Sampling issues, data quality & data protection

Jimmy Armoogum, Jennifer Dill

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2016) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

10th International Conference on Transport Survey Methods

Sampling issues, data quality & data protection

Jimmy Armoogum[a]*, Jennifer Dill[b]

[a]*UPE-IFSTTAR-DEST, 14-20 Boulevard Newton - Cité Descartes - Champs sur Marne, F-77477 Marne la Vallée, France*
[b]*Nohad A. Toulan School of Urban Studies & Planning,* Portland *State University, Portland, Oregon, United States*

Abstract

This workshop discussed various aspect of the mathematical part of survey methodology, as well as archiving and confidentiality issues aimed at improving data quality and its use through time. Participants identified ways to correct or minimize bias by dealing with incomplete sampling frames, using weighing and imputing procedures. We discussed methods to archive and share GPS-based survey data to preserve anonymity. Finally, we debated research needs on these topics for the next following years.

Peer-review under responsibility of International Steering Committee for Transport Survey Conferences ISCTSC.

Keywords: Sampling, weithing, imputing, nonresponse, bias, data quality, measurement errors, archiving, data privacy

1. Introduction

Issues related to sampling and data quality can affect the usefulness of any travel survey. This workshop addressed issues such as respondent bias, non-response, sampling issues, and declaration bias, along with models that can help reduce the impacts of these issues and preserve data quality for transport analysis. The workshop also addressed the challenges of data protection, including confidentiality and storing to ensure long-term usability.

2. Sampling Issues

A sampling frame provides the means to reach each member of the study population who would be eligible to be surveyed. Researchers have several options for sampling frames, including using household- or person-based sampling and phone or address sampling frames. A common approach for travel surveys has been to sample households and interview all members of a household. One advantage of this approach is that the researcher is able to analyze how household members interact in travel behavior and how they share the vehicles. The disadvantages

* Corresponding author. Tel.:+ 33 1 81 66 86 07; fax: 33 1 81 66 80 01.
E-mail address: jimmy.armoogum@ifsttar.fr

2214-241X
Peer-review under responsibility of International Steering Committee for Transport Survey Conferences ISCTSC.

of this approach are that is not easy to contact everybody in a household and interview them about the same day (or week) and that the burden of the household may be very high. This can result in a low response rate if a household is considered non-responsive if only one member is not interviewed. This will also bias the response rate to smaller households. Often, the survey will allow proxy responses—one household member can report data for another household member. However, this process introduces other sources of error and bias (Badoe & Steuart, 2002).

One problem with any sampling frame is that they are often incomplete; they are missing some of the eligible sampling units. For example, some people have phone numbers that are not in a phone registry. The practices of phone number registries vary significantly between countries. Random digit dialing (RDD) can address the issue of incomplete sampling frames for phone samples, but poses other problems. For example, the researcher cannot send the household an introductory letter prior to the phone survey – a technique that has been shown to improve response rate.

The decision of how to address incomplete sampling frames would be made easier with more information about who is missing. In most cases, it is unlikely that the missing people or households are random. Knowing, for example, the demographics of the missing people can lead to better decisions, such as choosing another frame, sampling strategy, or weighting, and allows the researchers to most appropriately interpret their findings.

One increasingly common approach to dealing with sampling bias is to augment the sample with additional sources, use more complex sampling frames. For example, it is becoming more common for surveyors to join mobile and land-line phone samples. Researchers might also use two-stage sampling approaches, stratified sampling, and over-sampling – all to help ensure that harder-to-reach individuals (e.g. younger adults, people who use less-common travel modes, etc.) are included. These approaches also help address another issue—collecting data on relatively rare behavior. For example, long distance mobility is very unevenly distributed in the population (e.g. 47% of individuals made no trip over 100 km during at least 3 months before the interview, according to the last National Travel Survey conducted in France). Thus uniform sampling is especially inefficient for understanding long distance mobility.

These more complex sampling frames pose other difficulties. For example, when two or more data sources are combined, care must also be taken to minimize duplicate entries and address potential inconsistency that may appear in both sampling frames. In addition, calculating confidence intervals is more challenging.

3. Data Quality

3.1. Non-response

Non-response is the inability to measure all the units of sample of all variables of interest. Two different types of non-response exist: (1) Total (or unit) non-response, where there is no information about the unit selected other than sample frame; and (2) Partial (or item) non-response, where the unit selected responded only to a part of the survey.

