MOVE 2011
2nd International Conference on the Measurement and Economic Analysis of Regional Tourism
BILBAO · SPAIN · OCTOBER 27TH–29TH · 2011

Organized by:

Improving time and spatial resolution of tourism statistics through imputation

5. Measurement and analysis tools

Elena Goni
EUSTAT Instituto Vasco de Estadística/Euskal Estatistika Erakundea; Methodology, R+D and Innovation Unit, Statistician
Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ, Spain
E-mail: elena_goni@eustat.es Phone: (+34) 945017522. Fax: (+34) 945017501

Fernando Tusell
Universidad del País Vasco/Euskal Herriko Univertsitatea, Facultad de CC.EE. y Empresariales.
Department of Applied Economics III (Statistics and Econometrics). Professor.
E-mail: fernando.tusell@ehu.es

Jorge Aramendi
EUSTAT Instituto Vasco de Estadística/Euskal Estatistika Erakundea; Methodology, R+D and Innovation Unit, Statistician
E-mail: j-aramendi@eustat.es

Maria J. Bárcena
Universidad del País Vasco/Euskal Herriko Univertsitatea, Facultad de CC.EE. y Empresariales.
Department of Applied Economics III (Statistics and Econometrics). Assoc. Prof.
E-mai: mariajesus.barcena@ehu.es

Abstract

A method is described to impute missing values in the EETR (Encuesta de Establecimientos Turísticos Receptores), the main source of information on hotel occupancy in the Basque Country. We discuss the objectives and alternatives considered, and describe the algorithm.

Keywords: imputation; spatial disaggregation; tourism statistics

1. Background

In the Basque Country, the EETR (Encuesta de Establecimientos Turísticos Receptores) is a statistical survey whose aim is to evaluate the number of visitors staying in hotels and guest houses, spent, nights length of stay, etc., broken by geographical origin. It has been in operation for over a decade now, administered and exploited by EUSTAT[1]. There has been a steady increase in coverage, as measured by the number of establishments surveyed.

In order to reduce the administrative load and inconvenience to the hotels and guest houses surveyed, a decision was made to ask for data of only a week per month; only large establishments provide data for each and every day of the month. (In recent years, with the widespread use of computer programs in the administration of even small hotels and guest houses, daily reporting is becoming more common: an electronic XML document is filled and sent to EUSTAT, providing daily information without manual intervention or administrative paperwork.)

[1] EUSTAT, Instituto Vasco de Estadística / Euskal Estatistika Erakundea, is the institution in charge of the official statistics in the Basque Country.

Aside from the “missing by design” pattern introduced by the sampling scheme (only one week per month with data for the majority of respondents), non-response is small, although very irregularly distributed in time and space.

The information received, either in paper or electronically, is processed to produce monthly totals of visitors received, average occupancy rates, etc. per stratum. Strata defined distinguish country guest houses from other hotels, and these are further segmented by category and location: Biscaye, for instance, is divided in Bilbao, coastal Biscaye and inland Biscaye regions, as these segments may exhibit different behaviour. The methodological note for the EETR operation, available at http://es.eustat.es, gives full details on the number and definition of strata.

It has been repeateadly the case that published figures are insufficient to some users for some purposes: details for periods shorter than one month, for regions smaller than one of the predefined segments, or both, are sometimes required. EUSTAT performs on demand *ad-hoc* processing of the raw information to meet those requirements; clearly, though, this is a time consuming and specialized task, as it requires computing different expansion factors in each case.

Rather than coping with changing coverage and expansion factors, a natural choice is to consider a full table *N x T* where *N* is the number of respondents and *T* is time in days. (We will refer in the sequel to the *N* sampling units as “hotels”, even though many are best described as guest houses or inns.) With the present sampling scheme, many cells in such table are missing by design; the idea is to impute them, so virtually any desired magnitude can be obtained simply by aggregation of the relevant rows or columns of cells. (The *N* rows of respondents can be selected by geographical location, category, size, etc.) In other words, starting from a table such as Table 1 (where the crosses denote available data and empty cells missing data for a sample of hotels (A, B, …, ZZ) and 31 days of a given month, our goal is to construct a table such as

