmission to an external central data repository, either via the phone's cellular data plan or by Wifi/Bluetooth connectivity when this is available.

Programmability: Apps can be readily downloaded onto smartphones to program the device to collect the desired data, perhaps do preliminary analysis onboard the device (e.g., aggregation of

[26] As opposed to, for example, portable GPS units which must be distributed (and retrieved from) respondents.

[27] As just three of many examples, Charlton, et al., (2011) and Grond and Miller (2016) used smartphone apps to gather data on bicycle route choice, while Lue and Miller (2018) used smartphone app data to model pedestrian route choice.

raw data to reduce onboard data storage and/or data transmission requirement), and to control data transmission from the smartphone to the central data repository. The app can also prompt the user in a variety of ways and query the user, either as part of an active data collection process (i.e., in real-time asking the user concerning the mode and/or purpose of a current trip), or as a stand-alone survey.[28]

As with any technology, smartphones are not without their technical challenges. These include:

- Very detailed collection of time-space traces can drain the phone's battery relatively quickly, much to the annoyance and inconvenience of the user. This often can be a primary reason for user's to discontinue use of the app. This problem is becoming less pressing with each new generation of smartphone with significantly improved battery lives. But it still remains a concern, with "optimization" of the precision of data collected and battery life is still an important app design consideration.
- Not all smartphone users have a data plan. In such cases, the smartphone must be able to store all collected data onboard until a Wifi connection can be made to upload the data to the central repository. Even if the user has a data plan, it is not necessarily a good design to make use of the data plan to transmit data without compensating the user in some way for the use of a resource that they are paying for.
- Unfortunately a universal smartphone operating system does not exits, and even within a particular proprietary operating system, succeeding generations of the operating system are rarely backward compatible. This makes app programming much more labour intensive and error-prone than it might otherwise be.
- Smartphone users must voluntarily load the app onto their phones and can delete the app at any point in time. Thus, this data collection method is subject to the concerns of sample frame definition, sample recruitment, response burden and retention common to all survey methods.

Smartphone app technology is improving almost daily, with apps increasingly in operational use by both private sector travel market research firms, academic researchers and government agencies. For a recent review of smartphone technology see (Rashed, et al., 2015). Building upon a preliminary pilot test of two apps in the late fall of 2014 (Miller, et al., 2016), the TTS 2.0 project has been very actively investigating the technology. In July 2016 it tested 17 different apps on a variety of smartphones in a series of controlled tests (Harding, et al., 2017b,c), and, at the time of this report's preparation, was about to go into the field for a major test field test of two state-of-the-art apps.

5.5 SENSORS

5.5.1 Introduction

An increasing variety of sensors suitable for observing various types of vehicle and person movements exist. These can be divided into three primary categories:

- Fixed road-based sensors.
- Transit vehicle-based sensors.
- Portable, person-based sensors.

[28] Smartphones are increasingly the device of choice for people completing web-based surveys.

These three sensor types are briefly discussed in the following sub-sections.

5.5.2 Fixed Sensors

A variety of sensor technology exists for detecting and counting vehicles (and, occasionally, depending on the technology, pedestrians and/or cyclists) passing the sensor. Such sensor-based count data is of somewhat limited use for travel behaviour analysis/modelling, since it does not provide any information other than the presence of a vehicle (or person) at a given point in space and time – the data generally tell us nothing about the trip origin, destination, purpose or route. Such data are largely used for more operational planning and control applications. Sensor count data are, however, briefly reviewed here, since they may sometimes be combined with survey or other data, and/or they may be useful as an independent data source for travel model calibration/validation purposes.

Sensor counting technologies that can be used to gather travel data include:

- Vehicle counters (loop detectors).[29]
- Video streams.
- Bluetooth sensors.

Loop detectors: Magnetic loop detectors embedded in roadways are commonly used in many cities to count the number of vehicles passing over the detector by time of day. These detectors normally are able to classify vehicles into cars (and light trucks) and heavy trucks. The time period for a single count varies from a few econds to 15-minutes or more, depending on the hardware and software used. Pairs of loops (double loops) placed closely together are usually used to gather speed data. The quality of the data and speed estimation are highly dependent on the calibration of the loops and will vary according to congestion level. Loop detectors are prone to failure and require replacing after roadway resurfacing. In many cities where these systems exist they are gradually be replaced by newer technology that often is more reliable, provides higher quality data and may also have lower lifetime costs. Similarly, many cities without such systems are opting for other systems as they upgrade their traffic monitoring and control systems, thereby by-passing this particular technology altogether.

Video streams: Cities often have parts of their freeway and arterial networks equipped with video cameras used to monitor or detect incidents and congestion levels. Cameras are also installed at many intersections to observe interactions between users. Various algorithms have been developed over time to extract data from traffic video cameras, by registration plate recognition (e.g., Friedrich et al. (2008) or by following moving objects based on there uniquely identifiable features (e.g., Saunier and Sayed (2006)). Although continuing to improve, as a generalization, however, such software generally does not yet appear to be ready for widespread, operational use for ubiquitous travel data collection (Hui, et al., 2017). This is should be an area for more active research and development, although, in addition to current software limitations, the cost of widespread deployment of video cameras (and communications for retrieving the video data) remains a serious obstacle.

[29] Traffic can also be counted by a variety of other means, such as pneumatic tube counters and manual (person-based) counts.

Bluetooth: Bluetooth is a proprietary open wireless technology standard for exchanging data over short distances. Multiple vehicles have systems installed that use this technology and many devices (computers, smart phones) are Bluetooth-ready. Bluetooth devices have a unique identifier that can be captured using an appropriate antenna. Typically, these antennas will be placed along a corridor and record time-stamped unit identification numbers. Using records from multiple antennas makes it possible to derive travel times between antenna locations. Hence, depending on the particular, data could be used to derive O-D trips and partial route choices of a sample of vehicles (i.e., typically for a travel corridor using a cordon of sensors). While the Bluetooth data usually have been used to date to monitor vehicle movements (Roorda, et al., 2009; Bachmann, et al, 2013), Malinovskiy et al. (2012) investigated the feasibility of using Bluetooth for pedestrian studies using two separate sites. Their results suggest that "*given sufficient populations, high-level trend analysis can provide insights into pedestrian travel behaviour*."

5.5.3 Transit APC and AVL Systems

Automatic passenger counters (APCs) are increasingly being used to provide automated counts of passengers boarding (and possibly alighting) from transit vehicles. The major challenge with such devices is to accurately differentiate between boardings and alightings when the same door is used for both movements. Similarly, the installation of GPS units onboard transit vehicles to accurately track these vehicles in time and space (AVL) is increasingly common in many transit systems. Onboard GPS serves many purposes (facilitate onboard stop annunciators, real-time tracking and operational control of vehicles, etc.), including coordinating with APCs to provide accurate spatial-temporal locations for the recorded boardings and alightings. APC and AVL systems can also be integrated with smartcard data to enrich these data.

5.5.4 Emerging Applications of Portable Technologies

Of growing significance is the miniaturisation of a number of technologies that, singly or in combination, promise to improve the travel survey toolkit. Some examples, among others, include:

- Wearable digital cameras (e.g., Kelly et al., 2011).
- Portable accelerometers (with GPS units or smartphones) to aid in detecting trip mode and stop locations (e.g., Schüssler, et al., 2011).
- New micro-electro-mechanical systems for vehicle and, possibly, person tracking (Noureldin, et al., 2009).
- Radar.

5.6 Transit Smartcard & Other Mobility Service Usage Data

5.6.1 Introduction

Transit fare payment systems and service reservation and fare payment systems for other types of mobility providers (taxis, bike-share systems, etc.) are potential sources of large, passive, "choice-based" data streams of usage of the given system. These data sources are briefly discussed in the following two sub-sections.

5.6.2 Transit Smartcard Data

It is very common for public transit agencies in large cities world-wide to use some form of smartcard for fare payments by their system users. Smartcard technologies and capabilities vary considerably from one implementation to another. Generally implemented for fare collection and accounting purposes, smartcard data obviously have the potential to provide useful data concerning both transit system ridership and service performance. Pelletier et al. (2011) list a range of possible smartcard data applications:

- ***Analyze travel behaviour of transit users****:* Agard et al. (2006) examined travel behaviour using multiple days of observation for individual smart cards in Gatineau, Quebec. Using a similar dataset and data mining techniques, Morency et al. (2007) measured transit use variability. Instead of relying on a typical weekday, analysis conducted with smart card data can allow observation of variability among days, weeks, seasons, and years. Park and Kim (2008) also demonstrated the possibility of using these data to better understand user habits.
- ***Assess turnover rates:*** Analyses were also conducted with smart card data to assess how well the transit network retains its users (Bagchi and White, 2005, Trépanier and Morency, 2010).
- ***Derive origin-destination matrices on the transit network*** (Munizaga et al., 2010).
- ***Forecast travel demand***: using historical data, Park and Kim (2008) created a future demand matrix.
- ***Understand impacts, on demand, of various incidents /events/context:*** Using one month of data Chu and Chapleau (2011) illustrated how transaction data are impacted by the occurrence of events. Descoimps et al. (2011) analyze the impact of weather on transit demand for various population segments.
- ***Enrich household travel surveys:*** Bayard et al. (2008) suggests that smart card data could be used to enrich classical OD survey data and that OD survey data could enrich smart card data. Trépanier et al. (2009a) compared these two datasets for a specific region and have identified many issues. There are few examples of data fusion in this context but fusion techniques have been examined by various researchers (see Chapter 6).
- ***Estimate transit performance indicators:*** Trépanier et al. (2007) proposed a method to impute a destination point and, using this information, developed a load profile for each transit route. Trépanier et al. (2009b) and Reddy et al. (2009) also demonstrated the usability of smart card data to estimate a set of transit performance indicators (veh-km, mean speed, veh-hr, mean trip distance, schedule adherence) for various contexts. Trépanier and Vassivière (2008) propose an intranet (internet tool only accessible to the organization's employees) tool with various operational statistics.

Smartcards obviously share the limitation of "choice-based" surveys in that they only provide information concerning current transit users. They also share the problem typical of most passive datasets in that they generally do not provide any information concerning the trip-makers. The representativeness of the smartcard data can also be an issue in systems in which smartcard users are a minority of all transit riders (although this concern obviously decreases as the "market share of smartcard users grows within the population). Also, as with many passive datasets, the sheer volume of data generated each and every day by smartcard transactions can be overwhelming to analyze without proper data management and analysis software.

On the other hand, smartcard data share most of the attractive features of passive data streams in terms of providing very large, continuous samples of transit usage across the transit system at very low marginal cost once the data collection and management system has been set up. In particular, since the smartcard data collection system is implemented (and paid for) for fare collection purposes, the data are available for "relatively free" for analysis and modelling purposes.

