Transportation Association of Canada

Primer on
Urban Passenger Data Collection: Keeping Up With a Changing World

It is increasingly recognised that the *status quo* in urban passenger transportation data collection and management is not sustainable: practice must evolve to meet changing needs, opportunities and challenges. We are in a highly dynamic era in which technology is changing rapidly, and picking "winners/losers" among data collection methods is difficult to do. Furthermore, appropriate methods vary by context, and this context is also changing as urban areas grow and evolve. Thus, it is essential that transportation agencies be willing to experiment with new methods, guided by a blueprint for a process of adaptation and evolution in a changing world.

The scope of urban passenger data collection today

As illustrated in Figure 1, high quality, comprehensive data on travel behaviour, transport network performance and associated land use characteristics are absolutely essential to the planning, design and operations of urban transport systems. Such information is derived from a variety of sources, but most comes from travel surveys and other data collection methods that provide the base data for analysis and modelling of both trip-making and transport system performance. Without adequate data we can neither understand our transport needs and issues nor can we design and evaluate service and policy alternatives.

Given this fundamental need for data, transport agencies at all levels of government need to invest considerable resources in on-going data collection, using many methods. Changes in Information Technology (IT), however, are rapidly introducing new static and mobile data collection options (Web surveys, GPS tracking, smart-phone applications, etc.), while at the same time reducing the effectiveness of some methods (e.g., the decline in effectiveness of land-line-based telephone interviews).

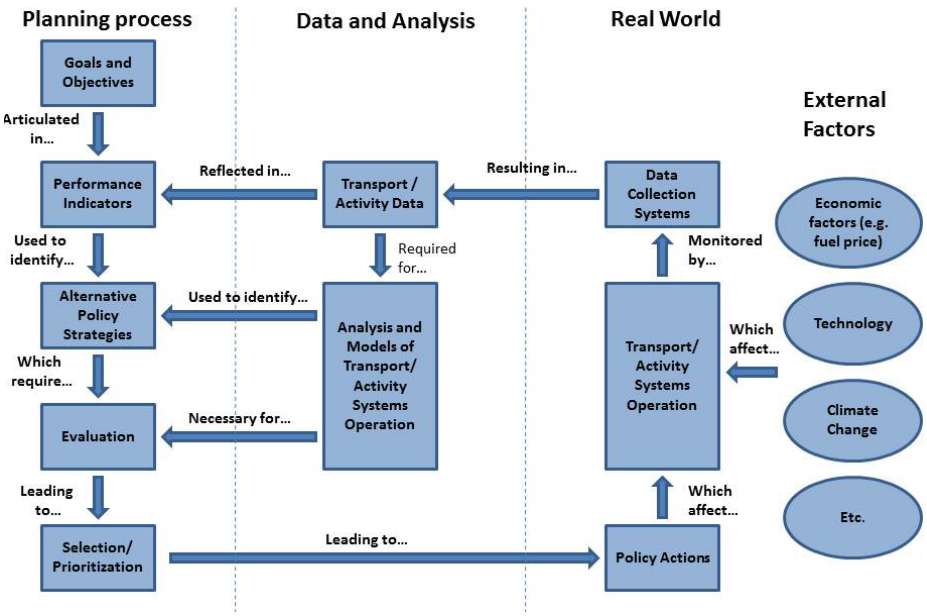


Figure 1 – Role of Data in the Transportation Planning Process

Similarly, advances in statistical methods for fusing and mining diverse data sets offer the potential to make much better use of a variety of data sources. In addition, the emergence of a wide variety of new data providers, from commercial data retailers to managers of open-source, Web-based datasets introduces a potentially rich array of new, useful sources of transportation data. On the other hand, the demise of the mandatory Census Long Form represents a significant loss in high quality, detailed data about our urban regions and poses new challenges for the validation and weighting of travel surveys.