Table 1. Raw data layout

Hotel	Day 1	Day 2	Day 3	Day 5	Day 6	Day 7	Day 8	Day 9	…	Day 28	Day 29	Day 30	Day 31
A	X	X	X	X	X	X							
B										X	X	X	X
C						X	X	X					
…													
ZZ	X	X	X	X	X	X							

Table 2, which for each hotel and day has data, either observed or imputed. The cells with “X” in Table 2 contain the same data as in Table 1, and the cells with an “I” contain imputed data: our reconstruction of the figures which might have been collected from the corresponding hotels, but were not, either because they were not asked or because the hotel did not respond at all (so the imputation cover also de case of non-response).

Table 2. Imputed table

Hotel	Day 1	Day 2	Day 3	Day 5	Day 6	Day 7	Day 8	Day 9	…	Day 28	Day 29	Day 30	Day 31
A	X	X	X	X	X	X	I	I		I	I	I	I
B	I	I	I	I	I	I	I	I		X	X	X	X
C	I	I	I	I	I	X	X	X		I	I	I	I
…													
ZZ	X	X	X	X	X	X	I	I		I	I	I	I

It is clear that once we have a table such as Table 2 constructed, every single question which might conveivably be asked can be answered by simply adding over the columns which make up the time interval of interest and over the rows corresponding to hotels in the geographical area, category, size, etc. of interest.

2. Imputation method

Design choices

To construct a table such as Table 2, we have to select a suitable strategy for imputation. Earlier in the development of the project it was decided to base our imputation in a donor method, rather than a model -based method, with missing values for one hotel being filled with those of a "like" hotel. A "likeness" criterion had therefore to be defined. The choice of a donor-based method in preference to a formal model-based method is justifiable on several grounds: it is traceable, affording a clear understanding of where and how the imputed values come from and, most importantly when several variables have to be imputed at once, it guarantees coherency of the imputed values (since imputed values are always actual values observed for the donor).

Further decisions at the onset of the project were: a) We would only allow "first generation" donors, i.e., an imputed value should always be an observed value, rather than a value previously imputed, b) We would produce single rather than multiple imputations (see e.g. Rubin(1987), Schafer(1997)) for each value missing, and c) The pool of donors for missing observations would be restricted to the same stratum. For the purposes of this survey, the Basque Country is divided in 14 well-understood strata, and it was thought unwise that a donor should ever come from a different stratum than the receiver.

Since data for each respondent assumes the form of a time series, a proximity or "likeness" notion must be defined among time series, so suitable donors can be chosen. The problem of clustering time series has received a lot of attention: for a survey, see (Liao, 2005). Other recent references include (Shingal and Seborg, 2006), (Coke and Tsao, 2010) and (Genolini and Falissard, 2010), this last dealing with missing values in a manner not unlike our approach here. Since we do not need a full clustering strategy, but rather a similarity measure to rank candidate donors from closest to furthest, a simple approach which immediately suggests itself is a distance between series *i* and *j* such as,

$$d_{ij}^2 = \sum_{t=1}^{T} \left| x_{it} - x_{jt} \right|^2$$

where x_{it} is the observation of the *i*-series at time t. The problem with this is that our series are very sparse, and for any given pair of series, the set of time points at which both have an observed value is very small. A Gower adjustment (Gower, 1971) might be envisaged. What we have done instead is to fit trajectories to each series and compute distances between the fitted trajectories, as described next.

Time series modeling

Examination of our series shows that they share some common patterns: there is a weekly variation pattern, which for the majority of the series implies higher occupancy at weekends, while for another group (conceivably, hotels frequented by business travelers, rather than tourists) weekends are the period of lowest occupancy. There is also, for most series, a marked seasonal pattern, which again for most series implies higher occupancy in July and August. Finally, there is also year-to-year variation.
In the light of this, a plausible model for occupancy[2] series (in relative terms, i.e. as a fraction of total capacity) is:

$$x_{it} = \beta_{i,Year(t)} + \beta_{i,DayOfYear(t)} + \beta_{i,DayOfWeek(t)} + \beta_{i,Easter(t)} + \xi_{it}$$

In the above formula, *t* is time (measured in days) in the period from January, 1, 2000 to December, 31, 2009. *Year(t)* is the year associated with *t, DayOfWeek(t)* is the day within the year associated with *t* (i.e., takes values from 1 for January, 1, to 366 for December, 31; account is taken of leap years). Finally, $\beta_{i,Easter(t)}$ picks up the effect of the Easter holiday season, and ξ_{it} is a random term.