Since smartcard systems are designed for fare collection purposes, however, they often are not "optimized" from a ridership analysis and modelling perspective. In particular, many (typically flat-fare) systems only require the rider to "tap on" with the smartcard when first boarding the transit vehicle; they do not have to "tap off" when disembarking from the vehicle. Similarly, within system transfers from one route to another may or may not always require a new "tap in" when boarding the next transit vehicle. In such cases, trip destinations (and possibly multi-line paths through the network) need to be inferred through complex analysis algorithms, since they are not directly observed in the data. Linkage between smartcard tap ons/offs with GPS/AVL vehicle location data also varies considerably from one system to another. Also note, even in the best case in which precise stop-by-stop tap ons and offs are recorded, the data only identifies where the trip-maker board and alighted from the transit vehicle. They do not identify the actual trip origin and destination locations, which must also be inferred using other data (land use, etc.) to inform these inferences.

Montevideo has a well-designed smartcard system, which currently is used by approximately 65% of its transit riders. A sample of the Montevideo smartcard data is being analyzed as part of this project to explore its usability for a variety of applications, such as those listed above. Report number 3 in this project's report series will report on findings to date of this investigation.

5.6.3 Other Mobility Service Usage Data

Computerized, passive datasets documenting usage of other modes of travel also are increasingly being collected by a variety of organizations. Car-sharing, ride-sharing and bike-sharing services all maintain detailed transaction records (e.g., where the bike is picked up, where it is dropped off, the time taken to travel from pick-up to drop-off, etc.). The implementation of GPS units in many taxi systems, along with computerized data collection and management means that large datasets concerning taxi movements are now being collected in many cities. And new mobility providers such as Uber and Lyft, of course, also maintain computerized datasets of service usage and performance. The availability of such datasets, however, to academic and public agencies for use varies dramatically from city to city and organization to organization.

5.7 3RD PARTY PASSIVE DATA STREAMS

In addition to transit smartcard and other mobility service usage data, many passive data streams exist that are collected continuously for a wide variety of purposes (both transportation- and non-transportation-related) by many different (typically private) organizations. These notably include:

- Road traffic data.
- Credit card data.

- Social media.

Real-time road traffic data are collected by a wide variety of private sector companies (Google, Waze, Inrix, TomTom, etc.). These data are largely "crowd sourced" from trip-makers' smart devices' (smartphones, etc.) GPS and other signal traces, aggregated by the company, and then "fed back" to trip-users in terms of real-time route choice and travel time information. Increasingly public agencies are purchasing such data from one or more provider for a variety of analytical purposes. The cost, quality, on-going availability and usability for a variety of purposes all vary from one case to another and also are changing over time as this "industry" evolves and matures and as governments evolve working relationships with them.

Although not currently often used for transportation analysis and modelling purposes, credit card data contain considerable information about household trip-making for out-of-home shopping (goods and services), recreation and social activities, as well as out-of-home versus in-home shopping (and, in general, household expenditure patterns). Credit card companies are clearly analyzing their data in-house, exploring commercial opportunities for use of these data. Whether such data might be made available to governments and academics for transportation analysis and modelling purposes is unknown at this time, but this may be a question worth exploring.

The proliferation of social media usage (Facebook, Twitter, Instagram, etc.) raises the question as to whether social media posts can be mined for useful information concerning travel behaviour: where do people go to lunch? Where do they socialize? Etc. Research is being undertaken exploring such issues, but, it would appear that this work is much too speculative at this time to consider incorporation into any practical, operational data collection exercise for an urban region.

Another possible use of social media is for recruitment of respondents for web-based (or perhaps other types of) surveys. In this approach, however, the sampling frame cannot be controlled. Respondents are generally recruited through snowball sampling (also known as chain referral sampling or respondent driven sampling, Gile and Handcock, 2010). An invitation to participate is first sent to a convenience sample that acts as seed for recruitment. The process is hence non-probabilistic and depends on the initialization process (selection of seeds). There are few documented examples of using these media to construct a sample. Benfield et al. (2006) discuss the use of various media to recruit respondents. According to these authors, "*different recruitment procedures can have different effects on the resulting sample and (b) the right recruitment procedure, with some luck, can yield interestingly large samples for the study*". In one of their surveys, the authors used a snowball sampling technique and sent an initial recruitment e-mail to 60 friends, colleagues and family. They managed to yield 189 responses in one month. They have estimated that the snowballing effect had stopped after the third or fourth iteration. While perhaps useful/convenient for research purpose, it is unlikely that sufficiently statistically representative samples can be generated by such methods for large-scale operational planning applications.

CHAPTER 6
THE CORE-SATELLITE DESIGN PARADIGM

6.1 INTRODUCTION

There is an increasing trend in travel surveys to use "multi-instrument" surveys in which two or more surveys (or combinations of a survey and other data collection methods such as GPS traces) to capture different elements of the behaviours being studied. This approach reflects both the complexity of these behaviours (which may not be able to be captured within a single instrument) and the need to keep response burdens within a single survey within reasonable levels. It also recognizes the variety of current and emerging data collection methods, each with their individual strengths and weaknesses, that, ideally, should be collectively exploited for optimal effect within a coordinated and comprehensive data collection program.

A particularly attractive and generalizable multi-instrument survey design is the *core-satellite design* proposed by Goulias, *et al*. (2011) (see Figure 6.1). This paradigm is defined in detail in the next section. Section 6.3 then presents prototypical "use case" to illustrate how the core-satellite design paradigm can be applied in practice. The core-satellite paradigm depends critically on advanced data science-based *data fusion* methods for combining two or more datasets. These methods are briefly introduced in Section 6.4.

6.2 PARADIGM DEFINITION

As illustrated in Figure 6.1, the core-satellite approach involves the following components:

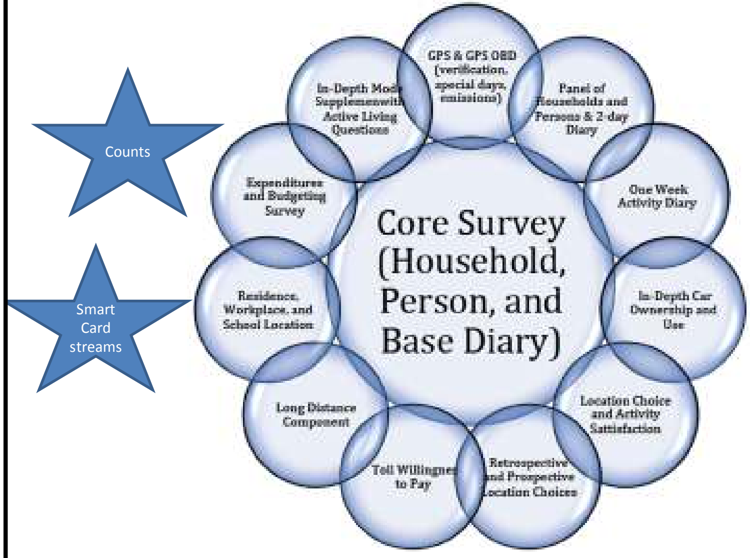
- A *core survey*, which is a large-sample survey which gathers primary information concerning the respondents and their key behaviours.
- Any number of *satellite surveys*, which are smaller-sample, more focussed surveys (or other data collection methods) designed to gather more detailed information about specific behaviours of interest.
- Additional, independent, *complementary* surveys/datasets that might be used to augment the core-satellite database, but may not be directly linkable to the core-satellite data.[30]

Characteristics of the core survey include:

- It includes key data that are fundamental to the agency's primary policy/planning needs (their core business).
- It includes attributes of the respondents that permit core data to be linked to common variables in the satellite surveys so that core and satellite data can be jointly used. This linkage is direct if the satellite survey is a sub-sample of the core survey. If the satellite respondents are not a sub-sample of the core respondents, then these common variables must permit fusing of core and satellite data.

[30] These independent datasets have been added to Goulias, *et al.*'s original paradigm, which only include the core and satellites.

- Its sample size is large enough to make statistical inferences concerning variables of interest.
- It is expandable to make statements about the full population.
- It is consistently applied over a large geographical region (e.g., the urban region of interest and, ideally, its immediate hinterland).
- It is stable (but not necessarily static) over time and is applied relatively frequently (or continuously) so as to provide consistent time-series data.
- It is relatively short, so as to minimize response burden and to permit large sample sizes to be cost-effectively collected.


Counts
Smart Card streams
Core Survey (Household, Person, and Base Diary)
GPS & GPS OBD (verification, special days, emissions)
Panel of Households and Persons & 2-day Diary
One Week Activity Diary
In-Depth Car Ownership and Use
Location Choice and Activity Satisfaction
Retrospective and Prospective Location Choices
Toll Willingnes to Pay
Long Distance Component
Residence, Workplace, and School Location
Expenditures and Budgeting Survey
In-Depth Mod Supplemenwith Active Living Questions

Adapted from: Goulias, et al., (2011)

Figure 6.1: The Core-Satellite Multi-Instrument Survey Design

Satellite surveys provide the opportunity to enrich the core dataset by filling gaps and adding detail to the core that would not be feasible and/or cost-effective to collect as part of the core. They can be used to gather data for special models that can be linked to core behaviours or to augment the core sample for small sub-populations of special interest. They must be statistically linkable to the core survey, either by being sub-samples of the core or via common attributes of the respondents in the two datasets. Types of satellite surveys include:

- Extra questions and/or instrumentation of a subset of the core sample.
- An additional survey (separately administered subsequently to the core survey) of a subset of the core survey.
- Increased (stratified) sampling of specialty populations (the elderly, transit riders, etc.).
- Passive data sources (e.g., smartcard data), where a reliable means of linking to the core survey (through common variables) exists.
- Surveys conducted on different samples but with common data items for linking/fusion.

Satellites may be either "one-time" surveys or on-going in nature. They may or may not cover the full geographic area of the core. They can be much more flexible in terms of choice of method and can be a means for experimenting with new methods. They may well involve greater response burden, more detailed methods, etc. Examples of possible satellite surveys might include:

- Car type/usage.
- Bicycle ownership and usage.
- Walking behaviour.
- Multi-day/week trip or activity diary.
- Residential mobility and dwelling type choice.
- Heath, fitness, physical activity, active transportation
- Route choice (chosen routes, choice sets considered; by mode
- Parking location/usage/cost

The third category in the paradigm: independent, complementary surveys/datasets consists of any other data, collected by any means or obtained from any other source which is part of the agency's overall database. Such data generally will not be fusible with the core-satellite database, except in special cases in which common variables and sufficient commonality in sampling frame, time of collection, geography, etc. permit some form of integration. This category is added to the Goulias, *et al*. paradigm for the sake of completeness and to emphasize the need to think comprehensively about one's overall data collection, management and analysis program. Complementary surveys and datasets might include:

- Special generator surveys (airports, hospitals, etc.).
- Visitor/non-resident surveys.
- Taxi usage records.
- Smart card records.
- Transit onboard/ridership surveys.
- Cordon/screenline counts.
- User satisfaction surveys.
- Census Place of Residence – Place of Work (POR-POW) questions, including the usual mode to work question.