In parallel to this changing technical landscape, the needs and capabilities of our transport planning agencies are changing. Among other factors, these vary considerably by the size of city or region. In a variety of ways, all urban areas face growing challenges such as:

- continuing urban growth and congestion
- greenhouse gas emissions (climate change)
- air quality
- ecosystem health
- economic productivity
- safety and security
- capital and operating financial burdens

Increasingly difficult fiscal environments for agencies require them to be ever more cost-effective within typically shrinking budgets. Data collection is often viewed as a “costly frill”, despite the critical role that

it plays in the planning and design of multi-billion dollar transport systems. Limited budgets also too often translate into limited staff resources for data collection and management. And yet it is exactly in times of fiscal restraint such as we currently face that the need is greatest for cost-effective decision-making, and this can only be achieved if the right information (including model-based forecasts, from models built on high quality base data) is available to inform and support these decisions.

For all these reasons it is essential that transport agencies have a clear and comprehensive understanding of the technical options available to them, of the strengths/weaknesses and benefits/costs of these options, and clear guidance on cost-effective data collection programs that will best deliver the required information in keeping with their local needs and budgets.

The "business case" for urban passenger data collection

The business case for transportation data collection, management and dissemination depends on obtaining the benefits associated with a better planned, monitored and regulated transportation system that, collectively, outweigh the cost of data collection.

As illustrated in Table 1, the annual investment that Canadians make in the transportation system is enormous, totalling $39.5 billion in 2010 (with $33.9B of this being spent by provincial, territorial and local governments). Significant costs also arise from the inability of transportation infrastructure supply to meet the demands of travellers, notably recurrent peak period delay to drivers, increased fuel consumption, and higher greenhouse gas emissions.

By contrast, the typical cost of undertaking a major household travel survey in Canada is very modest. For example, in 2006, 150,000 households were surveyed in Greater Toronto's Transportation Tomorrow Survey (TTS) for $3.09M, inclusive of all aspects of the survey, including data collection, data management and reporting of the results, a total cost of $20.38 per completed interview. Similarly, in 2008, Montreal's Household Travel Survey invested $1.5M, at a unit cost of $19.45 per household.

If 5% of all Canada's roughly 15 million households were to be surveyed at a cost of $20/interview once every five years (a typical frequency for such surveys), the total cost would be $15M, or $3M per year. Using the lowest cost estimates for infrastructure and congestion cited in Table 1, this annual cost is equivalent to less than 1/100th of 1% of the total annual budget spent on transportation in Canada, less than 30 metres of subway line, or less than the cost of one morning's congestion for commuters in large urban areas of Canada.

Table 1 – Representative Transportation Systems Costs

Canadian Transportation Expenditures, 2010 ($Billion)[1]

Mode	Amount
Roads	28.9
Public Transit	5.8
Marine	1.8
Air	1.2
Rail	0.4
Multi-modal/Other	1.4
Total	39.5

Highway Construction and Maintenance Costs[2]

Facility	Unit Cost
Rural highway construction	$380,000 per km
Annual road maintenance	$1,045-8,700 per km, 2 lanes
Bridge construction	$1.8-4.4 million 1000 m^2 bridge

Rail Transit Construction and Maintenance Costs[3]

Facility	Date	Amount ($M/km)
Canada Line, Vancouver	2005-09	105
Sheppard Line, Toronto	1994-2002	170
Spadina Line, Toronto	2009-15	306
Sheppard Extension, Toronto[4]		177
Orange Line Extension, Montreal	2003-07	143

Rail Transit Maintenance/Rehabilitation Costs

Facility	Date	Amount ($M)
Reno-System (subway upgrades), Montreal	2004-12	950
Reno-System, Phase 3		500

Cost of Congestion ($Billion/year)

Location	Date	Amount
Canada, major urban areas[5]	2006	2.3-3.7
Montreal Region[6]	2009	1.4
Toronto Region[7]	2006	3.3

References

1. CTC (2010)
2. Applied Research Associates (2008)
3. Toronto Transit Infrastructure (2012)
4. Estimate, not yet constructed
5. Transport Canada (2006
6. MTQ (2009)
7. Metrolinx, The Big Move (2009)

It is not difficult to imagine that a sound data collection effort across Canada could lead to infrastructure, congestion and pollution savings far in excess of these, by directly contributing to improved evidence-based transportation planning.