The specification above would use for each series 10 + 7 + 366 + 1 parameters, respectively for the year effect (10 years of data), day-of-week effect (7 days), day-of-year effect (366 days) and Easter effect, which is accounted for separately due to the movable nature of Easter[3]. Identification constraints would reduce the number of free parameters by two (constraining the sum of the day-of-week and day-of-year effects to be zero). Even so, the number of parameters required is far too large to make this a feasible model as it stands.

After accounting for the day-of-week effect, one may expect the year profile to be rather smooth. Hence, what we have done is to fit the day-of-year effect with a smoothing cubic spline which greatly reduces the number of free equivalent parameters (as defined, e.g. in Hastie and Tibshirani, 1990) used[4]. After some experimenting, we have settled for 12 equivalent parameters in the smoothing spline term[5].

It may help to understand the data to look at the decomposition afforded by our model in the case of one series, chosen at random, and representative of the behaviour common to the vast majority; it is presented in Figure 1. Only two years of the fitted trajectory components are shown, so that the weekly fluctuation can be clearly discerned. The first panel shows the effect of the year; for the particular series shown, year 2007 was on average worse than 2008, but the difference was not large; about four

[2] Occupancy can be defined in terms of occupied rooms or occupied beds with respect to the total available. Occupancy in terms of rooms has been used here, even if occupancy in terms of persons is of more direct interest. Bear in mind that the model is only used to select "close" donors. The use of either variable leads to very similar choices of donors.

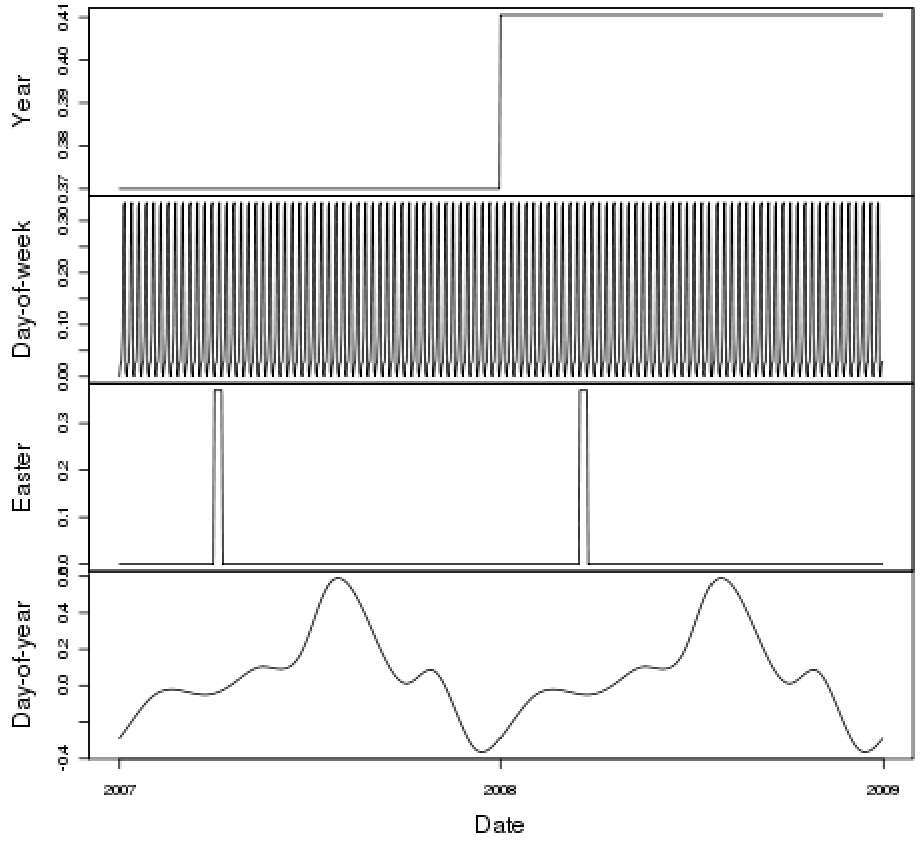
[3] The reader may find it easier to rewrite the model in terms of dummy variables. Our term $\beta_{i,Year(t)}$, for instance, is shorthand for $\sum \beta_{i,j} x_{t,j}$, where $x_{t,j}$ is a dummy variable taking the value 1 when *t* is a date in year *j* and zero otherwise; similarly for the other terms in our model. This notation more clearly shows that, as it stands, the model would need the number of parameters β stated above.

