

Regional Information Systems for Tourism Data: the experience of Piedmont¹

Roberto Fontana

Politecnico di Torino, Dipartimento di Matematica, roberto.fontana@polito.it

Abstract: The paper describes how the data concerning accommodation structures are collected, organized and used by *Regione Piemonte*. A short description of some of the current research activities is provided.

Keywords: Tourism statistics, Official Statistics, Imputation of missing data

1. The collection process

The collection process is carried on under the supervision of the *Assessorato al Turismo della Regione Piemonte*, supported by the *Osservatorio Turistico della Regione Piemonte* (OTRP) and involves all the accommodation structures of Piedmont (there were 4.823 structures at the end of 2008) . It is made by two parts:

1. on one side, with the support of the local touristic agencies, the administrative data and the list of services and equipments provided by each structure are collected;
2. on the other side, in co-operation with the provincial statistics offices, monthly tourism flows registered by each structure are acquired. This census is based on a monthly basis and takes 15 months to be fully completed: it starts in January and it finishes at the end of March of the next year, when the provincial statistics offices certify and make final the data that have collected.

As the result of the completion of the entire process, for each accommodation structure, the following data are available:

- name, address, phone, fax, email, website
- type and category (e.g. “3-star hotel”, “bed&breakfast”, ...);
- size, in terms of number of rooms, beds and bathrooms;
- available services and equipments, like “air-conditioning” or “pets allowed”; (there are around 200 of such “yes or no” variables);
- opening/closure periods and prizes;
- arrivals and bednights classified according to the origin of the tourist (Foreign Country or Italian Region)

With respect to tourism flows, each structure can transmit its own data or by surface mail or, since the end of 2007, using an on-line service (TUAP). In the first case data must be typed into the database while, in the latter case, data become immediately available and are also daily based.

¹ The present paper is an output of the “Analisi statistica dei flussi turistici” project, carried on the Department of Mathematics of Politecnico di Torino and funded by Sviluppo Piemonte Turismo (Torino)

2. Data analysis for operators

2.1 The database

All the regional data are stored into a database (TURF), maintained by CSI Piemonte. TURF represents the starting point for all the uses of these data. It stores all the data from 1991. With respect to tourism flows, there are around 1.650.000 records; each record provides arrivals and bednights of an accommodation structure, classified according to the origin of the tourist corresponding to a given year and month.

All the data, apart those concerning tourism flows, are used to build the directory of all the accommodation structures in Piedmont. The directory is also available on-line from some websites (e.g. www.torinopiemonte.com).

The data concerning tourism flows and some of the main characteristics of the accommodation structures are made available to the wide public via an online service. For interested researchers, it is possible to freely register and then to extract a file, in comma separated value (.csv) format, that contains tourism flows referred to some user-defined aggregation. It is not possible to extract data concerning a single structure (or a small subset of structures) for privacy preservation. Some basic statistics and graphical representations are automatically generated on the basis of the selected data.

Finally, once per year, in April, OTRP publishes a report, made of graphs and tables, that focuses on trends of tourism flows in recent years, for different sub-areas of the Piedmont region and for the different origins of the tourists.

2.2 Goals of data analysis

Tourism data can be analyzed for different objectives. For example for an entrepreneur that is interested in opening a new hotel in a certain area is useful to know how many hotels there are in the surroundings, how they perform and so on. Or for the major of a city it is important to know the number of tourists to properly organize waste collection and water and energy supply. For *Regione Piemonte*, one relevant goal is to assess if the adopted policies to help the tourist industry are giving the expected benefits. In this context, one of the key indicator is represented by tourism flows. Indeed bednights represent the sales that an accommodation structure has done in a certain time period. Figure 1 reports the yearly bednights series for the period between 1991 and 2008. It is evident the increasing trend since 2000 (in 2008 bednights are 43% more than in 2000).

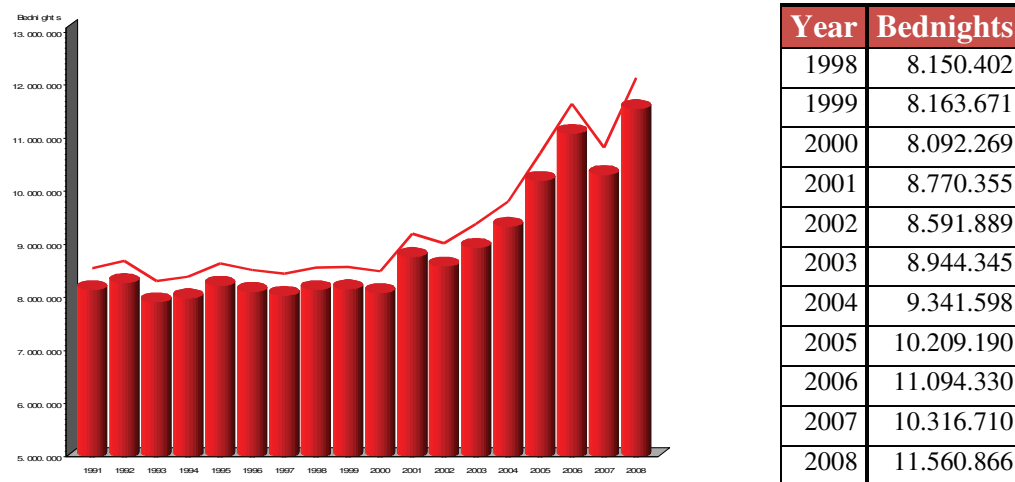


Figure 1: Bednights vs Year (1991 -2008)

2.3 Critical aspects concerning tourism flows data

The main critical issues concerning tourism flows data are timeliness and missing data. With respect to 2008, the analysis that has been done using the variable in the database that registers the date of insertion of a single record into the database itself, provides a mean value of around 5 months as an estimate of the time needed to collect the 90% of the data of a certain reference month (see column “P90” in Table 1). This delay, even if is decreasing over time (it was around 7 months in 2007), is often considered too large by the operators working in this sector.

