RENEWING THE EUSTAT TOURISM SURVEY: NEW COLLECTION METHODS AND DESIGN FOR MORE DETAILED ESTIMATES

Jorge Aramendi[1], Elena Goni[2], Anjeles Iztueta[3], Marta Salvador[4], Fernando Tusell[5]

[1]EUSTAT: j-aramendi@eustat.es; [2]EUSTAT: elena_goni@eustat.es; [3]EUSTAT: aiztueta@eustat.es
[4]EUSTAT: marta_savador@eustat.es ; [5]UPV/EHU: fernando.tusell@ehu.es

Key words: Tourism statistics, hot-deck imputation, time patterns; spatial and temporal disaggregation, XML-files, respondents burden

1. Introduction

The *Tourism Survey* carried out by EUSTAT constitutes one of the pillars of the tourism information system in the Basque Country. It is sent out monthly to some 1000 establishments, from large hotels to rural houses or inns. Variables investigated include night-stays, arrivals, occupancy (both in terms of rooms and beds) and length of stay, all broken down by different geographical strata and visitors origin, as well as the number of employees.

Data is collected for each day of the month for large establishments and, in order to minimize their response burden, for a random period of seven days within the month for smaller ones. Results are weighted, processed and published monthly, providing totals or, aggregated per stratum and province.

Even though published results provide considerable detail, users ask frequently for more detailed figures: occupancy rates or night-stays for periods shorter than one month (weekends, Easter, Bank Holidays,…), results for specific categories of hotels or smaller geographical areas. Meeting those demands requires customized processing of the data, with a considerable cost for each particular query.

EUSTAT has adopted a new approach that provides estimates with the maximum time and geographical disaggregation in order to be able to respond to any such information request. The new design is based on two aspects: first, the data is collected daily for the higher hotel categories (3 stars or more) through electronic data collection (XML-files). Second, we have opted for an imputation approach to obtain a macrotable that contains exhaustive information (available places, travellers and overnight stays) for the whole tourist establishment population for any single day of the month. A Hot-deck imputation method is applied where the donor is the nearest neighbour unit, determined according to the distances in the patterns of the time occupancy or in geographical location, for the same day.

2. Weighting approach and data analysis

The previous method to produce monthly figures for each stratum was obtained by a simple weighting operation. The weighting factor was computed using the number of offered beds, which was available for each hotel by other means, whether it answered the questionnaire or not. Therefore, this method assumed hat the average observed occupancy rate could be extended to non-observed beds in the stratum.

The whole dataset of the Tourism Statistics has been extensively analysed in order to establish a new approach that can accommodate a wide variety of information queries without requiring taylor-made weighting schemes.

There are two main features in the data set, seasonality and the irregular sampling pattern, that have heavily influenced the methodology finally adopted. Yearly, it has been observed a high occupancy during summer, particularly in August, and a higher occupancy in weekends, especially in Easter. The irregular sampling pattern over time, on the other hand, can be a consequence of the sampling design. Many hotels, particularly of small size, are only required to provide data for one period of seven days of the month. These periods are selected at random for each hotel, among six different choices, which partially overlap. The final result is that the number of hotels sampled is not completely uniform along the month.

3. An imputation approach

3.1. Background

Rather than coping with different weighting factors for each user demand, a natural choice is to consider a full N x T x K table, where N is the number of establishments, T is time in days, and K the number of variables; for simplicity, we will only consider a N x T table, i.e. the information corresponding to only one variable. Since we do not observe each hotel for each day, such N x T table would have many cells missing and, therefore, we have to select a suitable strategy for imputing.

It has been decided to base our imputation in a donor method, with missing values for one hotel being filled with those of a similar hotel. The choice of a donor-based method in preference to a formal model-based method or multiple imputation is justifiable on several grounds: it is traceable, affording a clear understanding of where the imputed values come from and, most importantly when several variables have to be imputed at once, it guarantees consistency of the imputed values.

Further decisions at the onset of the project were: a) Only “first generation" donors should be accepted, i.e., an imputed value should always be an observed value, rather than a value previously imputed, and b) Donors for missing observations should be taken first from the same stratum as the recipient.

3.1. Distance calculation

First of all, a proximity or “likeness" notion must be defined among time series, so suitable donors must be chosen. A full clustering strategy is not needed, but rather a similarity measure to rank candidate donors from closest to furthest.

