Improving the statistical process in the Hotel Occupancy Survey

Elena Rosa-Pérez [1]

[1] *National Statistical Institute, Madrid, Spain; elena.rosa.perez@ine.es*

Abstract
Tourism is one of the most important industries in Spain, therefore an appropriate tourism statistics system is needed. From the supply side, the Hotel Occupancy Survey disseminates monthly data about the main variables. This survey benefits from the administrative Tourism Registers that provide information about new establishments and modifications in the frame variables. Such frame variables include the hotel classification in stars, which is used to stratify the directory and to sample, and the number of bed-places, used as an auxiliary calibrating variable to estimate the number of travellers and overnight stays. Improvements in quality may involve using new data collection methods, for example an e-questionnaire and electronic data reporting, in particular, XML files obtained with an application automatically from the hotel management system. This, in combination with new challenges, like the use of Big Data or the standardization of statistical processes, enhances the quality.

Keywords: (1-5 words), tourism, administrative register, multi-mode data collection, electronic data collection.

1. Introduction

1.1. Tourism in figures in Spain.

The World Tourism Organisation (UNWTO) defines tourism as "*a social, cultural and economic phenomenon related to the movement of people to places outside their usual place of residence, pleasure being the usual motivation*" (UNWTO, 2008). Tourism is one of the most powerful global industries and one of the greatest contributors to employment, development, wealth and quality of life.

In Spain the value of tourism activity reached in 2012 10.9% of GDP. This sector was responsible for over a tenth of Spain's total output and employment[1] . In 2015 the number of arrivals of international tourists was 68.2 million visitors, spending 56.5 billion dollars. This ranks Spain third both in International Tourist Arrivals and in International Tourism Receipts according to the UNWTO Barometer[2]. All this provides an overview of the high importance of tourism in the Spanish economy, importance also present in many other European countries.

However, tourism can also bring with it problems and threats like seasonality[3], the use or abuse of natural resources, job insecurity, etc. These problems affect countries all over Europe, so the European Economic and Social Committee is promoting policies that focus on creating a new European tourism model. To design such adequate policies, high-quality statistics are essential, as they provide an accurate picture of the real situation.

1.2. The Hotel Occupancy Survey.

The Hotel Occupancy Survey (HOS)[4] is a survey carried out by INE Spain according to the Regulation 692/2011[5] with the main objective of providing data on two aspects:

- On the demand side, by giving information on guests, overnight stays and average stay, broken down by country of residence, category of establishments and destination.
- On the supply side, by estimating number of open establishments, bed-places and bedrooms, as well as occupancy rates and personnel employed by category.

The statistical units of analysis are hotel establishments included in the corresponding administrative register of Tourist Boards (ARTB) of the 17 Spanish Autonomous

[1] http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft35%2Fp011&file=inebase&L=1

[2] http://mkt.unwto.org/es/barometer

[3] http://ec.europa.eu/eurostat/statistics-explained/index.php/Seasonality_in_the_tourist_accommodation_sector and http://ec.europa.eu/growth/tools-databases/newsroom/cf/itemdetail.cfm?item_id=8336

[4] http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft11%2Fe162eoh&file=inebase&L=1

[5] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:192:0017:0032:EN:PDF

Communities (SACs). SACs are NUTS-2[6] regions within the European Statistical System. Hotel establishments are classified according to their category (gold and silver) and, within these, by their number of stars. The category of establishments is assigned by the SAC Tourism Office.

The complete HOS frame includes around 18.000 establishments. The periodicity of the survey is monthly and the sample varies from 9.000 units in winter to 11.000 units in summer. The sampling design is a stratified simple random sampling design where strata are defined by the frame variables (i) establishment category and (ii) NUTS-3 regions (provinces). Some categories, like 4- and 5-star hotels, are exhaustive, as they represent more than 50% of the total overnight stays. Others, like silver 1-star hotels, are sampled to minimise the overall response burden over this subpopulations. The remaining categories are exhaustive or sampled depending on the number of establishments in the frame. The sampled establishments remain in the sample for a maximum of four years.