[4] Intuitively, if each day could be modelled by an unrestricted parameter, we need at least one (and preferably many) observations for each day of the year. If, however, we impose smoothness, "close" days are restricted to have nearly the same effect. Even days of the year absent from the sample can be estimated if there is information on close days before and after. The smoothness restriction makes up for scarce information.
There is a downside; salient spikes (e.g., a peak of occupancy one particular day of the year, possibly coinciding with a local holiday) will be smoothed out; but they will be smoothed similarly for all hotels, and will still weight in the selection of donors.

[5] Smoothness is imposed at year ends by using a periodic spline rather than a natural spline; Dec, 31 is restricted to connect with Jan, 1st with no discontinuity.

percentage points of occupancy. The second panel displays the intra-week pattern, which for this time series shows that occupancy within the week fluctuates on average by over 30 percentage points. The third panel shows that the fitted effect of the Easter holiday is to boost occupancy by over 30 percentage points, while the fourth and last panel shows the effect of time within the year on occupancy[6].

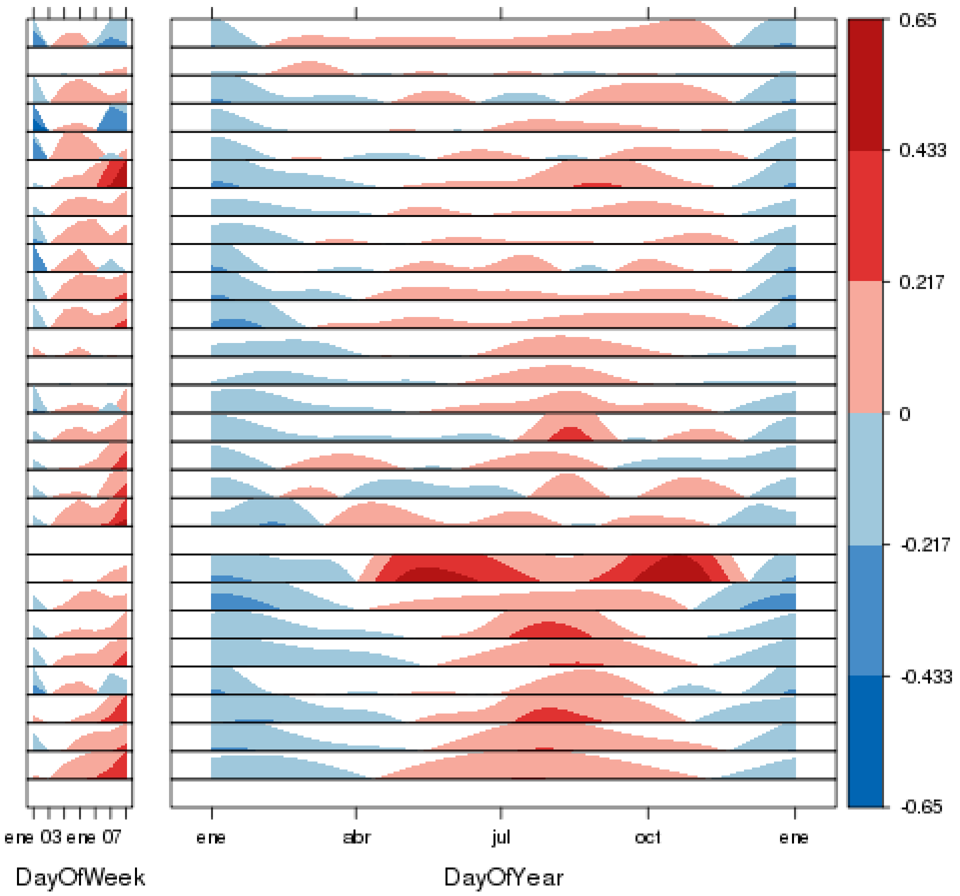
Figure 1. Decomposition of an occupancy series in its year effect, day of week, Easter and day of year constituents.



