Instituto de Estudios Turísticos
Secretaría General de Turismo
Estudios Turísticos, n.º 161-162 (2004), pp. 119-127 Secretaría de Estado de Turismo y Comercio

A NEW SURVEY METHODOLOGY FOR DESCRIBING TOURISM ACTIVITIES AND EXPENSES

Jean-Claude Deville* and Myriam Maumy**

I. INTRODUCTION

A "border survey", concerning the touristic frequentation in Brittany (excluding Britain people) has been competed for the period between April and September 1997. The "Observatoire Régional du Tourisme de Bretagne" and the "Comités D epartementaux de Tourisme" would like to launch another survey of the same type for the next years. Unfortunately they have no more the opportunity to get a lot of information collected at the regional and intra-regional frontiers, since the gendarmerie can no more help on the realisation of interviews on the border of roads.

That's why the "Observatoire Régional du Tourisme de Bretagne" with the help of a technical commitee constituted by scientists and experts of Britanny and of the "Système d'Information Touristique des Asturies de l'Université d'Oviedo (Espagne)" has decided to set a new methodology which should replace the former "border survey".

One of the main problems is the lack of survey base which should be used to interview tourists directly. The main idea of the work-around is to sample services targetting tourists and to investigate on some part of these different places. Obviously, one tourist can use one or many times on or many services of the survey base during the period of the survey. In order to estimate the parameters of interest relating to tourists, we must bind the set of weights of sampled services to the set of weights of tourists who have used these services.

The goal of this article is to present a method which can evaluate these parameters. This method is mainly based on the Generalised Weight Share Method (GWSM) set by Lavallée (1995, 2002).

II. THE GENERALISED WEIGHT SHARE METHOD

In this section, we recall the foundations underlying of the Generalised Weight Share Method (GWSM). For more details, we refer to Lavallée (2002) and Deville (1999).

To select the samples needed for social or economic surveys, it is useful to have sampling frames, i.e., lists of units intended to provide a way to reach desired target

* Laboratoire de Statistique d'Enquête.
** Laboratoire de Statistique de l'Université de Rennes 2.

populations. Unfortunately, it happens that one does not have a list containing the desired collection units, but rather another list of units linked in a certain way to the list of collection units. One can speak therefore of two populations U^A and U^B linked to each other, where one wants to produce an estimate for U^B. Unfortunately, a sampling frame is only available for U^A. It can then be considered to select a sample s^A from U^A in order to produce an estimate for U^B by using the correspondence existing between the two populations. This can be designed by *Indirect Sampling*.

Let the population U^A contain N^A units, where each unit is labelled by the letter j. Similarly, let the target population U^B contain N^B units, where each unit is labelled by the letter i. The correspondence between the two populations U^A and U^B can be represented by a *link matrix* $\Theta_{AB} = [\theta_{ji}^{AB}]$, of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. That is, unit j of U^A is related to unit i of U^B provided that $\theta_{ji}^{AB} > 0$; otherwise the two units are not related to each other.

With Indirect Sampling, we select the sample s^A of n^A units from U^A using some sampling design. Let π_j^A be the selection probability of unit j. We assume $\pi_j^A > 0$ for all j $\in U^A$. For each unit j selected in s^A, we identify the units i of U^B that have a non-zero correspondance, i.e. with $\theta_{ji}^{AB} > 0$. Let s^B be the set of the n^B units of U^B identified by the units $j \in s^A$, i.e. $s^B = \{i \in U^B;\ \exists j \in s^A$ et $\theta_{ji}^{AB} > 0\}$. For each unit i of the set s^B, we measure a variable of interest y_i from the target population U^B. Let $\mathbf{Y} = \{y_1, ..., y_{N^B}\}'$ be the column vector of that variable of interest.

We assume that for any unit j of s^A, the values of θ_{ji}^{AB} for i = 1, ..., N^B can be obtained. That is, we can collect all the values of θ_{ji}^{AB} by direct interview or by some administrative source for any sampled unit j. Also, for any identified unit i of U^B, we assume that the values of θ_{ji}^{AB} pour $j = 1, ..., N^A$ can be obtained. Therefore, the values of θ_{ji}^{AB} need not to be known for the entire link matrix Θ_{AB}. We need in fact to known the values of θ_{ji}^{AB} only for the lines j of Θ_{AB}, where $j \in s^A$, and also for columns i of Θ_{AB} where $i \in s^B$.