The questionnaire collects data about seven consecutive days within the reference month. In this week-long questionnaire data about the number of travellers, overnight stays, occupied bedrooms, personnel and prices are collected for each weekday broken down by SAC or country of residence in the case of travellers and overnight stays and by type of guests in the case of prices. There are two week-long questionnaires, depending on the category of the establishment and .the differences are that in the 3-, 4- and 5- star hotels questionnaire there is (i) a larger list of countries of residence and (ii) a wider breakdown of the occupied bedrooms.

Due to limitations in the estimation procedures arising from using week-long data for a full month of reference (imputation for the rest of the days is compulsory), a month-long questionnaire was put in place. The main differences between both questionnaires are:

(i) Only 3-, 4- and 5-star hotels are requested to fill in this month-long questionnaire.
(ii) The daily breakdown is suppressed and only monthly totals are requested.

[6] http://ec.europa.eu/eurostat/web/nuts/overview

As we shall explain below, both the daily breakdown and month-long coverage of data have been merged in a single questionnaire collected through electronic data reporting (EDR) in the form of automatically generated XML files. Not only do we gain in information detail for a better accuracy but also we reduce response burden and increase cost efficiency. All questionnaires can be accessed at the HOS website.

2. The Spanish ARTB and the HOS

The HOS takes advantage of the ARTB in several ways, in particular increasing the quality of both the survey frame and the estimates and also minimizing the response burden.

2.1. The Spanish ARTB and the frame population.

The ARTB constitute the original frame of the survey. The ARTB in Spain are managed by the SACs. In 2006 the Directive 2006/123[7] on services in the internal market was launched with the aim of eliminating barriers to the development of service activities and the reduction of administrative burden. This directive was adopted by all SACs and it had several implications in the field of tourism. One of them is that before starting a hotel documents demonstrating that the requirements of the current legislation are fulfilled are to be provided to the Public Administration.

The main consequence of this Directive, for statistical purposes, is the fact that when one hotel opens for the first time, all the information about the establishment has already been provided in the previous step and is already included in the corresponding SAC's Tourism Register. Once the hotel starts, a tourism inspector confirms the information provided about capacity and category. Any alteration in the initial values will be reflected in the ARTB. If the conditions for granting the authorization are no longer fulfilled, the Administration will withdraw this authorization, which will be reflected in the Register.

[7] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006L0123&from=en

INE Spain has an agreement with all SACs entailing the submission to INE of a list with a periodicity ranging from monthly to biannual depending on the SAC. This list includes newly opened establishments, recently closed establishments (not seasonally or occasionally, but due to withdrawal of the authorization or final closure) and those having any alteration in the frame variables. The information sent in those lists includes the identification variables of the hotel (name, address, location, municipality, telephone, tax number, etc.) as well as the capacity (number of bedrooms and bed-places) and the category (type and number of stars). All this information nurtures the survey frame and enables INE to obtain a monthly sample changing from one month to another due to the inclusion of new establishments and the removal of closed ones.

The quality of the ARTB is manifest inasmuch as (i) the high coverage rate derived from the legal mandate to report to the SCA before starting a hotel establishment; (ii) the possibility of INE Spain to influence upon the Registers' variables and their quality.

2.2. The Spanish ARTB and the estimators.

The population of analysis is the set of open hotel establishments across the Spanish national territory. The population aggregates to estimate are the monthly total number of travellers, of overnight stays, of occupied bedrooms, of staff members and the average daily rates. The ARTB include a preliminary variable $\delta_k^{reg.open}$ flagging each establishment as open or closed in the reference month. This frame variable is used to optimize cost efficiency by submitting the questionnaire only to open establishments. During the data collection field work, unanticipated closures must be taken into account in the estimation procedure by flagging those units closed according to the field work. We denote this by the flagging variable $\delta_k^{fid.open}$

The role played by the ARTB can be explicitly viewed as following. All estimators are ratio estimators calibrated by the total number of bed-places in (reg)open establishments in the frame population U_F for each stratum h. The estimated number of any variable y in each stratum U_h (we drop out any time reference for case of notation) is given by

$$\widehat{Y}_h^{rat} = \frac{P_{U_{F,h}}^{reg.open}}{\widehat{P}_{U_h}^{HT,\circ}} \cdot \widehat{Y}_{U_h}^{HT,\circ}, \qquad (1)$$

where

- $P_{U_{F,h}}^{reg.open}$ is the total number of bed-*P*laces in open establishments in the frame population for stratum *h* and is given by:

$$P_{U_{F,h}}^{reg.open} = \sum_{k \in U_{F,h}} \delta_k^{reg.open} \cdot p_k \,, \qquad (2)$$

  p_k being the number of bed-places for establishment *k* according to the ARTB or according to the field work in case of lack of coincidence.