The core-satellite paradigm is an extremely flexible and generalizable approach to meeting different agency needs. It is defined around content (i.e., what content is core, what can be collected via a satellite process) rather than method. Different agencies will use different methods for both their cores and their satellites, depending on their data needs, resources, etc. These methods can and should evolve over time (with satellite methods generally evolving more quickly and more often than the core), while recognizing the need to maintain data compatibility over time for time-series analysis and consistency in modelling. One approach for maintaining such compatibility is to use both old and new methods during transition periods so as to be able to test in a controlled way for the impacts of the changes in methods on survey results.

The core-satellite paradigm allows each agency to think through the structure of its own data collection and management methods in a comprehensive, consistent manner to best meet its own analysis, modelling and planning needs. Ideally, if all (or at least many) agencies within a given urban region are working within this same framework, this will facilitate consistent, effective collaboration in data collection and sharing, with each agency undertaking components of the overall data collection task that it is best suited to do and has the most direct interest in, while at the same time providing a basis for sharing data among agencies (my core becomes your satellite, etc.).

Further, the approach permits a systematic "enrichment" of the database over time by permitting the incremental addition of new satellites (and/or complementary data) over time as need, time and resources permit. That is, it permits agencies to define priorities within their overall desired data collection program and to build their data collection program in a structured, manageable way that recognizes these priorities and short-run constraints while working in a systematic way towards long-term goals. Satellites also potentially provide the flexibility to address "hot-button issues" in a more cost-effective, flexible and timely fashion that would otherwise be possible. In particular, the satellites provide the opportunity for both testing new methods and/or addressing immediate issues with a quicker response/turnaround time than the five-year cycle that is typical of many current Canadian travel survey programs. Finally, the core-satellite approach provides a flexible framework for the inevitable "stitching together" of datasets derived by various actors and various time for various purposes.

6.3 EXAMPLE APPLICATION: TRAVEL DEMAND MODELLING

All large and most medium-sized urban regions require a regional travel demand forecasting model for a variety of short-, medium- and long-range planning applications. This model needs to be multi-modal, deal with all trip purposes, cover the entire urban region, work at a spatially disaggregate level (at least at a traffic zone level), and deal with the time period(s) of planning interest.

Figure 6.2 sketches an example core-satellite data design for a typical travel demand modelling exercise. The core survey is a large-sample home-interview survey. This survey collects key data about households and persons across the urban region, as well as details for each trip made by each person surveyed. Also shown in this figure is the large number of ancillary data concerning the road and transit systems required for model development and calibration.

Typical home interview surveys focus on gathering detailed person trip information and selected, key person and household attributes. By limiting the survey to these core data items large sample sizes can be cost-effectively gathered that provide statistically reliable population-level estimates at traffic zone levels of spatial precision. The trade-off with this approach is that limited or no information is collected about certain types of travel behaviour (vehicle usage, walk/bike trips, HOV usage, etc.) and/or sub-groups within the population of particular policy interest (e.g., the elderly). Expansion of the core survey to include a large number of additional questions and/or sample size is generally not a practical, cost-effective option.

Satellite surveys can address this problem. Smaller-sample, targeted, special-purpose surveys can be undertaken to address special modelling/analysis needs. These can either be add-on/follow-up surveys of a sub-sample of respondents to the core survey or stand-alone surveys of a fresh sample of respondents. Add-on surveys have the advantage that they can be directly linked to the core survey responses and so can be fully integrated into the core survey dataset, whereas stand-alone surveys must be fused to the core survey dataset through the use of common (overlapping) variables in the two. On the other hand, stand-alone surveys can be advantageous in that they:

- Can be undertaken at a convenient point of time, independent of the core survey, in particular, after the core survey as need for the data arise.
- Permit specialized sampling frames, focused on the behaviour or sub-population of interest, to be used.
- Facilitate the use of relatively complex survey instruments.

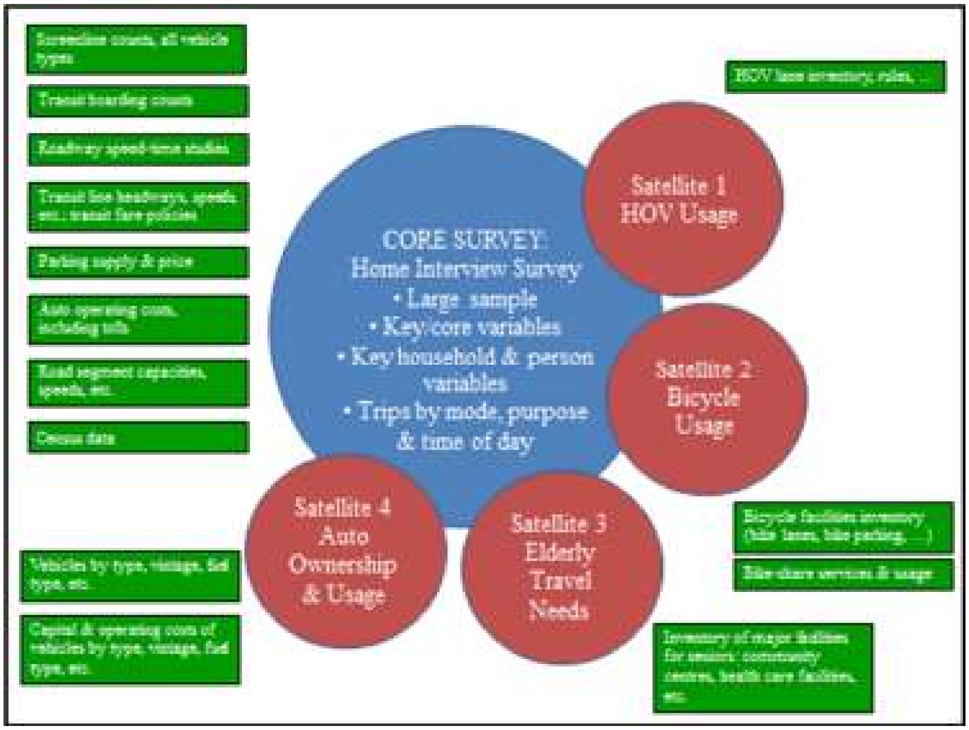

Screenline counts, all vehicle types
Transit boarding counts
Roadway speed-time studies
Transit line headways, speeds, etc.; transit fare policies
Parking supply & price
Auto operating costs, including tolls
Road segment capacities, speeds, etc.
Census data
Vehicles by type, vintage, fuel type, etc.
Capital & operating costs of vehicles by type, vintage, fuel type, etc.
HOV lane inventory, rules, …
CORE SURVEY: Home Interview Survey
• Large sample
• Key/core variables
• Key household & person variables
• Trips by mode, purpose & time of day
Satellite 1 HOV Usage
Satellite 2 Bicycle Usage
Satellite 3 Elderly Travel Needs
Satellite 4 Auto Ownership & Usage
Bicycle facilities inventory (bike lanes, bike parking, …)
Bike-share services & usage
Inventory of major facilities for seniors: community centres, health care facilities, etc.

Figure 6.2: Travel Demand Modelling Data Design

In the hypothetical example shown in Figure 6.2, four satellite surveys are shown for the sake of illustration. These deal with HOV usage, bicycle usage, travel behaviour of the elderly and auto ownership and usage behaviour. Gathering detailed information for each of these within a general core survey adds considerably to the cost of the survey and the response burden on the respondents. Also, many questions are not relevant to all respondents in the core survey (e.g., non-HOV users, non-elderly, etc.). Well-designed satellite surveys can much more cost-effectively address specific behaviours/issues such as these relative to trying to cover them within an omnibus survey. A key in all cases, however, is to ensure that variables are collected in each satellite survey that all satellite data to be linked to core data in useful ways.

6.4 Data Fusion Methods

6.4.1 Introduction

Many transportation datasets provide useful information about travel behaviour or transportation system performance, but they often contain data gaps that limit modelling or analysis capabilities. Chalasani and Axhausen (2005) classify transportation data into the following categories: travel survey data, transportation infrastructure data, transportation system performance data and geographic data. Each of these categories may contain data of different types and contexts. However, in almost all transportation analysis, we need data belonging to more than one type. Moreover, with increasing difficulty in constructing appropriate sampling frames and reaching diverse target populations, as well as the increasing amount and quality of data required for advanced transportation modelling (among other planning applications), it is becoming impossible to collect all the data required during a single survey or with a single method (Bayart et al. 2008). Given this, integrating or *fusing* multiple data sources together to create unified databases is increasingly important in urban transportation planning analysis and modelling. The purpose of this part of the report is to discuss the data fusion problem and the variety of methods available for integrating/fusing individual datasets to create new, more comprehensive datasets.

The concept of data integration/fusion involves creating a new, more comprehensive dataset by joining/combining multiple datasets to fill information gaps (Gautier, 1999; Saporta, 2002). A common example of this is the use of Census data to add socio-economic variables (income is a typical example) to travel survey records that are missing these variables. Compilation of comprehensive databases by fusing multiple sources can enhance the useful information available to transportation system users, transportation system operators/service providers and system planners (Amey et al. 2009), and, in some cases may even eliminate the need to undertake new surveys or data collection efforts by using available data to construct the needed information. It also may be possible to combine datasets collected in different time periods and in different spatial contexts, as long as there are some common elements among the individual datasets.

Section 6.4.2 provides an overview description of the data fusion problem and introduces several key issues that generally need to be addressed in data fusion exercises. Section 6.4.3 establishes a typology of data fusion contexts. For more detailed discussion of specific data fusion methods, see, among other, Miller, et al. (2012).

6.4.2 Data Fusion: Definitions & Issues

Figure 6.3 illustrates the typical data fusion problem in which two datasets (A and B) exist that share a certain number of common variables (X_A and X_B). One of the datasets contains additional variables (Y_A) that are not found in the second dataset. The problem is to impute the values of these missing variables (Y_B) given the relationships (correlations) between the X and Y variables observed in dataset A. In this problem, the X variables are called *common variables*, the Y variables are *target variables*, and datasets A and B are referred to as the *donor* and *receptor* datasets, respectively.