3.1.1. Unit Non-response

Falling response rates are a growing problem for travel survey researchers. The increased reliance on mobile phones is pointed to as a key contributing factor in this trend. While this trend is seen worldwide, there are still significant differences in response rates across countries. However, workshop participants noted that comparing response rates across surveys is not a simple task because of variations in what counts as a response.

While non-response results in a reduced sample size, a more important concern of researchers is the possible impact of non-response bias. Bias is introduced when those that do not respond to the survey are systematically different from those that do respond on key variables of interest. Researchers first must *understand* the unit non-response. For travel surveys aimed at representing the general population, this is commonly done by comparing the sample to a census. However, this will only identify the demographics of the underrepresented units, e.g. lower-income households or younger adults. It does not tell us how the travel behavior of the non-respondents might systematically differ from those of the respondents. For example, is it safe to assume that younger adults that do not respond to the survey travel the same as younger adults who did respond? Possibly not.

There are several ways to better understand the potential impacts of unit non-response bias. One is to compare the respondents who were easy to contact to those who responded, but required more contact and effort to get to

respond. The latter group might be more similar to the non-respondents than those who responded with minimal effort. A second way is through non-response surveys. The workshop presentation by Wittwer and Hubrich focused on a survey of 4,802 people who did not participate in the 2013 Mobility in Cities SrV survey in Germany. They found that non-participants were more likely to be in smaller households (particularly one-person), be between 15 and 44 years old, be employed, and not have a driver license. A third way to understand the unit non-response bias is to examine long-term surveys (e.g. Germany) to understand changes over time. Finally, researchers could compare different surveys (e.g. different countries, cities, regions).

There are two common responses to dealing with unit non-response: (1) increasing the response rate; and (2) weighting data to correct for non-response errors. A less common method is through imputation.

To increase response rates, we must first understand the many reasons people do not respond. Several factors are most directly related to the motivations of the person: the topic of the survey, including social desirability bias; respondent burden (e.g. length in relation to time available); survey mode (e.g. phone, web, mail, in-person, gamification, etc.); survey design (e.g. readability, aesthetics, etc.); incentives; the survey sponsor; culture; context (e.g. current political context and debates); and competition from other surveys. Other factors affecting response rates are more directly related to survey administration: the sampling frame and methods (e.g. stale sampling frames); the use and format of advance notice, initial contact, and reminders; address accuracy; language and literacy; the time frame; and data security, particularly GPS.

Reducing unit non-response typically requires more effort and resources that directly address these reasons for non-response. Common techniques include pre-notification, follow-up with respondents, incentives, response facilitators, and improving questionnaire design. Researchers may also want to consider the mode of the survey. In their poster on "Mixed-mode surveys on travel behaviour to reach different population segments" Roux, Tebar and Armoogum found that the demographics of respondents most willing to participate in an internet survey differed from those most willing to participate in a GPS-based survey.

The response mechanism is defined as "ignorable" when it could be modeled with the available characteristics (e.g. from sample base). For example, if the non-response mechanism depends on the level of income, we need to reweight the respondent sample according to the distribution of income to correct the non-response errors.

Even when all the best methods of increasing response rate are used, there may be some response bias. A common way to deal with unit non-response is to adjust the sample through weighting. The idea is to compensate the sample size reduction by modifying the weight of all respondents. Sample weighting could be used to accomplish the following objectives:

- To compensate for differential probabilities of selection among subgroups (in stratification procedures, geographical strata, age-gender)
- To reduce the effects arising from non-response
- To compensate for inadequacies in sample frame
- To bring sample data up to the dimension of study population

We can also correct total non-response by imputation procedures (for example in a household survey, by duplicating the response of a respondent household for a non respondent household). Whatever the method chosen, it is imperative that the researcher be transparent about the method, so that archived data are reused correctly over time..

Finally, workshop participants discussed the larger issue of response rates, asking the question, How important is a high response rate? We know that lower response rates can introduce a greater risk of bias in our findings. On the other hand, is the non-response telling us something important? It may not be "wrong" to have non-response, particularly if we can learn from it.