- $\widehat{P}_{U_h}^{HT,\circ}$ is the Horvitz-Thompson estimator using synthetic values p_k° (see below) for the total number of bed-places in stratum *h* and is given by:

$$\widehat{P}_{U_h}^{HT,\circ} = \sum_{k \in s_h} \omega_k \cdot p_k^{\circ}, \qquad (3)$$

  ω_k denoting the sampling weights ($\omega_k = \frac{n_h}{N_h}$ for all $k \in U_h$).

- $\widehat{Y}_{U_h}^{HT,\circ}$ is the Horvitz-Thompson estimator using synthetic values y_k° (see below) for the total number of variable *y* in stratum *h* and is given by

$$\widehat{Y}_{U_h}^{HT,\circ} = \sum_{k \in s_h} \omega_k \cdot y_k^{\circ} \,. \qquad (4)$$

Notice the use of the information provided by the ARTB to calibrate the estimation and thus to reduce the variance of estimators in comparison to direct expansion estimators.

To complete the view on the estimators letting us appreciate the quality improvement introduced by the extended questionnaire, we include details about how the synthetic values are chosen for both the month-partial and month-full data collected modes.

For the month-partial data collected mode through the week-long questionnaire (corresponding to strata involving gold and silver 1- and 2- star establishments), the synthetic values for the number of bed-places are given by:

$$p_k^{\circ} = \delta_k^{resp.week} \cdot p_k, \quad (5a)$$

$\delta_k^{resp.week}$ flagging each establishment as answering or not the week-long questionnaire.

For any other arbitrary variable *y*, they are given by:

$$y_h^{\circ} = \delta_k^{fld.open} \cdot \delta_k^{resp.week} \cdot \frac{D}{7} \cdot y_k^{week}, \quad (5b)$$

where *D* stands for the number of days of the reference month and y_k^{week} denotes the total of variable *y* for the reference week collected in the questionnaire.

For the month-full data collected mode using also the month-long questionnaire (corresponding to strata involving 3-, 4- and 5- star establishments), the synthetic values for the number of bed-places are given by:

$$p_k^{\circ} = \delta_k^{resp.month} \cdot p_k, \quad (6a)$$

$\delta_k^{resp.month}$ flagging each establishment as answering or not the month-long questionnaire.

For any other variable *y*, they are given by:

$$y_h^{\circ} = \delta_k^{fld.open} \cdot \delta_k^{resp.month} \cdot y_k^{month}, \quad (6b)$$

where y_k^{month} denotes the total of variable *y* for the reference month collected in the questionnaire.

Notice the difference: in equations (5b) we are assuming that the establishment remains open/closed during all the reference month in the same terms as during the data-collected week; whereas in equations (6b) no assumption is made, since it is taken from the data themselves.

3. EDR through automatically generated XML files

Given the quality improvement entailed by having detailed data along all the reference month, INE Spain decided to use an EDR collection mode, in particular, through the use of XML files automatically generated in the hotels' management systems.

Nowadays, almost every hotel and similar establishment have installed management package software in their check desks, where the main data of guests are recorded as they check in. These databases contain the necessary information about the visitors, which by means of a light-weight software application can feed an *ex profeso* automatically generated XML file. This file can be sent to the NSI by web service or being uploaded into the NSI's website by the establishments, which by just "pressing a button" can fulfill their statistical obligations.

Since May 2008, in Spain the hotels included in the HOS sample can submit the statistical information using such an XML file. This file contains arrivals, departures and overnight stays broken down by country of residence for each day of the reference month. The list of countries included is the ISO-3166-1 alpha 3 list of the UN[8], which comprises around 250 countries, dependent territories and special areas of geographical interest. Therefore, as shown in the preceding section, no hypotheses have to be made and more accurate estimators can be obtained.

