Annex 11
Dealing with Unknown Reference Populations in Border Surveys on Inbound Tourism [1]

1. Introduction

In parallel with the development of the travel industry, a growing interest in statistical tools for its measurement and characterisation is emerging. Following the approval of an international standard for the Tourism Satellite Account, significant emphasis has been put on the analysis of the economic impact of tourism on national economies. This implies a focus on the collection of information on the consumption of visitors, largely corresponding to their expenditure. Inbound border surveys are frequently used to collect information on the activity of non-resident visitors in the reference country. Since 1996, the Ufficio Italiano dei Cambi (UIC) has been carrying out an extensive inbound-outbound border survey on Italy's international tourism. The survey is run on a continuous basis through a representative sample of around 130,000 face-to-face interviews per year, allowing the observation of several qualitative and quantitative attributes (Ufficio Italiano dei Cambi, 1997). The data from this survey serve both the compilation of the Travel item of the balance of payments and the satisfaction of the information needs of tourism operators and analysts. The paper focuses on the consequences of the lack of knowledge on the reference population, a typical problem in tourism statistics but not yet adequately investigated. The solution adopted by the UIC in order to operationally tackle this issue is illustrated; counting operations are performed (> 1,000,000 per year), in order to determine the number and the nationality of cross-border visitors. An approach for the measurement of the additional sampling errors due to the procedure adopted in Italy is described. For the sake of briefness, two simplifications have been adopted. First, only inbound tourism has been considered, since the survey methodology is similar for the outbound side. Second, the survey was considered completely "independent" from external sources, even though, as explained subsequently, the UIC has recently started the collection of data from some administrative sources, in order to mitigate the mentioned effect of the uncertainty about the "true" size of the population.

2. Unknown populations and sampling errors

The reference (or target) population is the group of statistical units representing the real target of the survey. It should be defined univocally as it is closely connected with the main objective of the research. The survey population is defined by the statistician in the sample design. It may involve slight deviations from the reference population, justified by practical reasons, which should not alter significantly the final results. The sampling frame is a physical list in which the statistical units are represented and from which the observed sample is drawn. Any deviation between the frame and the reference population may be source of errors (Groves et al., 2004).

---

[1] Paper presented to the International Statistical Institute, 55th Session 2005. Sidney, Austria. The views expressed in this paper are those of the authors and do not involve the responsibility of the Ufficio Italiano dei Cambi. Although the authors share the overall responsibility for the paper, the three paragraphs have been written respectively by A. Alivernini, L. Buldorini and G. G. Ortolani., Ufficio Italiano dei Cambi, Italy.

The reference population of the UIC international tourism inbound survey is defined as the population of foreign (non resident) travellers that visited Italy during a certain period of time (month), as defined in the Balance of Payments Manual of the International Monetary Fund. The survey population is defined as the population of foreign travellers that crossed the Italian borders under observation during the reference period. This definition introduces two differences with respect to the reference population. The first stems from the difference between "travellers that visited Italy" and "travellers that crossed the Italian borders", which may cause differences in the time allocation of the expenditure. The second is due to the fact that the observed border points are a subset of the total Italian border points.

One of the most problematic aspects of the Italian inbound tourism survey is the lack of a set of consistent administrative records of inbound tourism flows with adequate coverage, timeliness and detail. Despite some recent progress that will be mentioned subsequently, it can be therefore assumed that a sampling frame is not available. The absence of a frame means that an essential element for the estimate of the total expenditure, the dimension of the total population of foreign travellers, is unknown. In order to estimate it there is the need to carry out counting activities, which represent, as a matter of fact, an additional survey nested in the main survey (the one measuring the expenditure of the travellers).

Counting activities are significantly differentiated in the various types of border considered (road, train, air and sea borders). For this reason it will be assumed that four different populations need to be estimated. It is assumed that the total population is given by

$$N = \sum_{v=1}^{4} N_v = \sum_{v=1}^{4} \sum_{p=1}^{P_v} N_{vp} \quad , \tag{1}$$

where *p* is the single border point and *v* can assume the values 1, 2, 3 or 4 representing road, train, air and sea borders respectively.