Data collection needs and methods

As illustrated in Table 2, many types of data are used in urban transportation planning, including:

- household activity / trip-making behaviour
- count data (traffic, riders, etc.)
- transportation system characteristics (speeds, lane widths, etc.)
- transportation cost and service levels
- land use characteristics (population, employment, etc.)
- population socio-economic information (income, auto ownership, etc.)
- attitudes / opinions / stated choices
- system impacts (e.g., emissions)

As is also shown in Table 2, numerous data collection methods exist that are suitable for collecting various types of data, depending on the data type and the eventual usage of the data. These data collection methods, however, can be usefully grouped into three broad categories:

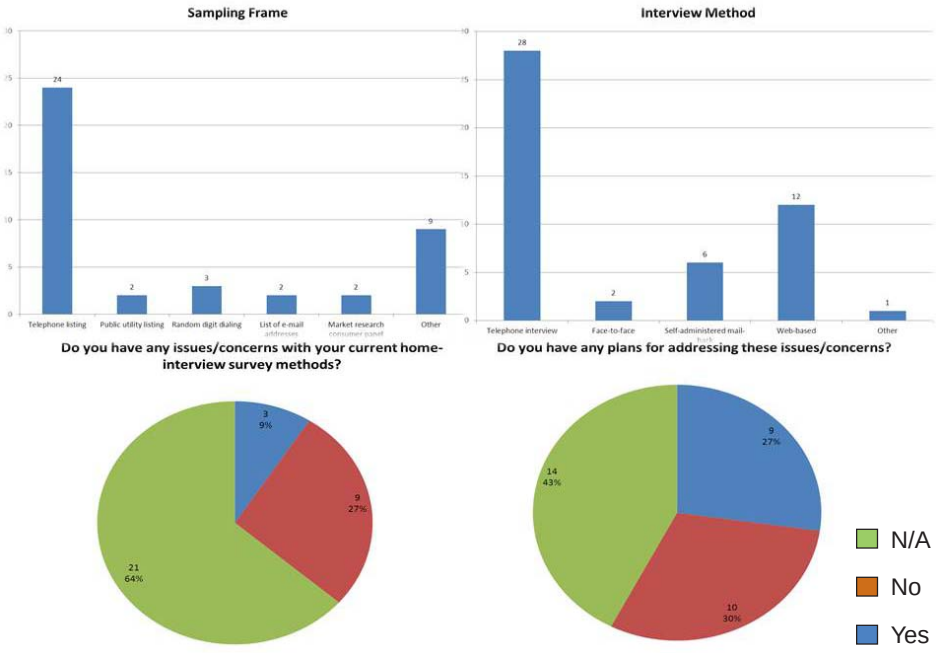


Figure 2 – Household Travel Survey Usage in Canada

Table 2 – Classification of Data Types & Methods

Conventional Methods

Data Type	Face to face	Household-Based Surveys			Choice-Based Sample Surveys		Standard Technology-Based Methods			
		Telephone	Mail-back	Web-based	Road-side	On-board	GPS	Roadside detectors	Smart-card	Mobile phone
Household activity / trip making behaviour		X	X	X	X	X	X	X		X
Count data						X	X		X	
Transportation system characteristcs							X	X	X	X
Transportation costs and service levels	X	X	X	X			X	X	X	X
Land use characteristics										
Population socio-economic information	X	X	X	X	X	X				
Attitudes / opinions / stated choices	X	X	X	X	X	X				
System impacts (e.g. emissions)								X		

Emerging Technology Based Methods

	Remote sensing (satellite / aerial)	Web "apps"	Social network software	Smart phone	Acceler-ometers	Personal health sensors	Environmental sensors	Blue-tooth
Household activity / trip-making behaviour		X	X	X	X			
Transportation system characteristics	X							X
Transportation costs and service levels	X			X				X
Land use characteristics	X	X					X	
Population socio-economic information		X	X	X		X		
Attitudes / opinions / stated choices		X	X	X				
System impacts (e.g. emissions)								

- population-based surveys (such as home-interview surveys)
- choice-based sample surveys (such as transit on-board surveys)
- non-survey data collection techniques (both "standard" and "emerging"), which typically involve use of IT of one sort or another