Suppose that we are interested in estimating the total T^B of the target population U^B, where

$$T^B = \sum_{i=1}^{N_B} y_i,$$

where the values of y_i are measured from the target population U^B. Now let $\theta_{+i}^{AB} = \sum_{j=1}^{N^A} \theta_{ji}^{AB}$ and let $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \theta_{+i}^{AB}$.

For estimating T^B, we want to use the values of y_i measured from set s^B. For this, we will use an estimator of the form

$$\hat{T}^B = \sum_{i=1}^{N^B} w_i y_i,$$

where w_i is the estimation weight of the unit i of s^B, avec $w_i = 0$ for $i \notin s^B$. Usually, to get an unbiased estimate of T^B, one can simply use as the weight the inverse of the selection probability π_i^B of unit i. As mentioned by Lavallée (1995) and Lavallée (2002), with Indirect Sampling, this probability can however be difficult, or even impossible, to obtain. It is then proposed to use the GWSM, which is defined as follows.

Starting from

$$T^B = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$$

we can directly form the following Horvitz-Thompson estimator:

$$\hat{T}^B = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \frac{t_j \tilde{\theta}_{ji}^{AB}}{\pi_j^A} y_i$$

The vector **W** is of size N^B and for each $i = 1, ..., N^B$, we have

$$w_i = \sum_{j=1}^{N^A} \frac{t_j \tilde{\theta}_{ji}^{AB}}{\pi_j^A}$$

The weights w_i of that vector are said to be obtained from the GWSM, as described by Lavallée (2002).

III. OPEN AREA SURVEY: SOME PRINCIPLES

The main principle of the survey consists in: *"to trap tourists (French and foreign people) thanks to services targetting their today needs"* like accomodation, food, leasure activities, and transport.

The statistical unit is the travel which is defined by the group of people who are having the trip together and that have a similar behaviour in the main variables of interest. Thus, a tourist may be having his trip with one or more companions, where n is the variable quantifies the number of people who are part of his traveling party and who are extensible to the answers given by the interviewee. We will use, for more practice, in this paper, the expression *Traveling party* to design the group of people who are travelling together.

The interviewee is the person who manage the expenses of the group of people who are having the trip together.

The survey plan should respect the following principles.

The periods of the survey are the moments when the touristic activity has great variations througout the year. We have defined **three main periods throughout the year:**

- july and august 2005 (tourist season);
- april, may, june and september 2005;
- School holidays in december 2004, in november 2005 and in february 2005.

The places where the survey will take place are:

- hotels and camp-sites (institutional accomodation),
- bakeries and pastry shops,
- 15 popular visitor places which are famous and are stated below
  - Belle Ile
  - Château de Fougères
  - Château de la Roche Jagu
  - Château de Suscinio
  - Fréhel
  - Ile de Bréhat
  - Ile aux Moines
  - Musée de télécommunications

– Océanopolis

– Pointe du Grouin

– Pointe du Raz

– Remparts de Saint-Malo

– Trévarez

– Vedettes de l'Odet

– Zoo de Branféré.

The survey base is constituted by 3 groups:

• institutional nights in hotels and/or in camp-sites;

• purchases in bakeries/pastry shops;

• crossing a popular place for the activity of the 15 visiting touristic places.

In *the first group,* we will realize a sample with three degree:

• a sample of hotels and camp-sites stratified with the usual procedures;

• a sample of days within the period of study;

• a sample of nights spent, ie tourists having spend one night or nights in the given hotel or in the camp-site given at the given day.

In *the second group,* similarly, we will realize a sample with three degree:

• a sample of bakeries/pastry shops;

• a sample of days within the period of study;

• a sample of clients in the given bakery/pastry shop at the given day.

Finally in *the third group,* we will realize a sample with two degree:

• a sample of days within the period of study;

• a sample of tourists who visit one of the 15 visiting touristic places refered at the given day.