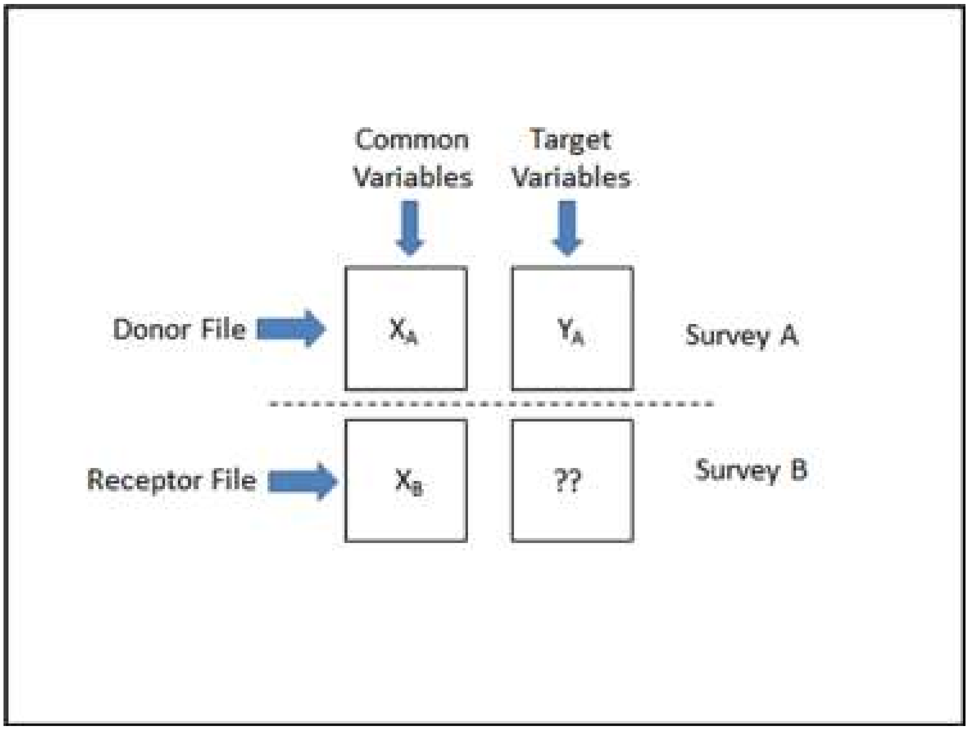


Figure 6.3: The Data Fusion Problem (D'Ambrosio, 2007**)**

All data fusion problems one way or another involve using the observed relationships (joint distribution) among the X_A and Y_A variables to *impute* appropriate values of the Y_B variables, given the observed X_B variables (Gilula et al. 2006). Building on this basic task, several major types of data fusion applications exist, including:

- Using the combined A and B datasets once the Y_B variables are known for modelling or analysis purposes. Using a combined revealed preference/stated preference (RP/SP)[31] dataset for mode choice model estimation is an example of this type of application.
- Using the updated B (receptor) dataset for modelling or other analysis purposes. In this case, the B dataset is typically a file of travel survey records and the A dataset contains information that is missing in the original B dataset but that is needed for the given application. Imputing incomes for trip-makers in a travel survey dataset from census data is a typical, simple example of this type of application.
- Creating a *synthetic population* for use in microsimulation models. In this case, the original B dataset typically consists of a set of aggregate tables of data (household size by income by census tract; person age by gender by labour force status; etc.) for the entire population being modelled. The dataset A contains much more detailed information for a sample drawn from this population (e.g., a set of individual person or household records, in which each record corresponds to a specific person or household and contains the actual attributes for this agent – age, gender, labour force status, etc.). A Public Use Microdata (PUM) file of actual individual census records (with spatial and individual

[31] A revealed preference survey gathers information on actual travel choices made by a respondent. A stated preference survey asks respondents to state what they would do in a given hypothetical situation. A combined RP/SP survey augments RP questions about actual choices with a set of SP experiments that are linked in way or another to the RP choices that the respondents make.

identifiers deleted) provided by most census agencies, or other types of surveys are typical examples of such a donor dataset. This dataset provides a sample estimate of the joint distribution of the full set of variables (X and Y) of interest and can be used in a variety of ways with the aggregate "marginal" population tables in the receptor file to synthesis the desired set of disaggregate agents for this population (Pritchard and Miller, 2012).

- If the A and B datasets are repeated cross-section surveys with the same sets of variables, then these datasets can be combined to create a *pseudo-panel*. In this pseudo-panel observations in one dataset are linked to observations in the second dataset that match sufficiently closely to be treated as if they were the "same" respondents in the two datasets. Thus, the linked observations can be treated as if they were observations from a panel survey. Such pseudo-panel data can provide the basis for dynamic, process-based modelling (Bernard et al. 2012). Construction and use of pseudo-panel data is not yet generally used in operational planning practice, but is receiving increasing attention in research applications. Pseudo-panels have the potential to provide deeper insights into changes in travel behaviour over time than might be obtained from simple analysis of repeated cross-section data while avoiding the cost and difficulties of panel surveys.

Thus, data fusion requires:

- Two or more datasets that collectively contain the required information.
- At least one common variable among the datasets.
- A suitable statistical method for undertaking the desired fusion.

A major challenge in data integration is the different levels of aggregation that often exists in different datasets. Differences in aggregation may be spatial (e.g., different zone systems), temporal (e.g., different time periods) or semantic, i.e., the categorization of the variables themselves (e.g., use of different categories for age, income, occupation/industry definitions, etc.). These differences need to be reconciled if the datasets are to be correctly fused. Ever-increasing planning requirements to address complex policy issues, combined with the availability of more detailed datasets and increased computational power leads to the need/opportunity for using high resolution/microscopic travel behaviour data across spatial, temporal and semantic dimensions to develop advanced analysis/modelling methods. This further accentuates both the need for data fusion techniques to construct the required complex, detailed databases and the technical challenges involved. The aggregation levels of different datasets play a crucial role in success and failure of such data fusion efforts (Polak, 2006; D'Orazio et al. 2006).[32]

6.4.3 Data Fusion Contexts

Data fusion for transportation analysis can be broadly classified into the following exhaustive set of categories:

- ***Spatial context***: Data fusion based on common spatial contexts (neighbourhood features, regions, cities, neighbourhoods, census tracts, traffic analysis zones, etc.). That is, the

[32] Whenever possible, care should be taken in designing data collection efforts to gather data at as disaggregate level as possible so as to facilitate matching of these data at various levels of aggregation (i.e., data can always be aggregated but it cannot be disaggregated with certainty). Similarly, accessing disaggregate base data (as opposed to published more aggregate results), should obviously be pursued whenever possible.

common variable between the two datasets is the zone or other spatial indicator that allows one set of zonal variables to be merged with a second set.

- ***Temporal context***: Data fusion based on a common temporal context (same year, month of the year, week of the month, day of the week, time of the day, etc.).
- ***Semantic context***: Data fusion based on common semantic contexts (same age groups, same gender groups, same household size groups, same auto ownership levels, same occupation groups, same income groups, etc.)
- ***Mixed spatial, temporal and/or semantic context***: Data fusion based on combinations of spatial, temporal and semantic contexts.
- ***Survey mode context***: Data fusion for datasets collected through different modes of surveys (face-to-face, telephone interview, web-based, etc.).
- ***Data type context***: Data fusion based on data types for advanced econometric modelling (e.g., combination of revealed reference (RP) and stated preference (SP) data).

Each of these contexts is briefly discussed below.

Spatial context: Expanding sample survey data to represent the whole population of a study area is the simplest and most widely used example of data fusion based on a common spatial context. Calculation of weighting factors for the survey observations based on census data of the study area allows the sample dataset to be expanded to represent the whole population. In such cases, calculation of the expansion weighting factors is very straightforward, since it is simply the inverse of the selection probability of the individual in the sample from the sample frame.

Matching datasets via postal codes, census tracts and traffic analysis zones to supplement household travel survey data with land use data, aggregate census information, etc. is another common example of data fusion used in transportation analysis. The major challenge in these applications occurs when the two zone systems (spatial aggregations) involve overlapping boundaries. For example, census tracts and traffic zones often are not simple aggregations of one another, so that a given traffic zone may intersect with multiple census tracts, or *vice versa*. In such cases, at least three options for reconciling the spatial aggregations exist:

- Aggregate both zone systems up to a new, common super-zone system in which each of the original zones is wholly contained within one and only one super-zone.
- Use GIS-based calculations to split the units of observation (persons, households, etc.) in one zone and allocate these units to the portions of the other zone with which the first zone intersects. This allocation of units is often done on the basis of the fraction of the zone's area that lies in each of the overlapping zones. This implies a uniform distribution of the units within the zone, a reasonable assumption in the absence of additional information. If information exists to alter this assumption it should be used. For example, the available of detailed information concerning dwelling unit distributions within the zone (from Google Earth©, DMTI Spatial datasets or other similar sources) may permit non-uniform allocations to be made with confidence.
- Use GIS-based calculations to split both zone systems into a common set of fine-grained grid cells to which the zone unit/attributes can be allocated. Again, either a uniform distribution assumption or data-driven allocations from Google Earth/DMTI type datasets can be used to perform these calculations.

Temporal context: Attaching transportation level-of-service variables (travel times, costs, etc.) for a given time period that have been computed using road and transit network assignment models to travel survey records is a very simple example of temporal context fusing, in which one of the common variables is the time period within which the trip occurred. Combining different count datasets by time period is another common and important temporal context fusing.

Although not often discussed, temporal aggregation is always an issue in any transportation analysis, given that trips occur continuously over time and travel times are continuous variables. When analysing trip-making by time period one is always making the assumption of uniform trip rates, mode splits, etc. throughout the time period. With respect to the fusion problem, inconsistent time period definitions obviously pose problems similar to those caused by inconsistent zone systems (e.g., it may be the case that two sets of traffic counts use different peak-period definitions).

Semantic context: Creating income variables to attach to travel survey data based on census information is a common example of data fusion within a semantic context. Income variables (average income, median income, etc.) can be generated from census data according to occupation types, job categories, neighbourhood or traffic analysis zone, etc. that are common between the Census and travel survey datasets. In such cases, income variables for each observation in the travel survey data can be directly imputed by assigning Census incomes to persons or households in the survey that have the same values of the common variables (same household size, occupation type, etc.). Direct imputation and more complex semantic fusion techniques are discussed further in Section 4.

Aggregation issues also exist in semantic fusion contexts. A very typical case involves two data tables that share a common variable (e.g., age) but use different aggregation categories. If the two categorizations map consistently into one another (e.g., one uses 5-year intervals and one uses 10-year intervals) then the finer categorization can simply be aggregated to the coarser categories. If the two categorizations overlap with one another, however, methods similar to the spatial aggregation case must be used, i.e., one of:

- Aggregate both categorizations to a common set of super-categories that both original categorizations unambiguously map into.
- Split one of the categorizations and allocate portions to the other categorization using appropriate assumptions (again often a uniform distribution assumption is made).
- Break down both categorizations to a finer, common system and statistically allocate data from both categorizations to the common finer system. For example, in the case of age categories, both categorizations could be disaggregated down to single years.