As an illustration of the patterns of occupancy in our sample, a horizon plot, Few (2008), of the first 28 hotels in our sample is shown in Figure 2. The left hand panel shows the $\beta_{i,DayOfWeek(t)}$ effect (from Monday to Saturday), showing that for most hotels weekends are the periods of the highest occupancy. The right hand panel displays the $\beta_{i,DayOfYear(t)}$ effects and shows that, again for the majority of 28 hotels, summer is associated to higher occupancy than the period before and after New Year.

[6] The perceptive reader will notice that for some days of the week at some dates in the year, the model may produce fitted occupancies of over 100%, or below zero, percentage points: we might have modelled a transformed variables rather than occupancy to avoid such inconvenience. Remember, though, that our modelling exercise does not attempt to produce final estimates, but rather occupancy trajectories whose similarities can assist us in the choice of donors. The use of transformed variables is further discussed in Section 3.

Figure 2. Profile of the day-of week and day-of-year effects for 28 occupancy series



Computation of distances

The model described has been fitted to each occupancy series; after estimating the effects, the distance between series *i* and *j* can be computed as

$$\hat{d}_{ij}^2 = \sum_{t=1}^{T} \left| \hat{x}_{it} - \hat{x}_{jt} \right|^2 ;$$

where $\hat{x}_{it}, \hat{x}_{jt}$ are the fitted trajectories for the respective series.

We might as well pause here to see what has been accomplished. By fitting a trajectory to each series, we can evaluate the sum defining $\hat{d}_{ij}$ for all time points *t* within the common range of the two series[7]; had we tried to use the raw data, we would have found that only a small proportion of time points have observations for both time series whose distance we need to compute.

[7] If two series do not overlap at all, their distance is set to infinity, which precludes any of them ever being selected as a donor for the other.

Out of 1061 occupancy series, 921 were fitted as described. The remaining series were too short and have been dealt with in a different manner.

An important factor affecting occupancy is geographical location. Local holidays, sport events, fairs and festivals are all factors that tend to boost occupancy in a municipality and, due to spill-over effects, those which are in close vicinity. Whenever the distance between trajectories can be computed, this effect can be expected to produce small distances for hotels which are spatially close to each other; when lack of sufficient data hinders such computation, direct resort to geographical closeness is made instead.

We have used the following approach: 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 share a boundary with each other; a 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) in going from one to the other. This distance is further corrected by an increment of 0.5 for pairs of hotels that are not of the same category (most strata include more than one category). The effect is 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.

Figure 3. Image representation of distances among the hotels with fitted trajectories, color-coded so that higher intensity of red means "closer"