One approach is a distance between time series (corresponding to hotels) i and j, for instance, the Euclidean distance,

$$d^2(i,j) = \sum_t \left| y_{it} - y_{jt} \right|^2 \qquad \textbf{(1)}$$

where y_{it} is the observation of y for hotel *i* at time *t*. For each hotel we have several candidate time series: night-stays, arrivals, or computed magnitudes such as occupancy. Since hotels are of very different sizes, it makes sense to adopt the occupancy rate, whose range is independent of the size of the hotel. Hence, we define our distance measure between hotels *i* and *j* as $d^2(i,j)$ in (1), the (square) Euclidean distance between bed-places occupancy profiles viewed as vectors in R^T.

Due to the observed marked seasonal pattern a plausible model for occupancy series of hotel *i* is:

$$y_{ij} = \beta_{i,Trend(t)} + \beta_{i,DayOfYear(t)} + \beta_{i,DayOfWeek(t)} + \beta_{i,Easter(t)} + \varepsilon_{it} \quad \textbf{(2)}$$

where *t* is time measured in days since January, 1, 2007. $\beta_{i,Trend(t)}$ is a smooth function of time to capture variation over the years; $\beta_{i,DayOfYear(t)}$ is a function of the day within the year associated with *t* (i.e., *DayOfYear(t)* takes values from 1 for January, 1, to 366 for December, 31; account is taken of leap years). Likewise, $\beta_{i,DayOfWeek(t)}$ is a term capturing the effect of the day within the week corresponding to *t*. Finally, $\beta_{i,Easter(t)}$ is a dummy variable taking value 1 in around Easter and ε_{it} and is a random term.

The model above cannot be fitted correctly in some case like, for instance, new hotels with no previous data. For such cases the following approach has been used. Hotels in the same municipality are taken to be at distance zero. Those which are in neighbouring municipalities are at distance 1, and so on. “Neighbouring" means that the two municipalities limit with each other (a full contiguity matrix was computed from digital cartography). Thus, two hotels are at distance d according to this notion if we have to traverse (d+1) neighbouring municipalities (including the origin and destination). This distance is further corrected by an increment of 0.5 for pairs of hotels that are not of the same category. This causes that among candidate donors in the same municipality, those of exactly the same category than the receiver are always preferred. Hotels in different strata are set at infinite distance.

We thus have two alternative distances for each pair of hotels: a Euclidean distance computed from (1) when data from both hotels permit the fitting of model (2) and a “geographical distance" as described in the previous paragraph, when for either (or both) hotels we do not have enough data to support the fitting of model (2). The Euclidean distance is used whenever possible.

3.3. Donor selection

It may appear that all that is left is, for any hotel requiring imputation, to pick the closest match (or an average of closest matches) in (1) and perform the imputation taking into account the required scale adjustment and some other details that will not be mentioned, to keep it simple.

But, due to the sampling scheme used (one week per month observed for the vast majority of the cases), the closest donor candidate may be able to provide values for some weeks but not for others. It will be exceptional that a single donor provides all required values; in most cases we need to pull data from several donors, which are chosen sequentially in order of increasing distance. Therefore, the donor selection needs to be multiple to accommodate different needs across time for the same establishment.

3.4. Employment imputation

The number of employees is imputed for the whole month by a two-step rule where past monthly values of the establishment are first analysed and, when not steady, year to year variations within the strata are applied to last observed data. The underlying assumption is that the evolution of the employment, regardless its seasonal component, is quite steady in the short term.

4. Results

The estimation models following the old and the new approach are:

$$\hat{Y}_P^{OLD} = \frac{\sum\limits_{(i,t)\in P} B_{it}}{\sum\limits_{(i,t)\in O} B_{it}} \times \sum_{(i,t)\in O} y_{it} \qquad \textbf{(3)}$$

$$\hat{Y}_P^{NEW} = \sum_{(i,t)\in O} y_{it} + \sum_{(i,t)\in P} \frac{B_{it}}{B_{jt}} \times y_{jt} \qquad \textbf{(4)}$$

In equation (3) we simply aggregate the magnitude of interest over observed hotels and multiply the total by the ratio of total offered beds (both observed and unobserved) to beds in the observed hotels.

