Stages of statistical strategy in Mobility Household Surveys. Challenges addressed

Alicia María Picco
Institute of Transport Research, National University of Rosario – Argentina.
aliciapicco@gmail.com – iet@fceia.unr.edu.ar
Clyde Charre
Institute of Transport Research, National University of Rosario – Argentina.
cytrabu@ciudad.com.ar
María Florencia Álvarez Picco*
Institute of Transport Research, National University of Rosario – Argentina.
mflorencia.alvarezpicco@gmail.com

All research on the issue of public and urban transport requires the design of simulation models, based on knowledge of the situation and on the forecast of its evolution from considering different scenarios. For the development of transport models it is necessary to reach a broad knowledge of the supply and demand system. Demand models, require a particular instance of gathering information about a set of variables associated with travel and socioeconomic characteristics of the users of the system, for which purpose, it has to be done survey using different statistical techniques among which are: household surveys, censuses and surveys of travel origin and destination in points of interception and opinion surveys. Surveys Consist of field operatives supported by different sampling designs depending on the type of survey used. Like any sampling process, requires evaluation to exercise control over the errors or biases admitted at various stages of production of information. This work focuses its development on household surveys, seeking to provide a brief description of the difficulties encountered and the solutions adopted at the various stages during the performance of this type of surveys applied to the mobility issue, highlighting the involvement of statistics throughout the process. In this sense, the approach to research through household surveys by sampling involves developing what is called "statistical design strategy" to ensure that the information obtained is capable of reproducing the characteristics of the area under study. To achieve this objective the activities to develop are disaggregated and worked in two groups: a first group "Tasks Before and during the survey" and a second "Post Survey". Among the tasks Before Survey are worked out in detail, the sampling frame, the sample design and the choice of the basic variables requires to obtain a demand model, as well as the design of the relational database.
Among the tasks Post Survey are prioritized the application of tools to maintain control, consistency and integrity of data, expansion and calibration.

Key Words: transport, sampling frame, no response, based design

1. Introduction

At the moment of managing and making decisions related to urban transport and mobility of people, reliable information is needed to serve as argumentative support in the implementation of the policy or decision.

In the collection of data in transport and mobility studies, the researcher seeks to achieve the objectives through a collision of qualitative and quantitative techniques and further analysis combining the results. In this way it seeks to increase the efficiency in the analysis and achieve a holistic treatment of the problem.

There are a variety of quantitative techniques to collect information related to this methodological perspective, including: household surveys, vehicle counts and composition in private and public transport, supplemented by surveys of origin / destination points interception travel and opinion surveys.

This paper addresses quantitative research, using survey technique household survey. Household surveys are traditionally used as a technique for gathering relevant information on transportation planning, analysis of its demand and supply, as well as to study the perceptions of users of the services.

With this tool, individuals are contacted at home, from a questionnaire formulated in a relevant and consistent way. Although, depending on the chosen study design, it is possible to complete surveys through self-filling, (the full interview is done in person), it is advisable that a pollster do the questioning. Consequently, for this study, it was decided to use personal interview surveys.

Among the advantages of household surveys, and particularly the personal interview, is the possibility of obtaining high response rates because, if necessary, you can ask several times and explain the questions until they are understood.

Also, depending on the objectives of the research, it is possible to obtain spontaneous responses, as well as greater flexibility in them. Finally, we highlight the advantage that the interviewer can maintain the interest of the respondent during the process (Ibeas Portilla: 2007). In particular, in mobility surveys where the main objective is to understand the behavior of people in relation to their displacement on a business day type, often all trips made can only be declared by each household member.

However it must be explained the potential disadvantages of this procedure. It highlights the high costs involved in doing household surveys, which require more effort and development of the field work, to interrogate all household members.

With this presentation we intend to provide a brief description of the experiences gained in conducting household surveys mobility in the metropolitan area of Buenos Aires and major cities around the country, highlighting the role of statistics throughout the process, so we can talk about "statistical design strategy"

2. Statistical Design Strategy

Sample surveys are a very useful tool for all the sciences which want to obtain information in an efficient way, in a fair time and controllable cost.

The approach of a research on mobility, through household sample surveys involves developing a strategy to ensure that the information: 1) is updated at the time of the survey, and 2) be able to reproduce the characteristics of the area under study and the analysis units.