As indicated in Figure 2, most Canadian urban regions currently are heavily dependent on household-based travel surveys as a primary source of information concerning travel behaviour. These surveys typically use land-line telephone directories as their sampling frame. This approach is becoming increasingly untenable as fewer households have listed land-line telephone connections, and people increasingly use call-screening or subscription to "Do Not Call" lists to avoid taking survey-related calls. Thus, an urgent need exists to find alternatives to this traditional approach to travel behaviour data collection. At the same time, smart-phone based apps that use either GPS or cellular signal triangulation, transit smart-card data, Web-based interview techniques, etc. are raising new possibilities for the collection of large amounts of useful data, and Canadian agencies are increasingly experimenting with such technologies (see Table 3). However, they are not without their own limitations – for example, much smart-phone data may not provide information concerning trip-maker characteristics – and their usage is not yet standardized or in widespread operational use.

Table 3 – Data Collection Technology Usage in Canada

	Use now	Plan to use in next 5 years	Do not use
Global positioning systems (GPS)	57	16	18
Other distributed or remote sensing technologies	42	3	43
Smartcards	25	19	46
Other transit pass technologies	16	6	63
Debit/credit cards	23	11	52
Social media	29	22	35
Other internet	45	12	30
Mobile devices	27	24	38
Other technologies	9	5	63

The need therefore exists for a new data collection framework for meeting Canadian transportation planning needs that robustly and cost-effectively:

- exploits current and emerging data collection methods and technology

- takes maximum advantage of existing datasets
- is transferable across different planning contexts within Canada

Elements of a new framework for data collection

Given the complexity and variety of data collection needs and methods a new data collection framework must include the following elements:

- An institutional (and political) commitment to on-going collection and management, that is not ad hoc or fragmented, of high-quality data to support urban transportation planning and analysis needs.
- A careful assessment of an agency's need for, and use of, data, leading to a comprehensive "model" of its data requirements. Ideally, multiple agencies within a given urban region should develop a collaborative, cost-effective and common data collection strategy.
- The wise exploitation of the enormous amount and variety of data that are potentially relevant to transportation planning applications, and that are collected by a vast number of public and private organizations. Given that such data are typically collected for non-transportation purposes, a given dataset may address a particular transportation modelling or analysis need only partially. However by the fusion of two or more datasets,(see below) it may be possible to construct more comprehensive datasets that can address this need.
- No one survey or other data collection package is likely to be able to collect all the data required today by a transportation agency. A multi-method approach, therefore, is generally required to comprehensively and cost-effectively meet an agency's full data needs. Again, data fusion techniques may be needed to integrate data gathered by the various methods used.

These framework elements, including brief example applications are discussed further below.

Data models

An "object-oriented" or "ontological" approach is recommended to develop a systematic model of data needs, the relationships among and between transport users and agencies, and the interconnections across applications and agencies (see example in Figure 6).[1] In addition to guiding the organization of a data collection program, such a data model can be the starting point for designing an efficient data management system, as well as a conceptual framework for the development of a consistent, comprehensive set of transportation models and other analytical tools needed for the range of an agency's planning and analysis needs.

Existing data sources

Canadian transportation agencies have always exploited a variety of data sources to support their modelling, planning, policy analysis and other analytical needs. Given the proliferation of government and private sector databases over the past number of years, available via the Internet among other sources, it is important to investigate the possibility for using new data sources (or existing data sources in new ways) in urban transportation applications. As shown in Figure 3, at least eight major sources of transportation-relevant data exist and are used by Canadian urban transportation agencies.[2]