Survey mode context: Tradition survey methods of face-to-face, mail-back and telephone interviews are no longer the only ways by which we collect data, with innovative survey methods such as computer- and web-based surveys and other new technologies (GPS, mobile phones, etc.) being used for collecting travel behaviour data. In many cases, multiple surveys are used to collect data from the same spatial and/or time contexts. Integrating data collected by different survey modes raises a number of statistical and practical issues. In particular, each survey method has its own set of instrument biases and other data quality issues. Thus, even in cases in which the fusion of data collected in a multi-mode framework may appear to be straightforward, analysis and modelling using the fused dataset should recognize the mode-specific randomness and errors existing within this dataset.

Data type context: Similarly, stated preference (SP) data are increasingly being used in place of or to supplement traditional revealed preference (RP) data. In particular, it is very common to combine RP and SP data in the joint estimation of travel demand models (typically disaggregate mode choice models) in order to take advantage of the strengths of both types of data. This is a particular form of data fusion in which two datasets are both used to determine the parameters of a single model, thereby incorporating information from both datasets within the same model. Not only does this represent a multi-mode survey application, which, as discussed above, implies varying survey instrument biases across the two dataset, but SP data are hypothetical responses gathered in an experimental setting, as opposed to RP data which are observations of actual choices made by respondents in real-world contexts. Thus, unobserved differences can be expected between the nature of the variables in the two datasets (stated versus actual choices; hypothetical versus real-world explanatory variables) that need to be accounted for in the statistical estimation process, as well as in the interpretation of the survey results themselves.

Mixed contexts: Data fusion problems more often than not involve multiple contexts. The example used above concerning matching trip travel times to travel survey records actually involves temporal (trip start time), spatial (trip origin and destination) and semantic (trip mode) components. Mixed contexts are not necessarily more difficult to deal with than single contexts (as the trip ravel time example illustrates), although the combinatorics of dealing with aggregation issues across multiple dimensions may in some cases be challenging. A simple example of the latter situation might be two datasets in which both age categories and trip mode choice definitions both involve different aggregations that make fusing the two datasets more problematic.

CHAPTER 7
TRANSPORTATION DATA COLLECTION IN LATIN AMERICA & NEXT STEPS

7.1 INTRODUCTION

This report has presented a summary review of the current state of the art and practice in transportation data collection methods. Section 7.2 builds on this review to highlight a number of Latin American-specific issues and challenges that need to be addressed when designing a data collection program for one or more Latin American urban regions. Section 7.3 concludes the report with a few comments concerning "the way forward" within the project towards developing a general transportation data collection program for Montevideo as a "prototype" for more general application by CAF within its OMU partner cities.

7.2 LATIN AMERICAN-SPECIFIC CHALLENGES

7.2.1 Introduction

As in any region, local context-specific challenges can exist in designing a transportation data collection program for a given urban region within Latin America. Such challenges become magnified if one wishes to establish a common data collection program for implementation within multiple urban regions, as would be desirable for the OMU program, given the heterogeneity in conditions across Latin American cities. At least three major issues exist that will affect data collection design and implementation:

- Heterogeneity in socio-economic and living conditions.
- Informal/privatized transit services.
- City-to-city and country-to-country differences in government structures and capacities.

Each of these issues is discussed briefly below.

7.2.2 Population Socio-Economic Heterogeneity

As discussed in detail in this report, the validity and utility of any data collection exercise depends on the representativeness of (lack of bias in) the data collected. While the exact definition of "representativeness" will vary from one application to another, generally representativeness means:

- All socio-economic groups need to be appropriately represented, consistent with their proportions within the overall population.
- Adequate spatial coverage of the trip-making of interest, at a level of spatial precision suitable for detailed analysis and modelling of travel patterns (traffic zone / census tract or better).
- Observation of travel by all modes of interest for all trip purposes.
- Data on travel for the entire time period of interest, typically at least a 24-hour weekday.

Each data collection method has its own challenges, especially with respect to ensuring socio-economic, spatial and modal representativeness. For any type of household/person-based survey (including web-based surveys), the critical, inter-connected concerns are:

- Establishing a representative sampling frame;
- Being able to contact people within this sampling frame in an unbiased and cost-effective way;
- Inducing a large percentage of those contacted to participate in the survey; and
- Ensuring that the survey responses are complete and accurate.

In North American surveys these are increasingly challenging requirements to meet, for a variety of reasons, resulting in decreasing response rates and increasing costs in recent years. In Latin America these challenges can be magnified by the significant socio-economic heterogeneity in many cities, ranging from the extremely poor to the extremely rich. Both ends of the socio-economic spectrum can pose significant challenges in terms of accessing potential respondents and inducing them to participate in a survey. The very poor, often living in informal/slum settlements, may in some cases at least only be contactable by face-to-face visits, which may pose security and other logistical concerns for interviewers. The rich, often living in gated communities/buildings, may also be very difficult to contact.[33] Both groups, even if contactable, may be difficult to motivate to complete a survey.

In the case of the poor, illiteracy may also be a constraint in some cases in terms of being able to complete any sort of writing-based survey.

Failure to survey the poor and/or the rich will also introduce spatial biases/gaps in the data, given the spatial segregation of these groups with inevitably exists. Modal biases will also exist, given that the poor are very transit-dependent and typically lack access to automobiles, while the rich are very auto-oriented and are likely to be resistant to using transit (unless, perhaps, it is a premium service).

Given these (arguably growing) challenges with person/household-based surveys, technology-based data collection methods (cellular data, smartphone apps, smartcards, other passive data sources), with their potential to provide very large amounts of data concerning trip-making, possibly over extended periods of time appear to be a very attractive approach for designing future data collection efforts. But, as discussed in Chapter 5, these come with their own set of challenges as well. With respect to the Latin American context, two in particular need to be highlighted.

First, these data generally are anonymously collected, and so the socio-economics of the trip-makers are unknown. As discussed in Chapter 6, data imputation and fusion methods can be used to enrich the trip data with imputed/synthesized attributes of trip-makers and/or their origin and destination attributes. Home, work and school locations can generally be imputed if multiple-day/week observations of a given (anonymous but "labelled") trip-maker are available. In such cases, at a minimum, given some knowledge of the socio-economic nature of different residential and employment neighbourhoods, some allocation of broad socio-economic status

[33] This comment can extend to the middle class in cities such as Sao Paulo, etc. experiencing major security issues.

might be attached to each trip-maker. Ideally, however, one would like to have recent, spatially detailed census data to provide the opportunity to synthesis a richer socio-economic representation of each trip-maker. The nature of census data, however, appears to vary considerably from one Latin American country to another in terms of: attributes collected, spatial precision, how frequently the census is conducted, and availability of detailed data for planning and modelling purposes. Thus, the opportunity to fuse anonymously-collected trip data with census data may vary considerably from city to city and country to country.[34]

Second, by definition, technology-based data collection methods require the trip-makers being tracked to be using the technology. Access to technology may vary considerably by socio-economic group. Cellphones now seem to be near-ubiquitous in many Latin American cities (even among poorer people), but smartphones still may display biases in terms of market penetration across socio-economic groups. Similarly, the socio-economic distribution of smartcard usage may be unknown in many cases.[35]

7.2.3 Informal/Privatized Transit Services

The delivery of transit services varies considerably across Latin American cities. In particular, in some cities, transit is largely or entirely privatized, with numerous private operators often competing for customers and with varying government oversight. In such cases data collection and analysis of transit ridership may be challenging for a number of reasons, including:

- Transit choice-based (onboard/rider) surveys may be difficult to implement due to the difficult in coordinating the survey across operators and/or in incenting the operators to cooperate.
- Obtaining consistent ride counts by transit line and operator may be difficult.
- Getting consistent, accurate information concerning transit line configurations, stop locations, fares, etc. by transit operator may be difficult.
- Smartcard implementations (and associated transit usage data) may not be feasible in the case of multiple and competing operators.

Given both the importance of transit in Latin American cities (notably among lower income classes) and the challenges with conventional travel survey methods discussed above, direct observations of transit riders through some combination of ridership counts, onboard surveys and smartcard records might represent an important component of an overall data collection program. As noted above, however, the feasibility and efficiency of such methods will generally depend in a critical way on the ability of municipal governments to implement and coordinate such data collection efforts, which, in turn, will depend on the organizational structure of transit within the given municipality.

[34] As also discussed in Chapter 5, these types of data are almost invariably collected by individual, rather than by household. Household contexts, however, are extremely useful in helping to explain individuals' travel behaviour. Data fusion is essential if household-based analyses and models are to be maintained, again reinforcing the need for high quality, detailed additional socio-economic data, such as a national census, to provide the data needed to support such fusion.

[35] Web-based surveys are subject to the same concern: a person needs internet access via some sort of device (computer, tablet, smartphone) in order to participate in the survey.

7.2.4 Government Structures & Capacities

More generally, any successful transportation data collection program requires a strong, sustained commitment by government to the collection and maintenance of the data. Data collection is often viewed as an expensive exercise that may be difficult to justify within tight municipal (or even national) budgets. The "business case" for high quality data to support cost-effective, impactful decisions concerning high-cost infrastructure investment, road and transit pricing policies, improving active transportation usage, and, in general, improving economic development, social equity and quality of life is very strong (Miller, et al., 2012). To achieve an effective, on-going data collection program requires recognition on the part of government of the importance of the data collection program. This must occur both at the technical staff level (who must champion the program) and the senior management level (who must find funding for the program and support technical staff in their activities).

It also requires multi-agency and multi-government coordination and cooperation. As with most large cities worldwide, Latin American urban regions usually consist of multiple municipalities and multiple agencies concerned with transportation planning, policies and operations. Travel and transportation services, however, traverse municipal boundaries and require a coordinated approach to all aspects of planning and decision-making, including data collection. Data collection by one municipality or agency within a large urban region, while perhaps better than no data collection, will have only limited impact on regional transportation planning and decision-making. Independent data collection by multiple municipalities or agencies within a region will generate inconsistencies and gaps in the overall data and will not be cost-effective, with the "whole being less than the sum of the parts".

Fragmented, uncoordinated data collection efforts will also result in general in smaller (and hence less useful) survey sample sizes, since the "overhead" in survey design, execution and management is spent over and over again with each survey, rather than once for a collaborative effort. It also is a barrier advancing the local state of practice with respect to data collection, since "lessons learned" in one survey or municipality/agency may be difficult to transmit to other groups within the region: the tendency is for the "same wheels being reinvented" over and over again, rather than the state of practice being incrementally advanced with each new survey or other data collection exercise.