3.1.2. Item Non-response

Even when people respond to a survey, they may not respond to every question. Item response rates often differ between objective and subjective (e.g. attitudes) items. Certain items are subject to higher rates of item non-response, such as income. Researchers need to better understand item non-response. For example, how does it vary by survey mode? by demographics? Identifying item non-response can be challenging. For example, is a missing trip a result of item non-response or measurement error? In their workshop presentation, Kagerbauer, Weiss, Streit, and Vortisch examined differences in survey respondents' perceptions of their mode choices and their actual behavior

and found that people underestimate their car passenger mode usage to a large extent, likely because it is a passive mode.

Researchers can reduce item non-response through improving survey question design and wording, such as improving the available answer responses, explaining why the question is important, and the order in which questions appear on the survey. More research is necessary on which of these strategies work in which circumstances. Computer-assisted surveys can make responses to questions mandatory. However, this can reduce overall response rate, if respondents get frustrated with this constraint.

As with unit non-response, item non-response is inevitable and must be dealt with. Researchers must decide which items on the survey are critical for response, given the survey objectives. This may differ at the household and individual level. There are two common approaches: not including the unit response in the analysis or imputation. Imputation procedures are common methods of adjusting data sets for missing values, resulting in "clean data" (complete rectangular data matrix), thus avoiding problems raised by estimations from response sets of various sizes. One issue with imputation is the fact that non-response is not random. For example, the poster by Souche found that non-response to the income question was linked to both a lack of resources and opinions about the policy in question, cordon pricing.

3.2. Measurements errors

A measurement error is the difference between a measured value of quantity and its true value. For instance when we ask the distance of a trip very few people know this information with a high level of accuracy. There are several sources of measurement error: Social desirability bias about the behavior, e.g. over-reporting bicycle use; interviewer effect (in person vs. paper/web); desires to influence policy; memory; Imprecision (e.g. rounding time or distance); proxy responses; participation in the survey that influences behavior; and survey topics. Measurement error is rarely random. For example, Kagerbauer et al.'s presentation found that teenagers underreported bicycling, while young adults underestimated car use and overestimated bicycling and transit.

Similar to item non-response, methods to reduce measurement error have relied largely on improved question and survey design. Including prompts is a common technique. For example, Crane and co-authors, presented a workshop poster about biases that exist in survey data using scales. They suggest that using an "anchoring vignette" can reduce scale bias due to age, sex, education, and income differences, thus improving the comparability of self-reported measures.

One particularly important source of error is people who report or record (via GPS) no trips on the assigned travel day. It is often not clear whether this is true or the result of measurement error.

4. Data Archiving & Confidentiality

The development surveys using GPS, Global System for Mobile Communications (GSM), and WiFi have such high spatial resolution that we must make additional efforts to anonymize the answers to preserve privacy. There are two particular reasons for this. First, it is important for survey respondents, as it shows interviewees the care we take to their answers and can give them confidence to respond again. Second, anonymization allows for greater sharing and use of archived data, which serves diverse stakeholder interests.

4.1 Data Privacy

Anonymizing data may allow greater dissemination of data and therefore its use. Getting a secure data center may benefit to data quality, indeed it also generated useful lessons in areas such as GPS data handling, processing, and user support. Gonder, Burton and Murakami's workshop presentation explained how the U.S. federal government established the Transportation Secure Data Center, which provides access to GPS travel survey data to researchers. The poster by Gehrke and Clifton presented a possible method for anonymizing GPS data through geographic perturbation, balancing the trade-off between disclosure risk and data utility.

4.2 Archiving

In the past, a survey was archived at the end of the survey process, after the analysis was complete. Now we archive surveys from the beginning and continue to enrich the information (e.g. by adding additional variables) to allow other analyses later. When we use old surveys, it is necessary to know clearly all phases of the survey including sampling (whether any subpopulation was surveyed or not) and the weighing adjustment (whether the sample is representative of what population). It is also important to document the imputation procedures that have been used and the variables that have been corrected (the imputed values may be invalid for some of the analysis and we need to change the imputation method).

When documentation is done well, it makes it easier to evaluate if surveys are comparable or not (on transversal or longitudinal point of view). Christensen and co-authors presented a poster on the challenges of harmonising archived national travel survey data to allow comparisons across countries. They found that if we pay close attention to issues such as differences in survey instruments, survey periods, reporting days, periods for trip data collection, and coverage of long-distance trips, we can make comparisons that reveal differences in behavior and not survey methods. However, they recommend that the survey design anticipate the post-harmonisation process (e.g. by not aggregating categories).

