Transportation Data Quality: What It Means And How To Get It

Ken Cervenka, North Central Texas Council of Governments

Abstract

Attempts to significantly improve travel demand forecasting procedures should consider the availability and quality of data in four primary areas: transportation supply (e.g., roadway and transit networks), land use information (e.g., population and employment estimates and forecasts), observed travel (e.g., time-of-day motor vehicle counts, transit ridership, and travel times), and behavioral information (e.g., the activities and travel of individuals). This paper and presentation will address all four areas, but fill focus on the recent collection of travel survey data by the North Central Texas Council of Governments (NCTCOG), the Metropolitan Planning Organization (MPO) for the Dallas-Fort Worth region.

The author's findings are based on first-hand experiences with the management of four projects (external travel survey, workplace survey, household survey, and transit onboard survey) administered by four separate consulting teams, as well as interactions with Los Alamos National Laboratory on the conceptualization of a "next generation" travel model. Any agency considering a new survey should first contemplate the issues that will impact the quality of the collected data, such as survey objectives; degree of risk; and trade-offs between the cost, quality, and quantity of the data collected.

Look up the word "quality" in the dictionary and you will see definitions such as "degree of excellence" and "superiority in kind." A more pragmatic definition of "data quality" for the transportation planning community is "information that leads to better transportation decision-making." And how do you get this kind of data? The answer seems simple: 1) figure out (as best you can) what you really need; 2) design a program that fits within your constraints; and 3) implement the program. The degree of success for any data collection program centers on how well these three tasks can be accomplished.

For the discussion that follows, the perspective of a Metropolitan Planning Organization (MPO) responsible for travel demand forecasting is taken. An MPO is charged with the responsibility of developing and maintaining the region's transportation plan, which serves as the blueprint for the region's future transportation system. Ultimately, an MPO's objective is to improve the "quality of life." Following some background information, the topics covered include a description of four types of data, hypotheses on data collection, data collection objectives, and some closing thoughts.

Background

The North Central Texas Council of Governments (NCTCOG) is the MPO for the 5,000-square-mile, four-million-person Dallas-Fort Worth Metropolitan Area. An alternative title for this paper could have been "Lessons Learned From Dallas-Fort Worth Experiences," for the author's knowledge of data quality issues is based on involvement in three major activities:

1. *Travel Model Development.* Since the late 1970s, NCTCOG has been running a highway/transit travel demand forecasting model that resides on an IBM mainframe computer. It is primarily a customized version of the Urban Transportation Planning System (UTPS) package

developed by the U.S. Department of Transportation in the 1960s and 1970s. The last major round of model calibration/validation was conducted in the late 1980s and was based on the 1984 household, workplace, and transit on-board surveys; the 1980 U.S. Census Journey-to-Work data; and highway and transit passenger counts. NCTCOG has spent considerable research time, in recent years, identifying the functional requirements and data needs for a near-term and long-term travel demand forecasting system that will reside on in-house computers.

2. *Regional Travel Surveys.* In the 1994-1996 time period, NCTCOG organized a number of surveys that had three primary objectives:

   - To obtain data needed for re-calibration of the existing four-step model process for the Dallas-Fort Worth Metropolitan Area;
   - To provide the data needed for testing new demand model strategies; and
   - To develop broader, more management-oriented (and policy-sensitive) forecasting procedures.

   Five major survey efforts, with an overall price tag of 1.5 million dollars, have been completed:

   - The External Travel Survey (by Wilbur Smith Associates) consisted of roadside interviews of 28,000 drivers at 38 locations (outbound direction, as the vehicles left the Metropolitan Area) in March and April of 1994.
   - The Workplace Survey (by Barton-Aschman Associates, Inc.) consisted of 20,000 visitor interviews and 7,000 completed employee questionnaires for 278 workplaces from September to November of 1994.
   - The Transit Origin-Destination Survey (by NuStats International) consisted of 4,075 completed questionnaires obtained from riders of the Fort Worth Transportation Authority's fixed-route services in May of 1996.
   - The Household Activity Survey (by Applied Management and Planning Group) consisted of the completion of one-day diaries for all members of over 4,000 households from March to May of 1996.
   - The Stated Preference Survey (by Applied Management and Planning Group, with Mark Bradley Research and Consulting as subcontractor) consisted of "trade-off choice" questionnaires mailed back by more than 500 individuals who had previously participated in the Household Activity Survey. The mail-out/mail-back survey was conducted in the summer of 1996.