The other problem concerns missing data due to the fact that some accommodation structures do not transmit their own data. It follows that the global values underestimate the size of this component of the tourism industry. Besides that, the difference between the response rates of different years, even if it is not very high, makes difficult to interpret data, especially for small geographical areas.

More formally, if T is the random variable that, for a given accommodation structure, measures the time that elapses between the end of a certain month and the insertion in the database, T *often* is greater than 10 days (the time limit prescribed by the law) and *sometimes* becomes $+\infty$, i.e. missing. *Often* and *sometimes* can be quantified specifying the reference month and the area. As an example, at regional level for the year 2008, the percentage of missing values is close to 30% and some statistics for the delay times are as in the following table 1.

Delay T [days]			
Month	Median	P90	Max
1	37	200	402
2	30	194	374
3	36	175	347
4	27	158	317

Delay T [days]			
Month	Median	P90	Max
5	19	156	288
6	24	133	259
7	24	185	228
8	21	164	197

Delay T [days]			
Month	Median	P90	Max
9	17	132	167
10	20	115	136
11	22	80	106
12	21	57	75

Table 1: Delays for tourism flows collection (Year 2008)

2.4 Statistical models for tourism flows prediction

The problem to estimate the missing data, that are present at a certain time in the database, has been faced during the INTERREGIII ALCOTRA project “Statistical models for tourism flows prediction”. Briefly, generalized linear models have been used to predict missing net occupancy rates, where net occupancy rates (R) are the ratio between the bednights (B) and the total capacity of an accommodation structure (C), that is the product between the number of days in which the structure is open and the number of beds. For a certain structure, $R = \frac{B}{C}$ can be computed as soon as the structure transmits the bednights B that it has registered in that month. R has been modelled (Mc Cullagh *et al.*, 1989) in the framework of linear models (LMs) and generalized linear models (GLMs), as a function of time (T), of the geographical area (G) and of the type of the accommodation structure (A), $\hat{R} = f(T, G, A)$. In particular binomial GLMs have been used to predict missing data, that is the net occupancy rate of the accommodation structures that have not yet provided their data. The number of bednights for an accommodation structures that has missing flows has been estimated using $\hat{B} = \hat{R} \cdot C$.

Finally, if we denote by $\mathbb{T}$ the set of all the accommodation structures that are open in a certain month, the total number of bednights for a certain area and a certain month (B_T) is estimated summing up the values of bednights corresponding to the accommodation structures that have provided their data ($\mathbb{A} \subseteq \mathbb{T}$) with the estimates of bednights corresponding to the accommodation structures that do not have provided their data ($\overline{\mathbb{A}} \subseteq \mathbb{T}, \mathbb{A} \cup \overline{\mathbb{A}} = \mathbb{T}$): $B_T = \sum_{\mathbb{A}} B + \sum_{\overline{\mathbb{A}}} \hat{B}$. An approximation of the variance of B_T has been derived under the standard hypothesis $R = f(T, G, A) + \varepsilon$ with $\varepsilon \sim \text{IIND}(0, \sigma^2)$. For a structure for which the bednights are missing at a certain time, it will be possible to compare the estimate, obtained with this method, with the true value, as soon as the structure will transmit its data.

The methodology has been implemented by CSI Piemonte and it is one of the tools used by the OTRP in the day by day activity. Current research is aimed at comparing the proposed approach with other methodologies to estimate B_T (Gismondi, 2007).

3. Conclusions

The database of *Regione Piemonte* that, for each accommodation structure, stores administrative data, equipments/services and tourism flows is one of the key tool for tourism analysis. A part of it is included in the national official data bank maintained by the National Institute of Statistics (Istat) and provides the Official Statistics of tourism flows in Piedmont. Some critical issues exist. Data imputation methodologies and new web-based tools for data collection are improving the situation concerning timeliness and representativeness. Since the end of 2007 an increasing group of accommodation structures is providing their tourism flows on a daily basis. This opens the way to a systematic analysis of the impact of events on tourism flows.

4. Acknowledgements

The author wishes to thank Giovanni Pistone (*DIMAT, Politecnico di Torino*), Maria Elena Rossi (*Sviluppo Piemonte Turismo*), Cristina Bergonzo, Livia Falomo, Valeria Varetti (*Osservatorio Turistico della Regione Piemonte*) and Serena Chiarle (CSI Piemonte) for their continuous support and the *Direzione Cultura, Turismo e Sport della Regione Piemonte* for having provided to the Politecnico research group the direct access to the database of *Regione Piemonte* (individual data). Data analysis has been performed using SAS.

References

- MC CULLAGH, P., NELDER, J.A (1989): Generalized Linear Models, Chapman and Hall, 2nd Edition.
- FONTANA, R., PISTONE, G., Statistical models for tourism flows estimation. *La Matematica e le sue Applicazioni*, n. 17, 2008 hard copy ISSN 1974-3041, on-line ISSN 1974-305X
- GISMONDI R.: Quick estimation of tourist nights spent in Italy, *Statistical Methods & Applications* (2007) 16:141-168
- SAS Institute Inc. 2004. SAS/Stat ® 9.1 User's Guide. Cary, NC: SAS Institute Inc.