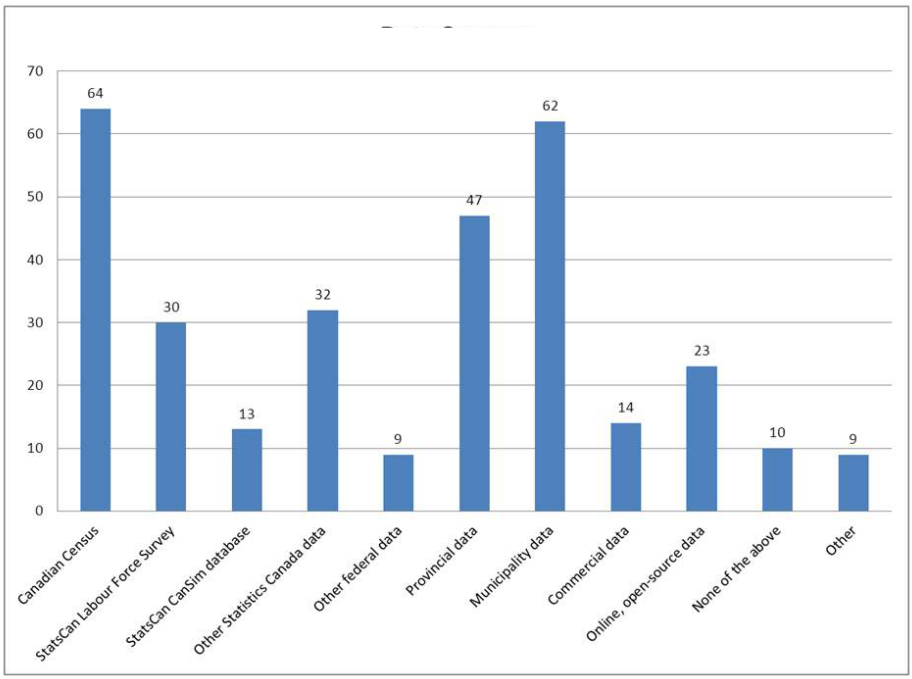


Figure 3 – Sources of Data for Urban Transportation Planning Applications

Data fusion methods

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. It is often possible to fill these gaps through the use of statistical *data fusion* methods to integrate two or more individual datasets to create a new, more comprehensive dataset. Data fusion involves:

- Combining two or more datasets together to create a combined dataset that is more comprehensive than either of the original dataset, thereby filling information gaps and extending the usefulness of the available data.

- The statistical imputation of missing variables in a "receptor" dataset by exploiting correlations between these variables and a set of variables that is common across the two datasets.

A common example of this is the use of Census data to add socio-economic variables (for example income) to travel survey records that are missing these variables. Another common example is the use of transit on-board ridership survey data in combination with a home-interview travel survey to develop a mode choice model. Figure 4 illustrates a third example in which transit smart card data are used to enrich OD travel survey records. Given the increasing use of "third-party" datasets, as well as multi-faceted data collection approaches (discussed below), the ability to fuse multiple data sources together to create unified databases is increasingly important in urban transportation planning analysis and modelling.[3]