3. *TRANSIMS Case Study*. TRANSIMS (TRansportation ANalysis and SIMulation System) is a "next generation" travel simulation and forecasting system being developed by Los Alamos National Laboratory (LANL) as part of the multitrack, multiyear National Travel Model Improvement Program. It is referred to as a "bottom-up" computational approach because the simulated interactions of individual behaviors are used to observe aggregate dynamic (i.e., emergent) behaviors. NCTCOG has been working with LANL since 1995 on a case study application of the first interim operational capability of TRANSIMS: Traffic Microsimula-

tion. The experience has given NCTCOG staff new insights about the large-scale needs for different types (and accuracies) of data.

Four Types of Data

The theme of this paper is that data quality refers to "information that leads to better transportation decision-making." As viewed by a travel modeler, there are at least four primary types of data that will impact a travel model, and, ultimately, the value of the transportation decisions that are based on the travel model results:

1. *Demographics (Land Use)*. These are the estimates and forecasts of all variables needed for calculating person trip (or activity) production and attraction rates and input values for mode choice calculations. Typical zone-based examples include population, households, average household income (or income distribution), auto ownership, employment, and area type.

2. *Transportation Supply*. Examples include the specification of all attributes of roadway links/intersections and transit routes/stops that are needed for a travel model run.

3. *Observed Travel (Aggregate Transportation Demand)*. Examples include time-of-day counts (for all relevant modes of transportation) and observed highway/transit travel times for specific time periods. The information is not used as input to a travel model, but rather as a means of calibrating (and ultimately validating/verifying) a travel model formulation. [Note: a mistake made by some modelers is to assume that the data used for model calibration can also be used for model validation].

4. *Behavioral Information (Disaggregate or Individual Transportation Demand)*. Examples include information about the actual activities and travel of individuals (revealed preference), as well as their predicted activities and travel under non-observable conditions (stated preference/stated response). For detailed information about the many kinds of surveys for obtaining behavioral information, refer to the June 1996 U.S. Department of Transportation and U.S. Department of Energy report, Travel Survey Manual.

Hypotheses on Data Collection

Here are six hypotheses (or assumptions) to be considered, prior to development of a detailed data collection program design:

1. There are uses for data that go beyond the direct needs of travel demand models. Demographic and land use data, for example, is used for a variety of planning purposes. Observed travel data can be used for preparing detailed summaries of transportation system performance and behavioral data can be used for policy analyses. For example, even if information about "work at home" patterns is not expected to be incorporated in a travel model, the information may be useful for preparation of a Travel Demand Management (TDM) program.

2. The ultimate value of any travel demand model is tempered by the availability and accuracy of existing/predicted data. A term coined in the 1960s, with the advent of increased computer usage, is GIGO: Garbage In/Garbage Out. Concerns with GIGO are just as relevant today as they were 35 years ago. Data is needed not only to calibrate and validate the equations and parameters contained in new travel model formulations, but must be forecastable for use as input in future model runs.

3. The ultimate value of any travel demand model is also tempered by how (or whether) the available data will actually be used. For example, a program to gather detailed signal timing/phasing data will not improve a travel model if there is no mechanism for incorporating this level of intersection detail. For another example, it is common practice to "throw out" travel time runs that occurred during unexpected events (e.g., a freeway accident), but yet this may give us good information about the frequency of non-recurring congestion and the reliability of a roadway segment.

4. It is not clear whether we should get the data to fit the models we want, or develop the models to fit the data we can get. Should our data collection program be designed to meet the requirements of a specific travel demand model construction, or should we instead be using the data to help us develop a new model structure?