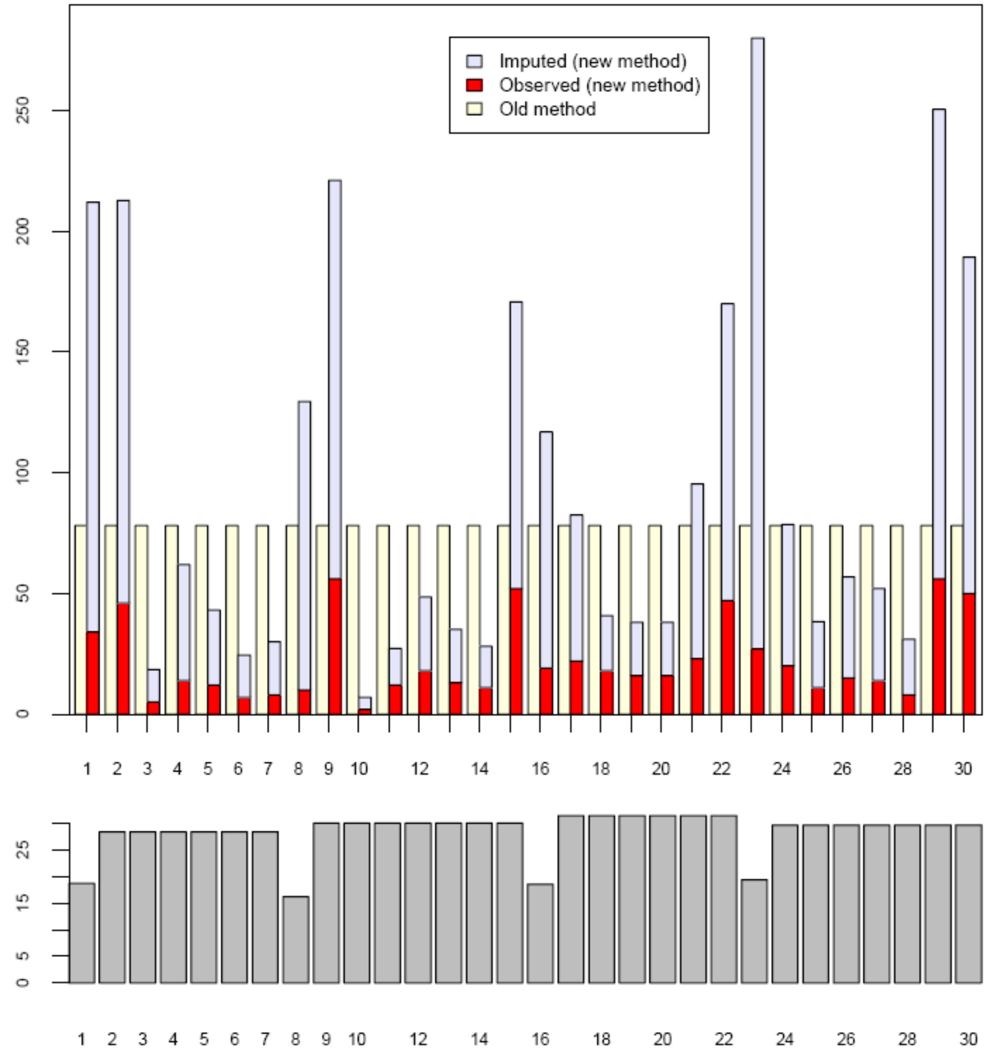
In equation (4), on the other hand, we aggregate the magnitude of interest over observed hotels and add imputed values of the non observed hotels. The imputed values for hotel i are obtained daily from those of a similar donor j, once adjusted for size.

The main source of discrepancies among the old and the new method can be traced to the irregular pattern of daily response along the month (due to both the sampling design and non-response from hotels), coupled with large differences in occupancy from one day to the next. As an example, Figure 1 graphs results obtained from one middle-sized stratum for June 2007.

The lower panel shows the percentage of observed beds, i.e. beds in hotels who were surveyed and answered, relative to the total number of beds in hotels known to be open in the stratum. It can be seen that in four days the number of beds observed is significantly lower, due to the sampling design.

It is clear that substantial day-to-day differences exist. The new, imputation-based, method successfully recovers large numbers of night-stays for the five weekends in the month and lower numbers in other days. Both effects will tend to cancel, but the balance may still show substantial differences, as is the case here: the total imputed night-stays for the considered stratum is 2340 when using the old method and 2827 using the new. The bulk of the discrepancy can be traced to the fact that days of high occupancy such as June, 1, 8, 16 and 23 were thinly sampled, which lowered the average occupancy rate when using the old method. When using the imputation method, this did not happen.

Figure 1. *Daily estimation of night-stays (upper panel) and percentage of beds in observed hotels (lower panel) for (June, 2007, stratum 15)*



5. Final remarks

A simple, donor-based method has been devised and implemented to ease the task of producing customized estimates of the EETR. The approach is simple and appears to work well. We have found, nonetheless, that with small samples and/or high non-response it may lead to instabilities. A single or a few hotels may become nearly universal donors for their stratum: estimates become then over sensitive to a small number of observations.

The new method constitutes a mayor re-engineering of the survey and other improvements have been introduced.

All the modelling has been programmed in R and has been correctly integrated into other processes of the survey.