Values collected through EDR reduce the need for imputation to construct synthetic values, which turn out to be (5b) for gold and silver 1- and 2- star establishments not using the XML file and (6b) for those using it.

Thus the detailed estimator for each stratum *h* for the monthly total of variable *y* and strata *h* involving gold and silver 1- and 2- star establishments is given by:

$$\widehat{Y}_h^{rat} = \frac{\sum_{k \in U_{F,h}} \delta_k^{reg.open} \cdot p_k}{\left(\sum_{k \in s_h^{nonXML}} \omega_k \cdot \delta_k^{fld.open} \cdot \delta_k^{resp.week} \cdot p_k + \sum_{k \in s_h^{XML}} \omega_k \cdot \delta_k^{fld.open} \cdot \delta_k^{resp.XML} \cdot p_k \right)} \cdot \left(\sum_{k \in s_h^{nonXML}} \omega_k \cdot \delta_k^{fld.open} \cdot \delta_k^{resp.week} \cdot \frac{D}{7} \cdot y_k^{week} + \sum_{k \in s_h^{XML}} \omega_k \cdot \delta_k^{fld.open} \cdot \delta_k^{resp.XML} \cdot y_k^{month} \right), \quad (7)$$

[8] http://unstats.un.org/unsd/tradekb/Knowledgebase/Country-Code

where the sample *s* is divided into subsamples s^{XML} and s^{NoXML} attending to the collection mode of their units.

For strata *h* involving 3-, 4- and 5- star establishments this estimator is given by:

$$\widehat{Y}_h^{rat} = \frac{\sum_{k \in U_{F,h}} \delta_k^{reg.open} \cdot p_k}{\sum_{k \in s_h} \omega_k \cdot \delta_k^{fld.open} \cdot \delta_k^{resp.month} \cdot p_k} \cdot \sum_{k \in s_h} \omega_k \cdot \delta_k^{fld.open} \cdot \delta_k^{resp.month} \cdot y_k^{month} . \quad (8)$$

The automated data collection system has several advantages: (i) non-excessive response burden, (ii) better timeliness, (iii) higher cost-effectiveness, (iv) improved editing and validation procedures (because checking rules for editing are effective before data enter the NSI's system) and better accuracy.

4. Forthcoming challenges

Several challenges related to the improvement of the quality in the HOS will be faced in the next years.

4.1. New e-questionnaire.

As part of the modernization of editing strategies, a new e-questionnaire has been launched. This e-questionnaire incorporates edits with parameters computed specifically for each variable and each unit (say, validation intervals built using both the individual historic and cross-sectional values of each variable). This more efficient, though computationally demanding, editing strategy has already been implemented in five short-term monthly business statistics, having reduced respondents' recontact rates by up to 20 percentage points upon their sample sizes.

4.2. Big Data.

INE Spain, as a request from the hotel sector, calculates and publishes monthly Indicators on the Profitability of the Hotel Sector. These indicators are the Average Daily Rate (ADR) and the Revenue per Available Room (RevPAR), and together with the Bedroom Occupancy Rate,

constitute an important source of information for hotel establishments, which enables them to evaluate their pricing policy or revenue management. A challenge ahead lies on the possibility of scraping the web to massively download these data. However, so far, risks of representativeness and other methodological problems remain to be solved.

5. Conclusion

Both the use of ARTB and the automatic data collection mode through EDR in the form of XML files clearly aligns with several principles of the European Statistics Code of Practice[9].

However, the implementation of this automatic data collection scheme through XML files generated at the respondents' information systems has to face some difficulties. Firstly, this scheme needs the installation of a light-weight software application in these systems and some reticence is common among respondents. Secondly, the development and deployment of this application may require an initial expenditure, thus the debate about whom (either respondents or NSIs) should take care of this burden rightfully arises. This points, as many Big Data sources of information, towards a new paradigm of private-public partnership within official statistics.

Based upon the former analysis, we firmly believe that a modernised and industrialised statistical production system must follow this avenue.

6. References

Wallgreen A. and Wallgreen B. (2007), Register-based Statistics. Administrative Data for Statistical Purposes, J.Wiley, N.Y.

UNWTO (2008). International Recommendations for Tourism Statistics 2008. United Nations Series M No. 83/Rev.1.

[9] http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15