5. Conclusion and research priorities

During the workshop we discussed issues that can have an impact on transport survey data quality. One potential limitation of a sampling frame is that it may provide only a partially complete list of all eligible sampling units and may thus require augmentation with additional sources. Regarding response rates, we need to facilitate the participation more (from simple things like sending a pre-notification letter, to more complex such as using multiple media to collect data) and motivate people to participate with incentives. A surprising way to increase the response rate is to improve the respondents' confidence. This could be done by using a Secure Data Center that would anonymize, document and archive the data. Such a center also facilitates the use of the data over time. At the end of the workshop, we suggested a research agenda covering three main topics:

1. Better understand why people do not respond to mobility surveys. How should we motivate people better to respond? How much do cultural differences explain differences in response rates? Can surveys be designed to match culture better to increase response rate?
2. Appreciate different types of bias that occurs. Is the sample really representative? Bias of non-response? It might match the census, but does it match behaviour? How to determine true non-trip makers from others? What are the reasons for reporting no trips (soft refusals, not understanding the definition of a trip, poor survey design, etc.)?
3. Understand the benefits of using data more innovatively. Topics in this area include: designing and archiving surveys using new devices, for a longer time over broader users, with automatic tools for analysis; data fusion; how to take advantage of passive data and combine these with survey data (see workshop B2: System based passive data streams systems: smart cards, phone data, GPS); and surveying/data collection in developing countries.

Acknowledgements

We wish to thanks the workshop participants for in-depth discussion on the topics of survey methodologies. The People who attended the workshop are: Michiel Bliemer, Linda Christensen, Regine Gerike, Jeff Gonder, Martin Kagerbauer, Doina Olaru, Olga Petrik, Stephen Roddis, Toky Randrianasolo, Stéphanie Souche, Tatjana Streit, Maria Tebar, Li Ming Wen, Rico Wittwer, Jennifer Dill and Jimmy Armoogum.

References

Badoe, D. A., and G. N. Steuart. Impact of Interviewing by Proxy in Travel Survey Conducted by Telephone. Journal of Advanced Transportation, Vol. 36, No. 1, 2002, pp. 43–62.

Workshop papers & posters

Linda Christensen: A method to join data from a National Travel Survey of individuals into travel behaviour of households – with the driving pattern of the cars as an example.

Steven Gehrke and Kelly Clifton: A Conceptual Framework and Proof of Concept for the Geographic Perturbation of Household Travel Survey Data.

Jeffrey Gonder, Evan Burton and Elaine Murakami: Archiving Data from New Survey Technologies: Lessons Learned on Enabling Research with High-Precision Data while Preserving Participant Privacy.

Martin Kagerbauer, Christine Weiss and Tatjana Streit: Do people really act the way they think? – Differences between perceptions and reality in mode choice behaviour.

Karen Lucas: Merseyside Local Area Travel Poverty Survey.

Edith Madsen: Measurement errors in discrete choice models.

Frank Milthorpe: Forty Years of Household Travel Surveys in Sydney.

Catherine Morency and Hubert Verreault: The Chronic Issue of Proxy Respondent Bias.

Doina Olaru, Brett Smith and Fakhra Jabeen: Combining samples to offset nonresponse and respondent biases.

Toky Randrianasolo, Maria Tebar and Jimmy Armoogum: Optimal choice of auxiliary variable to reweight a mobility survey.

Chris Rissel, Melanie Crane, Stephen Greaves, Chris Standen, Li Ming Wen, Klaus Gebel and Ding Ding: Using anchoring vignettes to correct perceptions of cycling safety and quality of life: findings from the Sydney Travel and Health Study.

Sophie Roux, Maria Tebar and Jimmy Armoogum: Mixed-mode surveys on travel behaviour to reach different population segments.

Stéphanie Souche: Why do respondents give non responses to the income question? Censored models helpful for reducing the bias of non responses in transport survey.

Rico Wittwer and Stefan Hubrich: Nonresponse in household surveys: A survey of non-respondents from the repeated cross-sectional study “Mobility in Cities – SrV” in Germany.