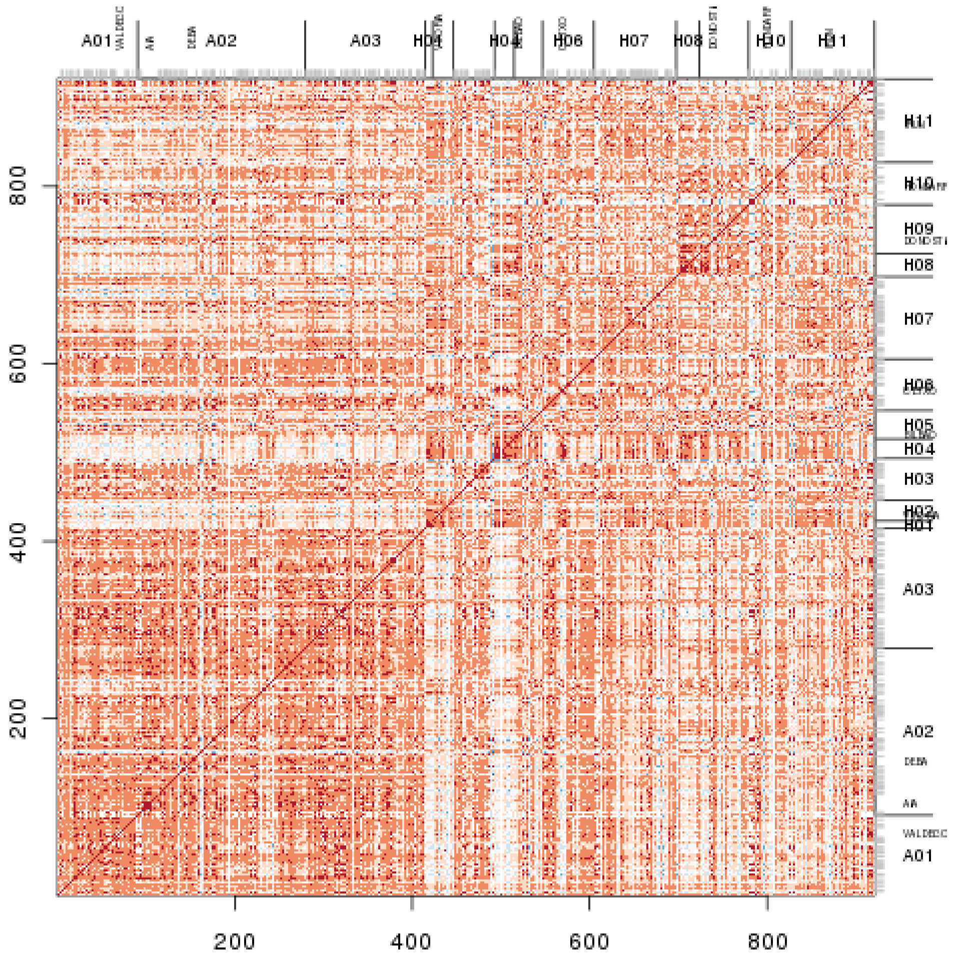


Figure 3 displays pictorially the (trajectory-based) distances between hotels (grouped in eleven strata, H01 to H11) and rural guest houses (grouped in three strata, A01 to A03, one for each of the Territories, Alava, Gipuzkoa and Bizkaia). Color red encodes the minimum distance, hence the bright diagonal (each hotel is closest to itself). Aside from that, rural guest houses clearly have a life of their own, clearly shown in the large region in the south west corner. The remaining strata are also reflected in the computed distances among trajectories, and evidence themselves in more intense squares over the diagonal, particularly H01-H02, H04 and H08 (which correspond to the capital cities of the three Territories).

Figure 3 is reassuring as it shows that the method is successful in picking the similarities in the data, and also that strata were judiciously chosen in the first place.

Imputation algorithm

We have so far two distances (one for pairs of occupancy series for which trajectories could be fitted, another one based in geographical location when the former could not be computed). 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) and perform the imputation.

This is not so; because of 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 is exceptional that a single donor provides all required values; in most cases we need to pull data from several donors.

Aside from that, we have to correct size effects; two hotels may have a very similar occupancy profile, and thus be closed in terms of the $\hat{d}_{ij}$ distance defined above (or else in terms of geographical distance), yet they may be of very different size, which would prevent direct donation of arrivals, etc. in absolute terms. Instead, arrivals, A_t, and overnight stays, N_t, are imputed in the receiver as being the same ratio of capacity observed in the donor. Variable "departures", D_t, is imputed so as to balance the relationship $N_{t-1} + A_t - D_t = N_t$ (on occasion, further adjustments may have to be made).

We can now summarize the procedure as follows:

1. Fit a smooth trajectory to each occupancy series with sufficient data.
2. Compute distances among smooth trajectories.
3. Compute geographical distances for hotels with insufficient data.
4. For each hotel with missing data, set the pool of metric or geographical donor candidates in order of increasing distance and iteratively, until no missing values remain:
   a) Pick next closest candidate donor.
   b) Impute multiplying the occupancy and arrival ratios of the donor by the size of the receiver.
5. Adjust and (optionally) round imputations to integer values.