IV. THE POPULATION OF INTEREST AND THE PARAMETER OF INTEREST

The population of interest is constitued of tourists who use at least one or more than one service of the survey during time frame.

The time frame of the survey begins in december 2004 and stops in november 2005.

The geographical area of the survey is divided into four areas which correspond to the four departments of Brittany.

Introduce the notations which we will use more later.

• soit A_1: the set of hotels of the survey labelled by the index a_1

• soit A_2: the set of camp-sites of the survey labelled by the index a_2

• soit A_3: the set of bakeries/pastry shops of the survey labelled by the index a_3

• soit A_4: the 15 visiting touristic places of the survey labelled by the index a_4

• soit D_l: the set of the survey days, labelled by the index d_l in an establishment a_l of the set A_l, for l from 1 to 4

• soit C_{d_l}: the set of the services in an establishment a_l of the set A_l of the day d_l from the set D_l labelled by the letter j.

We define an application F, which each service labelled by the letter j during the time frame labelled by the latter D in the four types of establishments of the survey, joins the traveling party using this service.

$$F : \{\text{services}\} \rightarrow \{\text{traveling party}\}$$
$$j \rightarrow F(j) = i$$

Let U^B, the population of traveling party labelled by the letter i of the time frame labelled by the letter D. This population of interest U^B is the figure by F of the set of services during the time frame D in the four sets of establishments of the survey.

For all $i \in U^B$, we define

$$R_i(B) = card(F^{-1}(i)),$$

the number of antecedents of the traveling party i during the time frame, i.e., the number of services j used by a given traveling party i.

Let now specify the word "services".

• In a hotel or in a camp-site, the service is a night.

• In a bakery or in a pastry shop, the service is a purchase in this shop.

• On the 15 visiting touristic places, the service is the visit of this touristic place by the traveling party or by a part of the traveling party.

The parameter of interest can be totals, effectives or ratios. We assume for instance, that we are interested in the estimation of a total related to a variable Y defined on the population U^B,

$$T^B = \sum_{i \in U^B} y_i.$$

For instance, T^B can be the number of people who have participated to a given activity, the total budget spent by the traveling party in Britanny, the region from where the traveling party come from, the number of days the traveling party has spent in Britanny We must note that, for a lot of variables, the total T^B depends on the size of the traveling party, i.e. the number of people who constitue the group, and the number of days spent in Britanny. From now on, we can write:

$$T^B = \sum_{i \in U^B} y_i = \sum_{l=1}^{4} \sum_{a_l \in A_l} \sum_{d_l \in D_l} \sum_{j \in C_{d_l}} z_j,$$

where

$$z_j = \frac{y_i}{R_i(B)}, \text{ for } j \in F^{-1}(i)$$

V. UNBIASED ESTIMATION OF A TOTAL

In the former section, we have shown that the total of interest can be written as a total on

a set of services of the domain. Let's assume we have a sample of services which answer j, to which we can associate some weights δ_j. These weights are assumed unbiased as we have shown in section 2.

In order to simplify the notations, we do not make appear all degrees of sample selection in function of the establishment a_l.

Let:

- s^B: the set of traveling party i which correspond to the set of services sampled during the period of survey
- s_{A_l}: the set of sampled establishments
- s_{D_l}: the set of sampling days for the establishment a_l
- s_{d_l}: the subset of services j which correspond to the day of the establishment a_l.

Having a set of weights δ_j for the services which answer, and as we know $R_i(B)$, we estimate T^B by:

$$\hat{T}^B = \sum_{i \in s^B} w_i y_i,$$

where

$$w_i = \frac{\sum_{l=1}^{4} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}$$

This estimator is unbiased. We are brought back to an estimation of the population of traveling party. This formula is the one given by the Generalised Weight Share Method cited in section 2. We remark that

$$U^A = U^{A_1} \cup U^{A_2} \cup U^{A_3} \cup U^{A_4} = \bigcup_{l=1}^{4} U^{A_l},\ \theta_{ji}^{AB} = 1$$

if the service j has been used by the traveling party i and then $\delta_j = 1/\pi_j^A$.