5. The best approach for one agency will NOT be appropriate for all agencies. Perhaps one way to deal with Assumption #4 is to realize that some agencies are willing to accept the risks associated with "pioneering research," whereas other agencies are content to follow established practice. Agencies (as well as their employees) simply have different opinions about operating outside of their "comfort zone."

6. No data collection program will ever be perfect. The "Holy Grail" of a perfect data collection program is simply not attainable, at any cost. Some compromises and risks will be necessary, for we cannot conduct new surveys every time we think of a new data item that might be of value to the next round of model development.

Data Collection Objectives

If the organization paying for and using the data (the client) is different from the organization collecting the data (the contractor), it is likely that the program objectives for these two organizations will be different. The ultimate value (i.e., quality) of the collected data will depend on how each party deals with their separate objectives. For example, consider the data collection objectives from the perspective of a client that will be performing travel model calibration/validation:

1. There is a purpose for collecting data that goes beyond simply collecting data. While the delivery of the data may be the contractor's final product, the client's real work is just beginning. If the data is not expected to improve the client's transportation decision-making process in some definable way, then there is probably no valid reason for collecting the data in the first place.

2. Time and/or cost constraints are most likely prevalent in all decisions. The client would, of course, like to find the contractor that can deliver the highest quality (and quantity) of data at the lowest possible cost, in the shortest possible time, and with no risk to the client. In reality, some compromises will need to be made, and the client's early task is to choose the contractor that (in the client's opinion) will most likely deliver the "best" overall product.

3. The contractor must ultimately deliver what the client considers the "best possible" product. From the client's perspective, the contractor should deliver all work that was promised, as well as "cover" any additional requests the client makes during the contract period. In reality, the client must work closely with the contractor to make various trade-off decisions and compromises, even after the final contract has been signed.

The contractor, on the other hand, may be working under a different set of objectives that center on the fact that a particular data collection effort is just one of many commercial transactions for the firm:

1. The contractor is running a business. To stay in business, the contractor must, over the long run, "make money" on many (although not necessarily all) projects.
2. The contractor is in competition with others. A particular data collection effort is "won" by offering a proposal that is most attractive to the client (in some way) then the competitors' proposals. It is therefore not always possible for the contractor to propose what he/she would really like to do (i.e., deliver the highest-quality product), but rather what the future client thinks should be done.
3. The client is expected to be "reasonable." Problems will most likely be encountered during any large data collection program, which means that a good client/contractor relationship must be established and all roles clearly defined.
4. The contractor wants to please the client and do meaningful work. It is generally "good for business," over the long run, for the contractor to not only deliver a product that meets all contractual obligations, but to deliver what the client will consider the "best possible" product-even if there is an extra expense that cannot be charged to the client.

Some Closing Thoughts

As noted at the beginning of this paper, transportation data quality can refer to "information that leads to better transportation decision-making." A data collection program should be designed so that it gets the data that's really needed, within the known time and cost constraints. Here are a few closing thoughts for an agency planning a new data collection program:

1. The ultimate objectives for use of the data should be defined, as much as possible, before any data is collected. Consider the use of a consultant "coach" or expert panel to help with the identification of needs.
2. Decisions must be made between potentially conflicting objectives: how much effort should be expended to get data to be used to update an agency's existing four-step model, and how much should be expended in the pursuit of an alternative approach? Also, how much effort should be expended for data that is needed for purposes other than travel modeling?
3. Consider a risk assessment: can data collection methodologies implemented in other regions be used with only minor revisions, or is a major new survey design effort (with extensive pre-tests) warranted?
4. Are data summaries already prepared for other regions of value to your agency? If so, it may be possible to reduce (or redirect) your own data collection program.
5. If a contractor is hired, be sure that everyone agrees on the roles and responsibilities for data collection design and administration. Even a binding "iron-clad" client/contractor agreement requires trust and respect among the parties, especially if new procedures are being tested. Also, be sure there is agreement on how "acceptable quality" for the final survey data is defined.

6. Rather than seek the “Holy Grail” of all data collection efforts, it is easier to simply accept (and plan for) the fact that no program is going to perfect.