The technical capacity of individual municipalities and agencies to support extensive, high-quality data collection efforts is often limited, and certainly varies considerably from one agency to another. Again, fragmented, uncoordinated approaches to data collection within an urban region will generally lead to a lack of a "critical mass" within any one agency to undertake major data collection efforts. Collaboration across agencies, on the other hand, can coordinate the use of all resources available across the agencies to create such a critical mass and to maximize the effectiveness of available staff to support data collection efforts.

An additional level of institutional complexity exists in terms of the strong role that national governments typically play in Latin American urban affairs. A supportive national government can play a very constructive supporting role in promoting an urban region's data collection efforts. Conversely, a national government which is politically at odds with a municipal administration can be equally obstructive. Ideally data collection should not be a political or ideological issue: it is in the public good regardless of the party in power. Establishing this attitude in many situations, of course, may not be easy to do. Elements of successfully doing this include:

- Making a strong, non-partisan "business case" for an on-going, comprehensive data collection process.
- Establishing a lead agency or (usually better) coordinating committee (with joint membership by all participating agencies) to manage the program.
- Keeping data collection as much as possible a technical exercise, managed by technical agencies, rather than a politicized one (i.e., one that is closely linked to a particular politician's or party's platform).
- Developing a multi-year program (with funding!) that is not tied to either municipal or national election cycles.
- Establishing some "quick wins" that demonstrate the usefulness and cost-effectiveness of the data collection program.
- Involving one or more local universities in the data collection and/or management process can promote a clear sense of objectivity and non-partisanship to the data and the data collection process. This can also promote government-university collaboration in analysis, modelling and the training of professional planners and analysts to build government analytical and planning capabilities over time.

A very successful example of this approach (albeit from Canada) is the Toronto region. The Transportation Information Steering Committee (TISC) was founded in the 1980's to coordinate data collection within the region. All major municipalities and transit agencies are members of TISC, with the Province of Ontario Ministry of Transportation (MTO) chairing the committee.[36] TISC members fund a collective data collection effort, notably a major travel survey every five years: the Transportation Tomorrow Survey (TTS). TTS and other shared transportation datasets are maintained by the Data Management Group (DMG) at the University of Toronto, which is funded by TISC. Funding for TTS and DMG is shared between the Province and the municipalities, with the municipal contributions being allocated in proportion to their populations. This program has established its credentials as a credible, objective, cost-effective source of valuable information, and is never subject to partisan political scrutiny, despite many changes of government at both the provincial and municipal levels. Municipalities and the Province may argue over projects, policies and the allocation of transportation capital and operating funding, but they never argue over the data being used to support these debates.

CAF's OMU could play an important role in addressing many of these institutional issues in terms of providing:

- An objective, non-artisan forum for collaboration and the sharing of data and expertise.
- Long-term technical and (possibly) financial support for data collection that is not tied to political imperatives and election cycles.
- Development of standard, state-of-the-art data collection methods and assistance to urban regions in their implementation.

[36] Canada is a federal state with generally three levels of government: municipal, provincial and the national (federal) government. In Canada, provinces play a role similar to most Latin American national governments vis-à-vis urban transportation, since the federal government has little constitutional authority in this area.

- Training of technical staff in state-of-the-art data collection methods.

7.3 Next Steps

This report has identified, in general terms, the strengths and weaknesses of the current transportation data collection state of art/practice. The on-going Montevideo iCity-South project is investigating several of these methods in detail within the Latin American context, using Montevideo as its case study region.

Specifically, at the time of this report's preparation the project is investigating:

- The recent Montevideo Household Mobility Survey (MHMS) as an example of a conventional household travel survey.
- Montevideo smartcard and other transit ridership data.
- A sample of Antel cellular data records (CDR) for Montevideo.

The results of these analyses will be published in a series of project reports as the results are finalized.

In parallel to the Montevideo project, the Toronto-based TTS 2.0 project is very actively investigating:

- Web-based surveys (including the development of new, advanced software for constructing web-based travel surveys).
- Smartphone apps for collecting trip records.
- Other passive data sources.
- Possible continuous survey designs.
- Data fusion techniques.
- Elaboration of a core-satellite data collection program design for the Toronto region.

The extent to which some of this research might be translated to the Montevideo case remains to be seen, especially within the current rather limited project budget and timeframe.

The final product of this project will be a recommended R&D program for developing a comprehensive data collection program for Montevideo. This program will build on the results of both this project and the TTS 2.0 project.

REFERENCES

Abt SRBI (2011) Household Travel Survey with GPS data loggers*,*

Agard, B., C. Morency and M. Trépanier (2006) "Mining public transport user behaviour from smart card data", in: *12th IFAC Symposium on Information Control Problems in Manufacturing* – INCOM 2006, Saint-Etienne, France, May 17–19.

Amey, A., L. Liu, F. Pereira, C. Zegras, M. Veloso, C. Bento, and A. Biderman (2009) "State of the Practice Overview of Transportation Data Fusion: Technical and Institutional Considerations", Paper# ITS-CM-09-01, Working Paper Series Working Paper Series, MIT-Portugal Program

Ashley, D., A. Richardson and D. Young (2009). Recent Information in the Under-Reporting of Trips in Household Travel Surveys, Australasian Transport Research Forum. http://www.tuti.com.au/atrf09-ashley.pdf

Bachmann, C., Roorda, M.J., Abdulhai, B. and B. Moshiri (2013) "Fusing a Bluetooth Traffic Monitoring System with Loop Detector Data for Improved Freeway Traffic Speed Estimation", *Journal of Intelligent Transportation Systems: Technology, Planning, and Operation* , 17|2, pp 152-164.

Bagchi, M., White, P.R., 2004. What role for smart-card data from bus system? Municipal Engineer 157, 39–46.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transport Policy 12, 464–474.

Baltes M. A. (2002). *Costumer Surveying for Public Transit: A Design Manual for Customer On-Board Surveys*, final report for National Center for Transit Research NCTR, NCTR - 416 – 083

Baltes M. A. (2003). *Conducting a Successful On-Board Survey of Public Transit Customers, Proceedings of the 2003 Mid-Continent Transportation Research Symposium*, Ames, Iowa, August 2003.

Bayart, C., Bonnel, P., Morency, C., 2008. Survey mode integration and data fusion: Methods and challenges. Resource paper for Workshop on Best Practices in Data Fusion, 8th International Conference on Survey Methods in Transport Annecy, France, May 25-31, 2008

Beirness D.J. and E.E. Beasley (2010). A Roadside Survey of Alcohol and Drug Use among Drivers in British Columbia. *Traffic Injury Prevention* 11: 215–221

Benfield, J. A., & Szlemko, W. J. (2006). Internet-based data collection: Promises and realities. Journal of Research Practice, 2(2), Article D1. Retrieved [December 1st 2011] from, http://jrp.icaap.org/index.php/jrp/article/view/30/51

Bernard, J.-T.; Bolduc, D.; Yameogo, N.-D. (2012). A Pseudo-Panel Data Model of Household Electricity Demand, forthcoming, *Resource and Energy Economics*

Blash L., J.D. Rogers and R. LeGates (2002). Urban Public Transit Riders: Surveying a Population on the Move, Technical report, American Association of Public Opinion Research.

Bohte W. and K. Maat (2009). Deriving and Validating Trip Purposes and Travel Modes for Multi-Day GPS-Based Travel Surveys: A Large-Scale Application in the Netherlands, *Transportation Research Part C: Emerging Technologies,* 17(3): 285-297

Bourbonnais, Pierre-Leo, Morency, Catherine (2011). Web-based travel survey: a demo, presented at the 9th International Conference on Transport Survey Methods, Chile, Nov. 2011.

Cambridge Systematics (1996). *Travel Survey Manual*, Washington, DC: US Department of Transportation, Federal Transit Administration, Federal Highway Administration, the Office of the Secretary and the U.S. Environmental Protection Agency.

Chalasani, V.S. and K.W. Axhausen (2004). *Mobidrive: A Six Week Travel Diary*, Zurich: Institute for Transport Planning and Systems, Swiss federal Institute of Technology Zurich. http://www.ivt.ethz.ch/vpl/publications/tsms/tsms2.pdf
http://www.fhwa.dot.gov/ohim/trb/mobidrive.pdf

Chalasani, V.S. and K.W. Axhausen (2005). Conceptual Data Model for the Integrated Travel Survey and Spatial Data, in: *Proceedings of ASC 2005 Maximising Data Value*, R. Khan, R. Banks, R. Cornelius, S. Evans and T. Manners (ed.), ASC, Chesham, pp. 123-135. (http://www.ivt.ethz.ch/vpl/publications/reports/ab302.pdf, accessed November 2011)

Charlton B., J Hood, E. Sall, M.A. Schwartz (2011) Bicycle Route Choice Data Collection Using GPS-Enabled Smartphones, *Transportation Research Board 90th Annual Meeting*, Papers DVD

Chu, X. (2006) *Using Local Transit On-Board Surveys for State-Level Measurements*, Report NCTR-576-07, BD549-16, National Center for Transit Research (NCTR) , University of South Florida

Chu, A., Chapleau, R. (2011). Smart Card Validation Data as a Multi-Day Transit Panel Survey to Investigate Individual and Aggregate Variation in Travel Behaviour, Presented at the 9th International Conference on Transport Survey Methods, Chile.

Chung, E., & Shalaby, A. S. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, 28(No. 5), 381-401.

Chung, B., S. Srikukenthiran, K.N. Habib and E.J. Miller (2017) “Development of a Web Survey Builder Platform for Household Travel Surveys”, presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

City of Calgary (2002) *2001 Household Activity Survey*, Calgary: Planning and Transportation Policy Forecasting Division. http://www.calgary.ca/Transportation/TP/Documents/forecasting/household_activity_survey_report.pdf

Cochran, W.G. (1977). *Sampling Techniques, 3rd Edition*, New York: John Wiley & Sons.

Colak, S., A. Lima, and M.C. Gonzalez (2016) "Understanding congested travel in urban areas", *Nature Communications* Vol. 2, 10793 (2016).

D'Ambrosio, A., Aria, M., Siciliano, R. (2007). Robust Tree-Based Data Imputation Method for Data Fusion. M.R. Berthold, J. Shawe-Taylor, and N. Lavrac (Eds.): IDA 2007, LNCS 4723, Berlin: Springer-Verlag, 174–183.

de Fabritiis C., R. Ragona and G. Valenti (2008). Traffic Estimation And Prediction Based On Real Time Floating Car Data, *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*: 197-203.

Descoimps, E., Agard, B., Trépanier, M. (2011). Comment les conditions climatiques influencent-elles l'utilisation du transport collectif ? Normalité des déplacements et impacts météorologiques, Dixièmes rencontres francophones Est-Ouest de socio-économie des transports, Montréal, Québec, Canada, 2-3 juin.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*, New York: John Wiley & Sons.