VI. SPECIAL CASE OF CERTAIN VISITING TOURISTIC PLACES: THE TOURISTIC POINTS IN THE OPEN COUNTRY

On some visiting touristic places, we unfortunately do not know the total number of people who visit the site. Indeed, in the set A_4, we do not know all the services (in this case the number of visitors) of the population. Then we can not have directly $\pi_j^{A_4}$ and then δ_j for $j \in A_4$. As a work-around of this problem, we estimate the daily number of visitors in order to reduce $\tilde{\pi}_j^{A_4} = n_{A_4}/\tilde{N}_{A_4}$. In fact, we will use a system which consists in counting from a strategic point the number of cars and the number of people who travel together in this car. Finally, we have to estimate this number of visitors in order to get $\tilde{N}_{A_4}$.

VI.1. Construction of an estimator of a function of interest from a sampling of cars

In this subsection, we are in the case where an investigator counts the number of people who travel together in a car, i.e., count the number of people in cars which cross the place where an electronic eye or an equivalent system has been placed in order to count the cars whose total number is known.

Let T_V be the total number of cars defined by

$$T_V = \sum_{j \in N^*} t_j, \qquad (6.1)$$

where t_j is the number of cars carrying j people. We can also define T_V by the following equality:

$$T_V = \sum_{i \in U_V} 1, \qquad (6.2)$$

where U_V represents the space of cars.

Remark 6.1. We note that we know the total number of cars T_V. Therefore, we do not need any estimator for the total number of cars T_V.

Let T_P be the total number of people who visit the site defined by

$$T_P = \sum_{j \in N^*} j t_j, \qquad (6.3)$$

By analogy, with the expression (6.2) for the total of cars T_V, we can also define the total number of people T_p by

$$T_P = \sum_{j \in U_P} 1, \qquad (6.4)$$

where U_P represents the space of people. We can also define the total number of people T_p by the following equality

$$T_P = \sum_{i \in U_V} v_i, \qquad (6.5)$$

where v_i is the number of people in the car labelled by the letter i. As we have mentionned at the beginning of this section, the total number of people T_P is unkown. Consequently, we must have an estimator of the total number of people T_P. Let $\hat{T}_P$ be the π estimator of the total number of people T_P defined by

$$\hat{T}_P = \sum_{i \in s_V} w_i v_i,$$

where s_V is a sample of cars and the weight w_i is equal to T_V/n, which allows to write the estimator $\hat{T}_P$ under the following form

$$\hat{T}_P = \frac{T_V}{n} \sum_{i \in s_V} v_i = T_V \bar{v}.$$

where $\bar{v} = 1/n \sum_{i \in s_V} v_i$ and where n refers to the size of the sample s_V.

Theorem 6.1. *$\hat{T}_P$ is an unbiased estimator of the total number of people T_P.*

Let Y be a variable of interest defined by

$$Y = \sum_{j \in U_P} y_j,$$

where y_j is a variable of interest which we can measure in the final questionnaire. Let $\hat{Y}$ be the π estimator of this variable of interest Y defined by

$$\hat{Y} = \sum_{j \in s_P} w_j y_j.$$

where the weight w_j is equal to $\hat{T}_P/m$. Consequently we can write $\hat{Y}$ under the following form

$$\hat{Y} = \frac{\hat{T}_P}{m} \sum_{j \in s_P} y_j = \hat{T}_P \bar{y},$$

where $\bar{y} = \frac{1}{m} \sum_{j \in s_P} y_i$.

VI.2. Variance of the estimator $\hat{Y}$ in the case of the sampling of cars

In order to calculate the variance of $\hat{Y}$, we will use the Huygens' theorem. As we condition $\hat{Y}$ according to the sample of cars s_V, we can establish the following theorem:

Theorem 6.2. *We have the following equality*

$$\text{Var}\,[\hat{Y}] = \bar{Y}^2\text{Var}[\hat{T}_P] + T_P^2\text{Var}[\bar{y}] + \\ + \text{Var}[\hat{T}_P]\text{Var}[\bar{y}]. \qquad (6.6)$$