So, for instance, if hotel B, with 50 rooms and 90 beds is the closest available candidate to become a donor for hotel A, with 20 rooms and 40 beds, for every day without data in hotel A and observed in hotel B we will use hotel B as donor. For a day *t* in which we have observed in hotel B 45 arrivals and 75 overnight stays, we compute:

- Size ratio of the receiver to the donor: 40 / 90 = 0.4444
- Imputed number of overnight stays: 0.4444 x 75 = 33.33
- Imputed number of arrivals: 0.4444 x 45 = 20

If for the previous night we had in hotel A, say, 30 overnight stays (either observed or imputed), the number of departures is imputed so that the restriction $N_{t-1}+A_t-D_t=N_t$ holds: 30 + 20 – 33.33 = 16.66. We could round numbers at this stage, but have chosen to allow non-integer values for arrivals, overnight stays and departures and round the aggregated figures.

On some ocasions no donor can be found in the stratum for a particular day *t*; in such cases, sufficiently rare not to be worried about, we simply impute the average occupancy and arrival rate of the stratum or even of the Territory.

The method has been prototyped in R with satisfactory results.

3. Alternatives and discussion

There are some rough edges in the method used; as usual, a compromise must be struck between simplicity and functionality and mathematical elegance; and in some instances we had to settle for a solution in order to preserve simplicity.

To begin with, the model fitted to occupancy is additive and the response is bounded, constrained to be in the [0,1] range. This means that fitted values are occasionally above 1 or below zero. While we could map the response variable to the [0,1] range[8], we would lose interpretability of the estimated effects, which would be on an unfamiliar scale.. On the other hand, as has already been pointed, the fitted trajectories are only used for the purpose of defining a distance between the time series associated to each hotel.

More elaborate distance notions between hotels could be entertained, in particular to account for the multivariate nature of the series associated to each of them. For instance, we may imagine two hotels with very similar occupancy patterns, yet an entirely different pattern of rotation of their customers. Our method would impute correctly occupancy, but might produce badly wrong timings for arrivals or departures. Again, we chose the simpler alternative; but were this thought to be a problem, a distance between multivariate time series might be computed along the lines in (Shingal and Seborg, 2006), and donors be selected accordingly.

One alternative to the simple donor method which has been outlined above was entertained for a while, and gave results deemed adequate most of the time. The idea was the following: since we are fitting trajectories to most of our hotels, we could use those trajectories to refine our imputations. If, for instance, hotel *i* is being imputed with data from hotel *j,* we might consider to estimate the occupancy x_{it} by

$$x_{it} = x_{jt} + \left(\hat{x}_{it} - \hat{x}_{jt}\right);$$

in other words, we would adjust the observed value in hotel *j (donor)* adding the expected discrepancy between the values of donor and receiver, given by $\left(\hat{x}_{it} - \hat{x}_{jt}\right)$, the difference in value of the fitted trajectories at time *t.*

This is an appealing idea which, among other things, makes imputation dependent on the history of both donor and receiver. Notice, though, that if we make the correction shown for *each* variable we need to impute (arrivals, occupancy in rooms, occupancy in persons...) the method is overly complex, as it requires fitting a trajectory to each variable; and coherency between the imputed values is not guaranteed.

[8] For instance, if the desired response is *Y* = "occupancy", constrained to be in the interval [0, 1], we might define $Z=\log_e(Y/1-Y)$, fit our additive model to *Z* and then recover *Y* using the inverse transformation $Y=(e^Z/1+e^Z)$.

The method which was adopted, and later abandoned to be replaced by the one described in the previous section, was to fit a trajectory per hotel for a single variable (occupancy measured in rooms), adjust the imputed value of such variable using the formula $x_{it} = x_{jt} + (\hat{x}_{it} - \hat{x}_{jt})$, and then generate imputations for all other variables multiplying x_{it} by suitable constants.