Dillman, D,A. (2009). *Mail, Internet and Mixed-Mode Surveys: The Tailored Design Method*, New York: John Wiley & Sons.

DMG (1991). *Analysis of TTS Data Bias due to Use of Informants*, Toronto: Data Management Group, University of Toronto.

DMG (1993*). Under-reporting of Trips in Telephone Interview Travel Surveys*, Toronto: Data Management Group, University of Toronto.

DMG (2007) *2006 Transportation Tomorrow Survey: Design & Conduct of the Survey*, Toronto: Data Management Group, University of Toronto.

Doherty, S.T. and E.J. Miller (2000) "Tracing the Household Activity Scheduling Process Using One-Week Computer-Based Survey", *Transportation*, Vol. 27, pp. 75-97.

Doherty, S.T., E. Nemeth, M.J. Roorda and E.J. Miller (2004). Design and Assessment of the Toronto Area Computerized Household Activity Scheduling Survey, *Transportation Research Records, Journal of the Transportation Research Board*, No. 1894, 140-149.

D'Orazio, M., M. Di Zio and M. Scanu (2006). *Statistical Matching: Theory and Practice*, New York: John Wiley and Sons.

El-Assi, W., K.M.N. Habib, C. Morency, and E.J. Miller (2017) "Investigating the Capacity of Continuous Household Travel Survey in Capturing the Temporal Rhythm of Travel Demands", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Friedrich, M., Jehlicka, P., Schlaich, J. (2008): Automatic number plate recognition for the observance of travel behavior, 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability, Mai 2008, Annecy, Frankreich.

Frignani, M., J. Auld, A. Mohammadian, C. Williams and P. Nelson (2010). Urban travel route and activity choice survey (UTRACS): An internet-based prompted recall activity travel survey using GPS data. *Transportation Research Record 2183*. pp. 19-28.

Gautier, J-M. (1999). Mégabase de consommateurs, sondages et statistique. In: Enquêtes et Gilula, Z., McCulloch, R.E., Rossi, P.E. (2006). A Direct Approach to Data Fusion. *Journal of Marketing Research*, Vol XLIII.

Geostats (2011). GeoStats (2011 Household Travel Surveys for Jerusalem, New York/New Jersey and Cleveland USA, http://www.geostats.com/service_travel.htm (page consulted Dec. 19, 2011).

Gile, K. J. and M. S. Handcock (2010). Respondent-driven sampling: an assessment of current methodology. Sociological Methodology 40(1): 285-327.

Golob, T., R. Kitamura and L. Long, (eds.) (1997). *Panels for Transportation Planning: Methods and Applications*, Dordrecht, the Netherlands: Kluwer Academic Publishers.

Gonzalez, M.C., Hidalgo, C.A., Barabasi, A-L (2008). Understanding individual human mobility patterns, Nature, 453:5, doi:10.1038/nature06958.

Goulias, K., R. Pendyala and C. Bhat (2011). Total Design Data Needs for the New Generation Large Scale Activity Microsimulation Models, forthcoming, *Proceedings of the 9th International Conference on Survey Methods in Transport, Scoping the Future While Staying on Track*, Termas de Puyehue, Chile.

Greaves S. And R. Ellison (2011) A GPS/Web-based Solution for Multi-day Travel Surveys: Processing Requirements and Participant Reaction, 9th International Conference on Survey Methods in Transport, Scoping the Future While Staying on Track, Termas de Puyehue, Chile

Greaves, S., S.S. Fifer, R. Ellison, G. Germanos (2010). Development of a global positioning system web-based prompted recall solution for longitudinal travel surveys.. *Transportation Research Record, 2183*. pp. 69-77.

Grond, K. and E.J. Miller (2016) "Analysis of GPS Smartphone Data for a Toronto Cycling Route Choice Model", presented at the 5th Symposium of the European Association for Research in Transportation, Delft, September 14.

Guy, B.P. and J.D. Fricker (2005*). Guidelines for Data Collection Techniques and Methods for Roadside Station Origin-Destination Studies*, Publication FHWA/IN/JTRP-2005/27. Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana, 2005. doi: 10.5703/1288284313368

Habib, K.M.N., L. Kattan and M.T. Islam (2009). "Model of Personal Attitude towards Transit Service Quality", forthcoming in *Journal of Advanced Transportation*: 10.1002/atr.106

Habib, K.N., E.J. Miller, S. Srikukenthiran, M. Lee-Gosselin, C. Morency, MJ. Roorda and A. Shalaby (2017) "TTS 2.0: A Research and Development (R&D) Project on Passenger Travel Survey Methods", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Harding, C., S. Srikukenthiran, K.N. Habib and E.J. Miller (2017a) "Evaluation of cost effectiveness and feasibility of in-person surveys as an augment to the regional travel survey: a case study in the Toronto area", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Harding, C., Z.T. Zhang, S. Srikukenthiran, K.N. Habib and E.J. Miller (2017b) "On the User Experience and Performance of Smartphone Apps as Personalized Travel Survey Instruments: Results from an Experiment in Toronto", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Harding, C., Z.T. Zhang, S. Srikukenthiran, K.N. Habib and E.J. Miller (2017c) "An Experimental Study for Evaluating Smartphone Apps as Personalized Passenger Travel Survey Tools in Toronto", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Haroun, A. and E.J. Miller (2004) Retrospective Surveys in Support of Dynamic Model-Building", presented at the 7th International Conference on Travel Survey Methods, Costa Rica, August.

Hartgen D. T. (1992). Coming in the 1990s: The Agency-Friendly Travel Survey. *Transportation* 19: 79-95

Hatzopoulou, M. and E.J. Miller (2010). Linking an Activity-Based Travel Demand Model with Traffic Emission and Dispersion Models: Transport's Contribution to Air Pollution in Toronto, *Transportation Research D*, 15:6, 315-325.

Hollingworth, B. and E.J. Miller (1996) Retrospective Interviewing and Its Application in the Study of Residential Mobility *Transportation Research Record 1551*, pp. 74-81.

Hoddinott, S.N. and M.J. Bass (1986). The Dillman Total Design Survey Method: A Sure-Fire Way to Get High Survey Return Rates, *Canadian Family Physician*, 32, Nov. 2366-2368.

Hui, N., M.J. Roorda, C. Davies and E.J. Miller (2017) "Using video data to evaluate pedestrian, bicycle and vehicle conflicts", presented at the Joint ITE/CITE 2017 Annual Meeting, Toronto, July.

Inbakaran, C. and A. Kroen (2011). Travel Surveys – Review of international survey methods, Australasian Transport Research Forum 2011 Proceedings. http://www.atrf11.unisa.edu.au/Assets/Papers/ATRF11_0106_final.pdf

Isaacman, S., R. Becker, R. C´aceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky (2011) "Identifying important places in peoples lives from cellular network data", in *Pervasive Computing*, pages 133– 151, Springer.

Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. Gonzalez (2013) "A review of urban computing for mobile phone traces: current methods, challenges and opportunities:, in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 2. ACM.

Jiang S., J. Ferreira, Jr. J. and M.C. Gonzalez, (2016a) "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore", paper presented at ACM KDD UrbComp'15, published in *IEEE Transactions in Big Data* Issue 9.

Jiang S.,Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M.C. Gonzalez (2016b) "TimeGeo: a spatiotemporal framework for modeling urban mobility without surveys", *PNAS* 113(37), E5370-E5378.

Kelly P., A. Doherty, E. Berry, S. Hodges, A. Betterham and C. Foster (2011) Can we use digital life-log images to investigate active and sedentary travel behaviour? Results from a pilot study, International *Journal of Behavioural Nutrition and Physical Activity*, 2011, 8:44

Kestens, Y. (2011). Panelist , Obesity, Diet, and Physical Activity session, 2011 mHealth Summit,Washington, D.C. (http://www.mhealthsummit.org/program_speakers_ykestens.php).

Kurth, D.L., J.L. Coil and M.J. Brown (2001). Assessment of Quick-Refusal and No-Contact Nonresponse in Household Travel Surveys, *Transportation Research Record: Journal of the Transportation Research Board*, 1768, 114-124.

Lee-Gosselin, M.E.H (2005): "A Data Collection Strategy for Perceived and Observed Flexibility in the Spatio-Temporal Organisation of Household Activities and Associated Travel", in Timmermans, H.J.P. *Progress in Activity- Based Analysis*. Elsevier, pp. 355-371

Lee-Gosselin, M.E.H., Doherty, S.T. & Shalaby, A. (2010). Data collection on personal movement using mobile ICTs : old wine in new bottles?, in: Wachowitz, M. (Ed) : *Movement-*

Aware Applications for Sustainable Mobility, Information Science Reference, IGI Global, pp 1-15

Lestina D., M. Greene, R.B. Voas and J. Wells (1999). Sampling Procedure and Survey Methodologies for the 1996 Survey with Comparisons to Earlier National Roadside Survey, *Evaluation Review* 23: 28-46

Loa, P., S. Srikukenthiran, K.M.N. Habib and E.J. Miller (2015) *Current State of Web-Based Survey Methods*, Transportation Tomorrow Survey 2.0 research report, Toronto: University of Toronto Transportation Research Institute, October.

Lue, G. and E.J. Miller 2018) "Estimating a Toronto Pedestrian Route Choice Model using Smartphone GPS Data", submitted for presentation at the 97th Annual Meeting of the Transportation Research Board, Washington, DC, January.

Malinovskiy, Y, Saunier, N, Wang. Y. (2012). Pedestrian travel analysis using static bluetooth sensors. In Transportation Research Board Annual Meeting Compendium of Papers, 2012. 12-3270 Considered for publication in Transportation Research Record: Journal of the Transportation Research Board

Manski, C.F. and S.R. Lerman (1977). The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica* 45 (8): 1977-1988.

Market Research World (2011). http://www.marketresearchworld.net, accessed December 25, 2011.

Meyer, M.D. and E.J. Miller (2001). *Urban Transportation Planning: A Decision-Oriented Approach, 2nd Edition*, New York: McGraw-Hill.

Miller, E.J., F.F. Calderón Peralvo, J. Vaughan and B. Yusuf (2017a) *SATA: Simulador de Actividad de Transporte de Asunción, Development of the SATA Prototype Volume I: Final Report*, report to CAF, Toronto: University of Toronto Transportation Research Institute.

Miller, E.J., F.F. Calderón Peralvo, J. Vaughan, B. Yusuf and A. Faghig-Imani (2017b) *SATA: Simulador de Actividad de Transporte de Asunción, Development of the SATA Prototype Volume II: Technical Appendices*, report to CAF, Toronto: University of Toronto Transportation Research Institute.