On road borders the counting of travellers is based on the observation of the flow of vehicles crossing the borders. The design of the counting activity (survey) rely on a sampling of the time, where the month is the "population", the time units (i.e. hours) are the sampling units and the number of foreign travellers crossing the border during the time unit is the observed characteristic. The time units should be selected in order to ensure the representativeness for all time periods, within the day and within the week. Once the time units are chosen, an actual observation of the traffic in those periods is carried out. The estimate of the total number of travellers N_{1p} of the road border point p is therefore given by

$$\hat{N}_{1p} = T \frac{\sum_{i=1}^{t} n_{1pi}}{t} ,$$

where *T* is the total number of time units in the month (i.e. 30 days or 720 hours), *t* is the number of actually observed time units and n_{1pi} is the number of passengers observed, during the time unit *i*, at the border point *p*, belonging to the type 1 (road).

The estimate of N_{1p} is subject to an error due to sampling variability given by the expression.

$$Var(\hat{N}_{1p}) = T^2 \frac{Var(n_{1pi})}{t}.$$

Therefore, the error depends on the variability of the number of passengers observed in the time unit and on the sample size *t*.

On rail, air and sea borders, counting activities are helped by information on the time schedule of transport vectors, known in advance. The schedule is a list of trains, flights or ships to or from international destinations that serves as a sampling frame from which a sample is selected. Under the hypothesis that *NT* is the total number of scheduled transport vectors in the reference month, *nt* the number of observed transport vectors and n_k the number of travellers on board of transport vector *k*, the estimate of the total population for the border point *p* of the border type *v* is given by

$$\hat{N}_{vp} = NT \frac{\sum_{k=1}^{nt} n_{vpk}}{nt} \text{ , } v= 2,3,4.$$

Also in this case, the estimate of N_{vp} is affected by an error due to sampling variability given by the expression

$$Var(\hat{N}_{vp}) = NT^2 \frac{Var(n_{vpk})}{nt},$$

which depends on the variability of the number of travellers on board and on the sampling size.

It should be noted that, although the formulas presented above refer to simple random sampling design, the actual sampling design applied is often stratified, in order to reduce sampling variability. In the case of time sampling (road borders), useful strata are deemed to be different time periods within the day (i.e. day and night, morning and evening) or within the week (i.e. weekdays and holidays). For other type of borders (train, air or sea) transport means are often grouped in strata of homogeneous destinations.

In expression (1) it has been shown that *N* is given by the total number of travellers crossing each border point. Under the hypothesis that the sampling variability in each border is independent, the total sampling variability of the total number of travellers *N* is

$$Var(\hat{N}) = \sum_{v=1}^{4} \sum_{p=1}^{P_v} Var(\hat{N}_{vp})$$

The most common error measure in sampling surveys is the sampling error, normally expressed by the mean square error of the estimate. In the case of the estimate of the total of a quantitative character *x* (that in our case is the individual

expenditure of the foreign traveller), in a population of *N* elements, using a simple random sample of *n* elements and the estimator

$$\hat{X} = N \frac{\sum_{i=1}^{n} x_i}{n} = N\mu_x \,, \quad (2)$$

the sampling error is the square root of the estimator variance

$$MSE(\hat{X}) = \sqrt{N^2 \frac{\sigma_x^2}{n}} = N \frac{\sigma_x}{\sqrt{n}} \,. \quad (3)$$

As it has been shown in the previous paragraphs, in the UIC survey the total number of passengers, *N*, is estimated through the counting activities. The value of *N* shown in expression (2) cannot be considered, in our case, a constant but an estimate. This will unavoidably contribute to an increase of the sampling error expressed in (3).

Expression (2) can be re-written as

$$\hat{X} = \hat{N} \frac{\sum_{i=1}^{n} x_i}{n} = \hat{N} \bullet \bar{x}$$

where the symbol $\bar{x}$ indicates the sample average of individual expenditure and $\hat{N}$ is the estimate of N (no longer a constant). The new formulation highlights that the estimator of the total traveller expenditure, $\hat{X}$, is a product of two random variables, i.e. the number of travellers, $\hat{N}$, and the individual average expenditure, $\bar{x}$. Under the assumption that the two latter variables are independent, the variability of the total expenditure is given by the variance of the product of two independent variables

$$Var(\hat{X}) = \hat{N}^2 \bullet Var(\bar{x}) + \bar{x}^2 \bullet Var(\hat{N}) + Var(\bar{x}) \bullet Var(\hat{N}) \,.$$

It can be noticed that the square root of the first term of previous expression would be identical to the sampling error shown in expression (3). Consequently, the other two terms indicates the sampling error due to the variability of N, namely the variability due to the absence of a frame.