To achieve these two objectives it should be considered the following aspects of statistical strategy: a) Target population and units of analysis, b) Sampling Frame, c) type design and sample size, d) selection probabilities, e) Selection of the sample, f) errors control no associative to sampling, g) Form Preparation , h) Structure of the database, i) Validation and consistency of the data base, j) Preparation and subsequent analysis of the pilot test, k) Preparation of final fieldwork, l) Expansion of data, m) Load, processing and reporting.

These issues can be grouped into: a) Tasks Before and during the survey b) Tasks after the survey.

2.1. Tasks Before and during the survey

Identifying the Target Population and obtaining the Sampling Frame are critical to achieve the two objectives. In household surveys of Origin / Destination travel is essential, not only to define the boundaries of the land where the survey will be performed, in other words, identify the domain of study, but also, and very precisely, the geographic area or related area to the trips, called plane area. Both definitions are the guidelines to guide the design of the sample.

In household surveys on issues such as sampling frame, stratification and allocation of probabilities, except in years immediately after a population census, the information available is old.

Consequently, for the definition of the sample, it is necessary first to have an updated sampling frame. This implies a previous analysis to achieve the maximum possible update, necessary both for territorial coverage and also for the probabilities adjudication and obtaining features that serve to stratification.

To update the Framework there is available geographic information on the Internet, which facilitates the researcher task. Once, defined the study area and nestled in a Geographic Information System (GIS) should be validated with field up survey. This coverage involves identifying and analyzing, within the land area defined for the study, areas of high population density, unpopulated areas or with very small population.

The land visualization ensures the quality of the first stage of the design, since it shows the characteristics of the whole area (residential, commercial, industrial and recreational) and improves the stratification criteria and probabilities assignment.

To satisfy the requirements in mobility studies and, as a result, achieve a matrix of trips between zone pairs connected by land uses is desirable to use a sample design probabilistic, stratified and two-stage type.

Probabilistic, to establish desired precision in the main results, calculate the accuracy observed in the results, and thus, gives the decision maker a measure of the interval in which are estimates the parameters of interest, with the known confidence level.

Stratified, to reduce the estimated variance and thus obtain better reliability of the sample. Better stratification input, the more related to the problem of mobility are the variables used to define the strata. The main variables are related to the offer and demand of current passenger transport, mainly: number of lines and population density

Two-stage, as it lets you use census divisions, in this case radios with closest census data that provide information both for probabilities assignment as for stratification. These divisions are called First Stage Units (LEU). In each UPE that will integrate the sample, households are selected, called second-stage units (USE)

The UPE are selected with probability proportional to size, measured in number of houses by systematic selection method independently in each stratum.

The units of second and final stage (USE) are selected systematically on a list made in the field, with equal probability for all homes. The list of houses that is performed on each selected UPE is a new update, this time of the secondary frame.

The determination of the number of households to be selected in each UPE varies on the stratum, trying to have more UPE with clusters with fewer households in stratum widely scattered and fewer UPE with more households in strata less dispersed geographically. It also takes into account for this determination, have more sample in those strata where more no response is expected.

Statistical participation in areas falling under the thematic-conceptual area of sample surveys is less common and comprehensive. It was necessary to show how the relationship between the questionnaire, the database and the final results requires a statistical methodological view.

As an example we can mention the need of differencing the identification of the different units, the statistics (housing) of the analysis (household, people, travel) and that identification must be reflected in the various forms that constitute the questionnaire used, as well as in the related databases that integrate the structure of the database. This differentiated identification is necessary to calculate the expansion factors.

Validation and data consistency must be done with automatic controls system. The system must respect the jumps on the form prescribed and allowable values for each field on the form and the relationships between fields that enable the consistency of the response.

It must be defined control criteria and statistical validation to be incorporated in the structure of the database, so as to allow them to reproduce, early, the behavior of structural variables and the substantive. These partial quantifications allow improvements in the way of releasing the information while survey is still in the field, asses in advance the processes that will be required for the expansion and size allocation needs.

Statistical participation during the preparation and implementation of the Pilot Test is essential for the further development of the final survey. Quantification of difficulties and successes that statistical tools allow to apply not only in purely quantitative issues, but also on the conceptual and field application themes, allow systematic evaluation easily capitalized to make the adjustments to improve both aspects of the sample, the questionnaire, training, and the structure of the personnel which will perform the operation in field.

During field work is necessary to introduce, in the middle of the necessarily long training on the topic of the survey, guidelines that allow to control, avoid, and if possible quantify different types of bias.

The training courses by experts should provide to the supervisors and surveyors, awareness of the main objective of the survey, a good understanding of the survey form and its fill up and relational aspects between variables in order to gain greater accuracy and thoroughness of the field data.