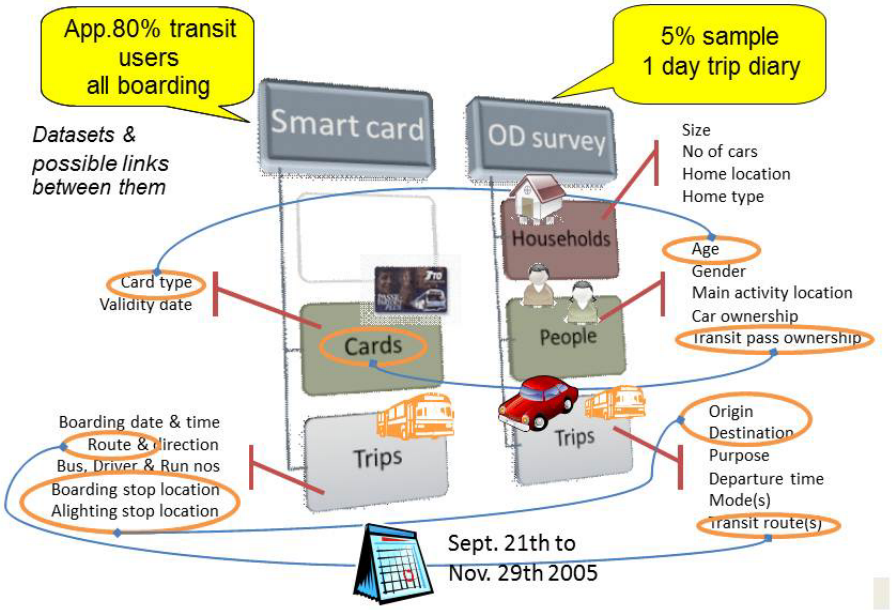


Figure 4 – Fusing Smart Card & OD Survey Data

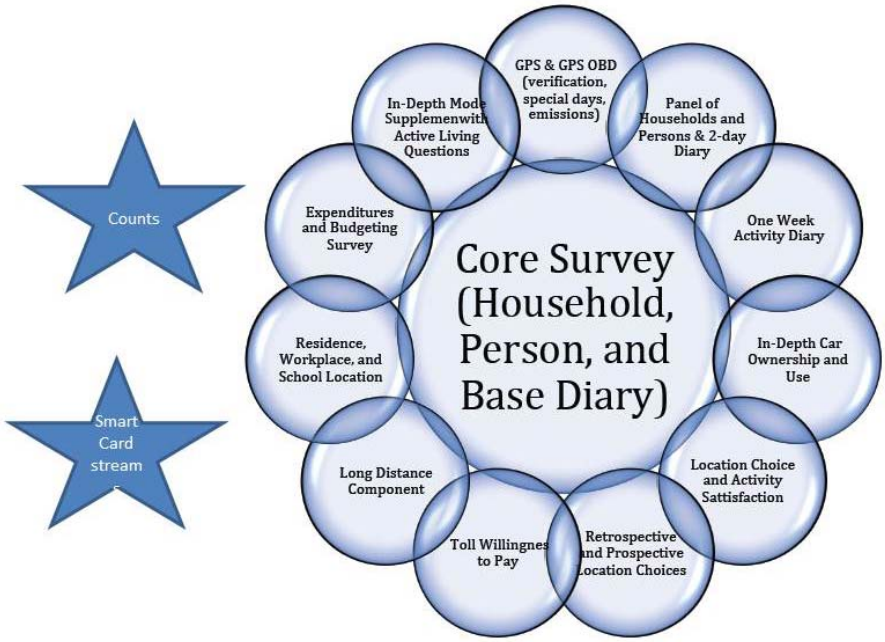
"Core-Satellite": A new data collection paradigm based on "Fusion by Design"

As already noted, a multi-faceted approach to data collection is generally required, reflecting both the complexity of the behaviours being observed (which may not be feasible to capture with a single sample or instrument), and the need to keep response burdens for a particular survey within reasonable levels. A particularly attractive and generalizable approach is the core-satellite design illustrated in Figure 5, involving the following components:

- *A core survey, which is a large-sample survey to gather primary information concerning the respondents and their key behaviours.*

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. It is expandable to make statements about the full population; consistently applied over a large geographical region; stable (but not necessarily static) over time; and applied relatively frequently (or continuously) so as to provide consistent time-series data.

- *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*. 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.
- *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.



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

Figure 5 – The Core-Satellite Survey Design Template

The core-satellite paradigm is an extremely flexible and generalizable approach to meeting different agency needs, whether it be the building of an evidence base in the longer term, or throwing light on "hot-button" issues. Although multiple methods are necessary, the main guiding principle is the integration of content rather than method (i.e., defining what content is core, and what can be collected via a satellite process). Different agencies will use different methods for both their cores and their satellites, depending on their data needs, resources, etc.

The focus of core-satellite approaches is the creation of *integrated databases*, and for this fusion techniques are essential. But of central importance is that fusion, more familiar as a way to merge existing data opportunistically, is undertaken under the core-satellite paradigm *by design*. The core-satellite paradigm (ideally combined with the object-oriented data model discussed above) allows each agency to think through and design the content and structure of its own integrated data collection, acquisition and management methods in a comprehensive, consistent manner to best meet its own analysis, modelling and planning needs.

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.

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.

A Hypothetical Example Application

Figure 6 uses an object-oriented data model to illustrate the variables currently collected in a typical Canadian home interview survey (highlighted in red in this figure). As indicated, such 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.

As illustrated in Figure 7, 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. The core-satellite data can also be further enriched by a variety of other non-linked (but still useful) datasets from a variety of sources (the blue boxes in Figure 7).