An additional problem caused by the absence of a frame is the bigger difficulty in controlling the selection of sampling units. The frame plays in fact a crucial role during the selection of the sample, allowing the correct calculation of selection probability for each sampling unit and therefore the definition of a non biased estimator for the target variable. In the UIC survey, the selection of sampling units (travellers) is necessarily made on the field, instructing interviewers to follow standard "approach rules" in order to help a random selection. Although "approach rules" are carefully codified and interviewers are duly instructed, the actual selection process is always partially out of control of the design planner, due to difficult logistic conditions under which the approach is realised. On road borders, for instance, interviewers have to stop the vehicle and submit the questionnaire, literally, "on the road", is easy to imagine how an actual random selection might be basically

impossible. A bias of the sample (auto selection) could be caused, for example, by the practical difficulties to stop heavy commercial vehicles. Drivers of such vehicles could therefore be under represented in the sample and, as their average expenditure is normally lower than tourists, total expenditure might result over estimated. In order to control the bias produced by auto selections of the sample, stratification techniques can be applied even after the selection has been made. In the previous example one might use data derived from counting activities in order to distinguish travellers by type of vehicle. The separate imputation of the average expenditure according to the type of vehicle would mitigate the bias effect on the final results.

3. Operational implications

The discussion above leads to two main operational implications. On the first hand, this work deals with some aspects of the multidimensional concept of data quality. The study stresses the importance of a measurement of sampling errors that takes into account the complexity of the survey design involved by the existence of a survey on physical flows "nested" in the main survey on expenditure. A more precise assessment of the error level involves positive outcomes, such as a more correct feedback to survey managers on the consistency of the survey process and an improved communication to the users of the reliability of the statistics produced.

On the second hand, a "statistical policy" issue emerges. The paper indicates the need to rethink collection strategies on tourism, by assigning a greater role to administrative records on cross-border visitor flows. This paper has shown that the knowledge of the size and basic characteristics of the reference population greatly helps the production of statistical information on inbound expenditure. As illustrated above for the Italian case, the lack of that knowledge implies extra-costs and a certain loss of accuracy of the output (increase of sampling error), because of the need to set up specific additional surveys.

Consequently, administrative records on physical inbound flows (outcome of controls at borders, databases of airport and road authorities, etc.) represent precious sources. Positive results have been recently achieved in the Italian survey, through the use of external administrative sources as a supplement to counting operations for some border points. Although the administrative records may require some adaptation, in order to correctly address the reference population (visitors) and meet the timeliness and detail requirements of tourism statistics, they are usually accurate and relatively inexpensive. The co-operation between statistical agencies and transport authorities should be therefore strengthened, both at the national and international level, with the objective of maximising the usability of the information originated as a by-product of administrative processes.

The World Tourism Organisation (Canadian Tourism Commission et al., 2004) has recently promoted an inventory of the practices of various countries in the collection of data on physical tourism flows at borders. The study shows that many countries, especially outside the European Union, can still rely on administrative records, e.g. in the form of entry / departure cards that visitors are requested to fill in while crossing the borders. The progress of the liberalisation of cross-border movements of persons pushes towards the reduction of such "formalities", which,

nonetheless, as explained in this paper, provide a relevant contribution to tourism statistics. The dismantling of this type of administrative sources should, therefore, be considered on the basis of a comprehensive cost-benefit analysis.

REFERENCES

Canadian Tourism Commission, Instituto de Estudios Turisticos, Swedish Tourist Authority, World Tourism Organization (2004). Comparative Study of International Experiences in the Measurement of Traveller Flows at National Borders, WTO, Madrid.

R. M. Groves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, R. Tourangeau (2004). Survey Methodology, Wiley, Hoboken.

Ufficio Italiano dei Cambi (1997). Methodology for the Elaboration of Statistics on Tourist Movements at Land Borders, UIC, Roma.