For instance, if the variable imputed with the correction $(\hat{x}_{it} - \hat{x}_{jt})$ is X_{it} = "number of rooms occupied", the number Z_{it} of overnight stays was computed as $Z_{it} = c_i X_{it}$, where c_i is a proportionality factor which gives the average number of persons per room, presumably related to the mix of single and double rooms offered by hotel *i.* As it happens, c_i fluctuates considerably over time: the persons per room ratio that c_i was meant to capture increases considerably during summer time, when double occupancy of rooms is more common. Consequently, the idea had to be dropped.

A further comment refers to the comparability of results using the imputation method outlined here and the current method which uses expansion factors.

The current method computes the monthly total of e.g. overnight stays in hotels that answer the survey, and then multiplies that figure by an expansion factor. This expansion factor is the ratio of two numbers:

- Number of bed-nights offered: the sum over days of the month of beds offered by hotels open each day. In other words, the number of overnight stays had all hotels been fully occupied each and every day they remained open.
- Number of bed-nights surveyed: the sum over days surveyed of all beds offered by respondent hotels.

In other words, a hotel with 70 beds open for the whole of a month of 30 days, contributes 70 x 30 = 2100 bed-nights to the first magnitud ("Number of bed-nights offered"). If, however, answered the survey for only one week, contributes 70 x 7 = 490 to the second magnitude ("Number of bed-nights surveyed"). If that were the only hotel in existence in a given stratum, the expansion factor would be 2100 / 490 = 4.286.

The imputation method outlined "fills in" missing data using information from one or several donors. When aggregating the imputed table (Table 2 above) for a whole month we would expect similar results as obtained with the standard method of expansion. However, discrepancies do occur, a matter that perplexed us for a while and which nonetheless has a simple explanation.

The current, expansion method, treats all days similarly, and randomizes the week of the month each hotel is asked about. In a large stratum, with no great discrepancies in the number of beds per hotel, this should give a representative sample of beds over the different days of the month. However, strata are not always very large, hotels are very different in size and then there is non response.

The net result is that the number of bed-nights surveyed may not be uniformly spread over the days of the month. Weekends may be over or under represented, and as we have seen (Figure 2, above) weekend days are very different from work days: this will likely lead to over or under estimation of the monthly totals.

Consider now the imputation method. When we fill the missing cells in Table 2 for day *t,* we do it with information from donors for the same day *t. It doesn't matter anymore that weekends are over or under represented, for in all cases missing days will be imputed using observed occupancies in those days, rather than using the average monthly value.*

From that perspective, we might expect closer to unbiased results from the imputation method. The downside of it is that some days of the month may be very thinly sampled, and the imputed values for all open hotels may depend on observations of just one or two hotels, if the stratum is small.

As a final comment, one of the shortcomings with the imputation method is that, as it stands, makes no provision for the estimation of sampling errors. This is a common feature of single imputation methods, and the very reason why multiple imputation came into existence. On the other hand, it is hard to see how a multiple imputation model could be constructed for the problem at hand which would take into account that several, mutually constrained variables have to be imputed.

Disclaimer

The present work is a development in methodology whose status at present is that of a mature experiment. It may or it may not be implemented in the future exactly in this way, either as a complement or as a replacement of current methodology. For a final, authoritative description of the methodology used at each time in the EETR the user should turn to the Methodological Note, available from http://www.eustat.es.

References

Coke, G., Tsao, Min. (2010) Random effects mixture models for clustering electrical load series, *Journal of Time Series Analysis,* 31, 451-464.

Few, S. (2008) Time on the Horizon, *Visual Business Intelligence Newsletter*, 1-7.

Genolini, C., Falissard, B. (2010) KmL: k-means for longitudinal data, *Computational Statistics,* 25, 317-328.

Gower, J. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857-871.

Hastie, T.J., Tibshirani, R.J. (1995) *Generalized Additive Models.* Chapman & Hall, London.

Liao, T. W. (2005) Clustering of time series data—a survey. *Pattern Recognition*, 38, 1857-1874.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, Wiley.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Shinghal, A., Seborg, D.E. (2006) Clustering multivariate time-series data. *Journal of Chemometrics,* 19, 427-