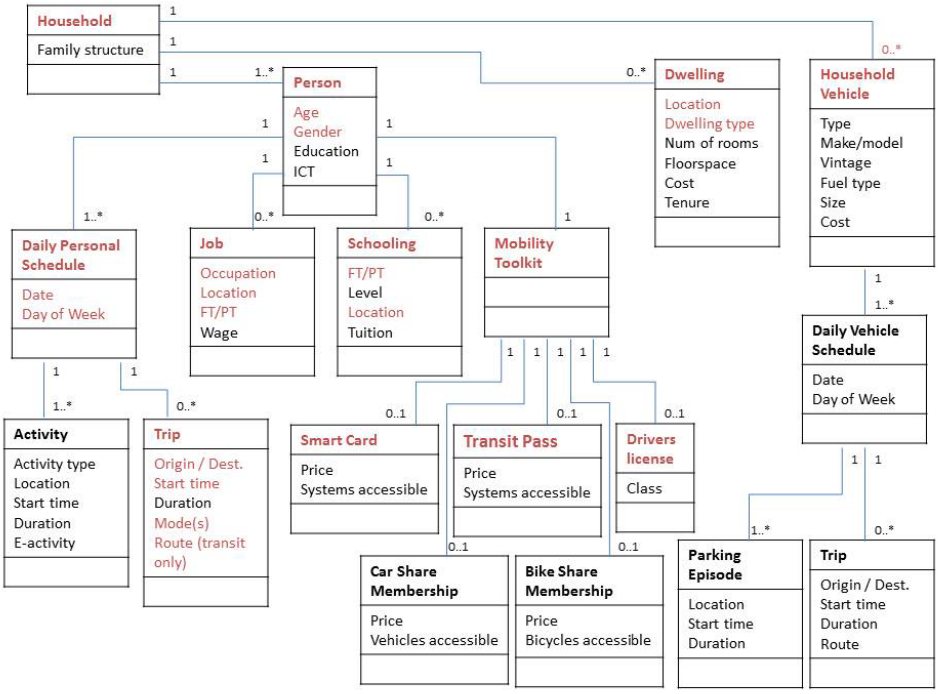


Figure 6 – Object-oriented Data Model for a Typical Core Home Interview Survey

Committing to evolutionary change in data collection methods

Knowing that the status quo in Canadian urban passenger transportation data collection is unsustainable, data collection methods must adapt to fundamental changes in travel behaviour, technology and planning needs. In particular, household travel survey methods that have been successfully employed for decades are being increasingly challenged by changes in the nature and use of telephones. Increasing analytical and modelling demands are being placed on survey data to support complex planning analysis and decision-making needs. At the same time, technological options for data collection are emerging and evolving rapidly. While many of these new technologies are very promising, they are at various levels of maturity, and alone cannot meet all the challenges. Like any tool, they have their strengths and weaknesses that can be addressed best by embedding them in a stable environment of sound survey and data management.

Given the need for both change and continuity, what is required is a process of careful (but on-going) evolutionary change in our data collection methods and state of practice. If this evolutionary change is to occur (and in a cost-effective way) it will require agencies that explicitly adopt an "R&D mentality" in their data collection program. Such agencies would recognize that every survey without an experimental component is an opportunity wasted, that planning for next time is part of the plan this time, and that it is essential to share what is learnt from the experiences of other agencies across the country and internationally.