It has emerged the need to raise awareness, both the supervisor and the surveyor general guidelines of the sample design. The warning to the surveyor about the implications of a

change of boundaries in a sampling unit in the final sample encourages him to be aware that has responsibility also in the activity of the sample and not just filling the questionnaire.

It also happens that the surveyor changes the selected house for one with a home with fewer members, or that they are able to contact without repeated visits on different days and times introducing biases that affect population structure.

An issue that is growing is **the lack of response**. It is known that many strategies have been developed to solve the issue of absence, but remains the problem of rejection requiring ad hoc measures. Despite efforts over the lifting of the surveys, the **lack of** response reaches the final data where it is necessary to make corrections in the expansions.

There are many techniques and methods to address the non-response in the expansion, but it is important to have in mind that the data we have, in other words those arising from survey respondents, are not distributed in the same way as those of a non responding, therefore the data used for the calculations has biases. It is an important task introduce from the sample design and in the following stages, particularly in the field, actions that allow identifying these biases and if possible quantifying them in each of the stages.

2.2. Tasks after the survey

There is a growing consensus on the need that the data load it is as close as possible in time to the data lift, to the point that it is enforcing the use of electronic data capture instead of paper.

Consequently, the main advantage lies in the consistency rules that are introduced during charging, thereby significantly improving the final quality of the data.

Relational databases go through a process which is known as normalizing a database, which is understood as the process required for a database to be utilized optimally.
A relational database:
- Ensures tools to avoid duplication of registers by key or key fields.
-Guaranteed referential integrity: thus in the moment of deleting a register deletes all related registers dependents.
-Promotes normalization for being more understandable and applicable.

In a consisted and validated base (clean), it joins the expansion factor that adjusts the information to the target population. It is then able to analyze each variable using all the statistical techniques that report on precision measurements.

As in this type of research sample design is complex: two-stage stratified cluster, where UPE are selected with variables probabilities and USE with equal probability, for expanding the values of the sample to the population is considered **Expansion Factor** for each UPE which consists of the product of the factors calculated for each design stage and by a correction factor for non-response:

The probability of selection of each primary unit is the corresponding number of houses divided by the total residential of the stratum.

Thus the expansion factor first stage is:

F_1 (hi) = Nh / (Nhi * mh)
And the expansion factor of second stage is: F_2 (hi) = Nhi / ñ
Where:
Nhi = dwellings by count in the ith stratum h UPE
ñ = fixed cluster dwellings selected in each UPE

The correction factor for non-response (FNR) allow adjusting the simple expansion factor considering rejections and homes with permanent residents, circumstantial or temporarily absent.

At the time of counting, it is not possible to know with certainty whether or not a household lives in each statistical or selection unit. In case of any dwelling be selected with that condition, occurs a lack of response due to what is commonly called "framing effect". For the purposes of expansion, it should be excluded from any coefficient to correct non-response, as their inclusion would incur in the error of overestimating the population.

Given the above characteristics, it is called "effective" to any unit of selection (housing) containing a unit of analysis, a home, that could be response or no response for absent or rejection.

Once expanded the data it is able to perform an analysis of the basic structures to see their behavior, for this, comparisons of structural variables are made with an external source, in particular with the results of the Household Survey to assess the need to calibrate the survey data. The variables that best reproduce the behavior of the population with respect to mobility are: sex, age group, number of household members and their relationship with the Principal Household member.

The proposed methodology allows users to characterize each of the modes of transport, identify demand habits, particularly the reason, schedules and journey times, system coverage, perception of rates and quality of each mode, and reproduce the sociodemographic characteristics of the study area.

3. Conclusion

Although, it may have difficulty performing household surveys mobility, Tasks performed before, during and after the survey in a responsible way and with statistical view at all stages ensures that the information obtained is capable of reproducing the characteristics of the area under study and the units of analysis.

With this statistical participation at all stages of the Strategy assure a homogeneous view of the survey process and the delivery to the user of a data set with not only precision measurements but also with a description of the biases, information that will allow making decisions with greater confidence.

References

-Hansen M.H; Hurwitz W. N; Madow W. G. (1953), Sample Survey Methods and Theory. John Wiley. New York -London –Sydney

- Ibeas Portilla, A (2007), Mobility survey manual (Preferences Revealed). University of Cantabria. Spain

-Kish L. (1965), Survey Sampling. John Wiley. New York - London – Sydney

- Ortuzar,J.de D. y Willumsen L.(2006); Modelling Transport, 3° Edicion. Wiley, UK.