Miller, E.J. and D.F. Crowley (1989) "Panel Survey Approach to Measuring Transit Route Service Elasticity of Demand", Transportation Research Record 1209, 1989, pp. 26-31.

Miller, E.J., K.N. Habib, M. Lee-Gosselin, C. Morency, M.J. Roorda and A.S. Shalaby (2012) *Changing Practices in Data Collection on the Movement of People, Final Report*, report to the Transportation Association of Canada, Île d'Orléans, Québec: Lee-Gosselin Associates Ltd.

Miller, E.J., C. Harding, M. Nasterska and Y Zhang, *Waterfront Toronto Transportation Carbon Model System Update*, final project report to Waterfront Toronto, Toronto: University of Toronto Transportation Research Institute, February, 2016, 40 pages.

Mogeng, Y., M. Sheehan, S. Feygin, J-F Paiement and A. Pozdnoukhov (2016) "A Generative Model of Urban Activities from Cellular Data, *IEEE TRANSACTIONS IN ITS*.

Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transport Policy 14 (3), 193–203.

MTO (2009) *City of Toronto and Regions of Durham, Peel and York 2008 Travel Time Study*. Final Report prepared by IBI Group. August.

Munizaga, M., Palma, C., Mora, P., 2010. Public transport OD matrix estimation from smart card payment system data. In: Presented at the 12th World Conference on Transport Research, Lisbon, Paper No. 2988

NCHRP (2007) *Technical Appendix to NCHRP Report 571: Standardized Procedures for Personal Travel Surveys*, Contractor's Final Report for NCHRP Project 8-37, Web-Only Document 93, Washington, DC: National Cooperative Highway Research Program, revised December, 2007. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w93.pdf

NCHRP (2008) *Standardized Procedures for Personal Travel Surveys, NCHRP Report 571*, Washington, DC: National Cooperative Highway Research Program. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_571.pdf

Noureldin, A., Karamat, T., Eberts, M. and El-Shafie, A. (2009). Performance Enhancement of MEMS Based INS/GPS Integration for Low Cost Navigation Applications, IEEE Transactions on Vehicular Technology, V58 (3), pp: 1077 – 1096, March 2009

Park, J.Y., Kim, D.J., 2008. The Potential of Using the Smart Card Data to Define the Use of Public Transit in Seoul. Transportation Research Record: Journal of the Transportation Research Board, No. 2063, Transportation Research Board of the National Academies, Washington, DC, pp. 3–9

Pelletier, M.P, Trépanier, M., Morency, C. (2011). Smart card data use in public transit: A literature review, Transportation Research Part C 19 (2011) 557–568.

Pendyala, R.M., K.G. Goulias, R. Kitamura and E. Murakami (1993). Development of Weights for a Choice-Based Panel Survey Sample with Attrition, *Transportation Research Part A* 27A: 477-492

Polak, J. (2006). OPUS: Optimising the Use of Partial Information in Urban and Regional Systems, Presentation, and Research Methods Festival Programme. http://www.ccsr.ac.uk/methods/festival/programme/opus1/polak.ppt (accessed November 2011)

Pritchard, D. and E.J. Miller (2012) "Advances in Population Synthesis: Fitting Many Attributes Per Agent and Fitting to Household and Person Margins Simultaneously", *Transportation* 39(3), pp. 685-704.

Rashed, M., S. Srikukenthiran, K.M.N. Habib and E.J. Miller, *Current State of SmartPhone Survey Methods*, Transportation Tomorrow Survey 2.0 research report, Toronto: University of Toronto Transportation Research Institute, October, 2015, 47 pages.

Reddy, A., Lu, A., Kumar, S., Bashmakov, V., Rudenko, S., 2009. Application of Entry-Only Automated Fare Collection (AFC) System Data to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles. 88th Annual Meeting of the Transportation Research Board, Washington, 21 p. (CD-ROM)

Roorda, M. J., M. Lee-Gosselin, S. T. Doherty, E.J. Miller and P. Rondier (2005). Travel/Activity Panel Surveys in the Toronto and Quebec City Regions: Comparison of Methods and Preliminary Results. CD Proceedings of the *PROCESSUS Second International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications*. Toronto, June.

Roorda, M.J., and E.J. Miller (2004). Toronto Activity Panel Survey: Demonstrating the Benefits of a Multiple Instrument Panel Survey. CD Proceedings of the *Seventh International Conference on Travel Survey Methods*. Costa Rica, August.

Roorda, M., B. Sharman, C. Sekula and P. Masters (2009). Preliminary Analysis of a System for Real-time Monitoring of Bluetooth Device Data on an Urban Freeway. Paper presented at the Translog 2009 conference. Hamilton, June 17-18.

Saneinejad, S., M. J. Roorda and C. Kennedy (2011). Modelling the Impact of Weather Conditions on Active Transportation Travel Behaviour. *Transportation Research Part D: Transport and Environment*, 17(2) 129-137.

Saunier, N., Morency, C. (2011). Comparing data from mobile and static traffic sensors for travel time assessment, ASCE Conf. Proc. T&DI Congress 2011: Integrated Transportation and Development for a Better Tomorrow Proceedings of the First T&DI Congress 2011

Saunier, N., Sayed, T. (2006). A feature-based tracking algorithm for vehicles in intersections, Proceeding CRV '06 Proceedings of the 3rd Canadian Conference on Computer and Robot Vision IEEE Computer Society Washington, DC.

Schüssler, N. and K. W. Axhausen (2008) Identifying trips and activities and their characteristics from raw GPS data without further information. *8th International Conference on Survey Methods in Transport.* Annecy France.

Sharman, B. and M.J. Roorda (2011). Analysis of Freight GPS Data: A Clustering Approach for Identifying Trip Destinations. *Transportation Research Record No. 2246*. 83-91.

Sharman, B., Roorda, M.J. and K.M.N. Habib (2014). Comparison of Parametric and Non-Parametric Hazard Models of Stop Durations on Urban Commercial Vehicle Tours. Paper accepted for publication in *Transportation Research Record*, No. 2269.

Sharp J. and E. Murakami (2005). Travel Surveys: Methodological and Technology-Related Considerations, *Journal of Transportation and Statistics*, 8(3) 97-113.

Seskin I.M. and P.R. Stopher (1998). Spatial Variations in Attitude Towards Expanded Public Transit Service, *Transportation* 15: 211-232

Singer, E. *The Use of Incentives to Reduce Nonresponse in Household Surveys* (2002). Report no. 051, Ann Arbor: Survey Methodology Program, The University of Michigan Institute for Social Research Survey Research Centre.
http://www.isr.umich.edu/src/smp/Electronic%20Copies/51-Draft106.pdf

Srikukenthiran, S., K.N. Habib and E.J. Miller (2017a) "Impact of a Multiple Survey Frames on Data Quality of Household Travel Surveys: The case of the 2016 Transportation Tomorrow Survey", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Srikukenthiran, S., K.N. Habib and E.J. Miller (2017b) "Impact of a Multiple Survey Frames on Data Quality of Household Travel Surveys: The case of the 2016 Transportation Tomorrow Survey", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Srikukenthiran, S., K.N. Habib, T. Lin and E.J. Miller (2017c) "Inverted Sampling Frames to overcome Under-Coverage of specific Population Cohorts: Examining the viability of recruiting households via employers and institutions", presented at the 11th International Conference on Transport Survey Methods, Quebec, September 24-29.

Stopher P. (1992) Development of Route Level Patronage Forecasting Method. *Transportation* 19: 201-220

Stopher P. R., C Prasad, L. Wargelin and J Minser (2011) Conducting a GPS-Only Household Travel Survey, 9th International Conference on Survey Methods in Transport, Scoping the Future While Staying on Track, Termas de Puyehue, Chile.

TCRP Synthesis 63 (2005). *Conducting On-Board and Intercept Transit Survey Techniques*, final project report, Washington, DC: US Federal Transit Administration, Project J-7, Topic SH-05.

Tooley, M.S. (1996). Incentives and Rates of Return for Travel Surveys, *Transportation Research Record: Journal of the Transportation Research Board*, 67-73.

Transportation Association of Canada (TAC) (2008). *Best Practices for the Technical Delivery of Long-Term Planning Studies in Canada Final Report.* Project Report ISBN 978-1-55187-261-7
Transport Canada (2009). *Canadian Guidelines for the Measurement of Transportation Demand Management Initiatives - User's Guide*.

TRB Travel Survey Methods Committee -ABJ40 (2011). *The On-Line Travel Survey Manual*, http://www.travelsurveymanual.org/, accessed December 26, 2011.

Trépanier, M., Chapleau, R., Tranchant, N., 2007. Individual trip destination estimation in transit smart card automated fare collection system. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 11 (1), 1–15 (Taylor & Francis).

Trépanier, M., Morency, C., 2010. Assessing transit loyalty with smart card data. Presented at the 12th World Conference on Transport Research, Lisbon, Paper No. 2341.

Trépanier, M., Morency, C., Blanchette, C., 2009a. Enhancing Household Travel Surveys Using Smart Card Data? 88th Annual Meeting of the Transportation Research Board, Washington, 15 p. (CD-ROM).

Trépanier, M., Morency, C., Agard, B., 2009b. Calculation of transit performance measures using smartcard data. Journal of Public Transportation 12 (1), 79–96.

Trépanier, M., Vassivière, F., 2008. Democratized smart card data for transit operators. In: 15th World Congress on Intelligent Transport Systems, New York, USA, 12 p

Wissen, L.J.G and H.J. Meurs (1989). The Dutch Mobility Panel: Experiences and Evaluation, *Transportation* 16:2, 99-119.

Wolf, J. J. Wilhelm, J. Casas and S. Sen (2011) A case study: Multiple data collection methods and the NY/NJ/CT regional travel survey. *9th International Conference on Survey Methods in Transport.* Termas de Puyehue, Chile.

Yang, J., Lu, H., Liu, Z., Boda, P.P.(2010). Physical Activity Recognition with Mobile Phones: Challenges, Methods and Applications. In: L. Shao et al. (Eds.): Multimedia Interaction and Intelligent User Interfaces: Principles, Methods and Applications, Springer-Verlag, London, pp. 185-213 (2010).

Zimowski, M., R. Tourangeau, R. Ghadialy and S. Pedlow (1997). *Nonresponse in Household Travel Surveys*, Washington, DC: US Federal Highway Administration. http://tmip.fhwa.dot.gov/resources/clearinghouse/docs/surveys/nonresponse/non.pdf

Zmud, J. (2007). Washington Full Survey Design Documentation, Unpublished manuscript, Austin, Texas: NuStats.

Zmud, J., B. Chlond, T. Kuhnimhof and A. Richardson (2011). *Feasibility Study of a Continuous Household Activity Survey Program (CHASP),* prepared for PTV NuStats and City of Calgary.