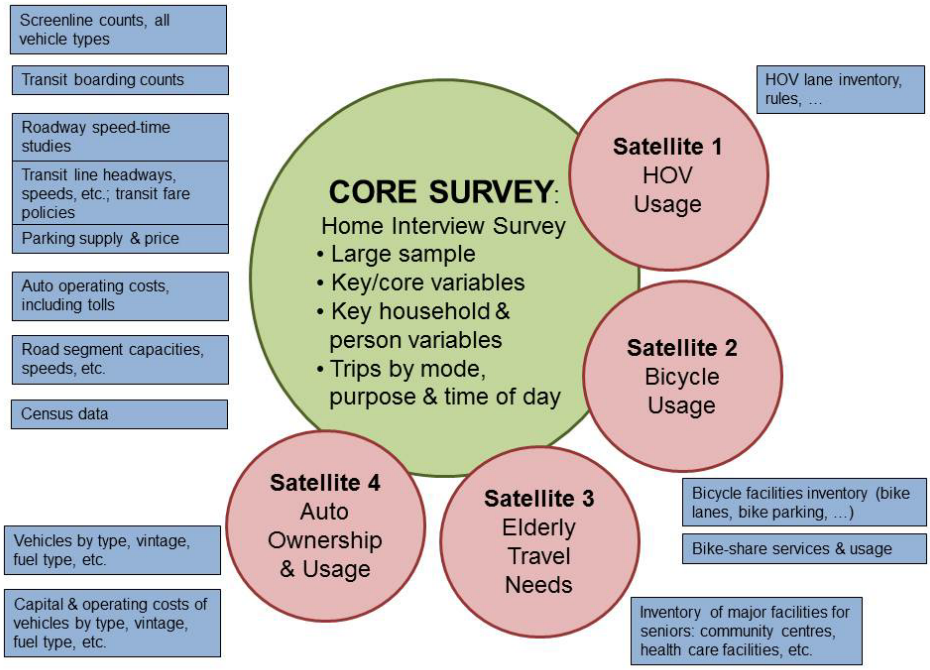


Figure 7 – Example Core-Satellite Data Collection Design

What about Canada-wide data?

Finally it is worth noting that, unlike many countries, Canada does not have a national urban travel survey covering all modes of travel. This is not particularly surprising, given the constitutional constraints on federal involvement in urban affairs which have historically limited the role of the federal government in urban transportation. Given this, a direct, "top-down" development of a national survey is not likely to be a feasible proposition, at least within the short- to medium-term. Nor is it clear that it is necessarily the best approach to improving data collection standards and practice across the country.

Given, however, that all major urban regions in the country routinely undertake major household surveys, and given that, at least in the largest regions, these already encompass a significant proportion of

provincial populations, a much more promising approach to "growing" data collection capabilities and improving/standardizing best practice would be to expand the current urban surveys to province-wide levels. This approach would have significant benefits to each province in terms of:

- Providing uniform travel data across the province.
- Providing smaller and medium sized cities and towns with vastly improved travel data.
- Eliminating "boundary" effects in the analysis and modelling of urban areas with constantly expanding development boundaries.
- Providing a framework for gathering longer-distance intercity and rural travel behaviour as well as local/urban. For example, a satellite survey could be added to the local travel core survey to gather long-distance travel.

Nationally this approach would provide an "organic", voluntary, one-step-at-a-time approach to evolving a national data collection program that would be driven "bottom up" by the provincial and municipal organizations, who are both closest to the problem and the greatest beneficiaries of access to improved data. It would promote and facilitate collaboration and the sharing of data and experience among provinces and their constituent urban areas across the nation. And it would encourage experimentation through the spreading of risk and, possibly, pooling of funds.

(Endnotes)

1 *CPiDC, Part VI, Section 3.5 "An Object-Oriented Model of Urban Activity/Transportation Systems".* See also Gruber, T. R., (1993) "Towards Principles for the Design of Ontologies used for Knowledge Sharing", *Proceedings of the International Workshop on Formal Ontology*, Padova, Italy., and Campbell, A.E., and Schapiro, S.C., (1995) "Ontologic Mediation: An Overview", *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, Menlo Park CA, USA: AAAI Press.

2 A large inventory of currently available, transportation-related Canadian datasets is provided in *CPiDC, Part IV: Survey of Data Sources for Urban Transportation Applications, Appendix.*

3 For a detailed discussion of the data fusion problem and the variety of methods available for integrating/fusing individual datasets to create new, more comprehensive datasets see *CPiDC, Part III: Data Integration/Fusion Methods.*

More information

The information in this primer has been extracted from the Transportation Association of Canada (TAC) publication *Changing Practices in Data Collection on the Movement of People* (2014), which is available for purchase in TAC's online bookstore.

Disclaimer

Every effort has been made to ensure that all information in this primer is accurate and up-to-date. The Transportation Association of Canada assumes no responsibility for errors or omissions. The primer does not reflect a technical or policy position of TAC.

Transportation Association of Canada
2323 St. Laurent Blvd., Ottawa, ON K1G 4J8
Tel: (613) 736-1350 Fax: (613) 736-1395
www.tac-atc.ca