Theorem 6.3. *In case we study a simple random sampling without replacement, the equality (6.6) then becomes*

$$\text{Var}\,[\hat{Y}] = \left(\bar{Y}^2 - \frac{1}{T_P}\,S_{\bar{y}}^2\right) T_V^2 S_V^2\,\frac{1}{n} + \\ + (T_P^2 - T_V S_V^2) S_{\bar{y}}^2\,\frac{1}{m} + T_V^2 S_V^2 S_{\bar{y}}^2\,\frac{1}{nm} + \\ + \frac{T_V}{T_P}\,S_V^2 S_{\bar{y}}^2 - \bar{Y}^2 T_V^2 S_V^2 - T_P S_{\bar{y}}^2 \qquad (6.7)$$

The next step consists in getting the allocation of sizes of samples s_P and s_V which minimizes the variance of the estimator $\hat{Y}$ for given sizes of population T_P and T_V.

We have to minimize

$$\text{Var}\,[\hat{Y}] = \left(\bar{Y}^2 - \frac{1}{T_P}\,S_{\bar{y}}^2\right) T_V^2 S_V^2\,\frac{1}{n} + \\ + (T_P^2 - T_V S_V^2) S_{\bar{y}}^2\,\frac{1}{m} + T_V^2 S_V^2 S_{\bar{y}}^2\,\frac{1}{nm} + \\ + \frac{T_V}{T_P}\,S_V^2 S_{\bar{y}}^2 - \bar{Y}^2 T_V^2 S_V^2 - T_P S_{\bar{y}}^2$$

of n, m under the constraint

$$C_V n + C_P m = C.$$

After some developments, we get an equation of third degree in n:

$$\lambda C_V^2 n^3 - \lambda C_V C n^2 - C_V T_V^2 S_V^2 \left(\bar{Y}^2 - \frac{1}{T_P}\,S_{\bar{y}^2}^2\right) n \\ + T_V^2 S_V^2 \left(C\left(\bar{Y}^2 - \frac{1}{T_P}\,S_{\bar{y}^2}^2\right) + C_P S_{\bar{y}}^2\right) = 0$$

This equation of third degree in n has a real solution which we have to resolve with numerical analysis. Similarly, we have

$$\lambda C_P^2 m^3 - \lambda C_P C m^2 - S_{\bar{y}}^2 C_P (T_P^2 - T_V S_V^2) m + \\ + S_{\bar{y}}^2 (C(T_P^2 - T_V S_V^2) + C_V T_V S_V^2) = 0$$

To solve this problem, we can do an approximation in the equality (6.7). Indeed, we assume that $\frac{1}{nm}$ is slight with regard to $\frac{1}{n}$ and $\frac{1}{m}$.

After some calculus, we get:

$$n = \frac{C}{\left(C_V + \sqrt{C_P C_V \dfrac{T_P S_{\bar{y}^2}(T_P^2 - T_V S_V^2)}{T_V^2 S_V^2 (T_P \bar{Y}^2 - S_{\bar{y}})}}\right)}$$

and

$$n = \frac{C}{\left(C_P + \sqrt{C_P C_V \dfrac{T_V^2 S_V^2 (T_P \bar{Y}^2 - S_{\bar{y}})}{T_P S_{\bar{y}^2}(T_P^2 - T_V S_V^2)}}\right)}$$

BIBLIOGRAPHY

[1] Deville, J.C. (1999): Les enquêtes par panel: en quoi différent-elles des autres enquêtes? suivi de: comment attraper une population en se servant d'une autre, *Actes des journées de méthodologie statistiques,* INSEE Méthodes n.° 84-85-86.

[2] Lavallée, P. (1995): Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids, *Techniques d'enquête* vol. 21, p. 27-35.

[3] Lavallée, P. (2002): "Le Sondage Indirect, ou la méthode généralisée du partage des poids",

Éditions de l'Université de Bruxelles, Bruxelles.

[4] Torres Manzanera, E., Sustacha Melijosa, I., Menéndez Estébanez, J. M., Valdés Peláez, L. (2002): "A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places". Ákos Probáld (Ed.): Proceedings Of The Sixth International Forum On Tourism Statistics. Hungarian Central Statistical Office, Budapest.

[5] Valdés Peláez, L. et al. (2001): "A methodology to measure tourism expenditure and total tourism production at the region level". Lennon, J. (Editor): Tourism Statistics. Continuum, London.