hard copy ISSN 1974-3041
on-line ISSN 1974-305X

La Matematica
e le sue Applicazioni
n. 6, 2010

Detection of high impact tourist events from occupancy data. An application to Piemonte, Italy.

R. Fontana, G. Pistone

Quaderni del
Dipartimento di Matematica
Politecnico di Torino
Corso Duca degli Abruzzi, 24 – 10129 Torino – Italia

Edizioni C.L.U.T. - Torino
Corso Duca degli Abruzzi, 24
10129 Torino
Tel. 011 564 79 80 - Fax. 011 54 21 92

La Matematica e le sue Applicazioni
hard copy ISSN 1974-3041
on-line ISSN 1974-305X
Direttore: Claudio Canuto

Esemplare fuori commercio
accettato nel mese di Maggio 2010

Detection of high impact tourist events from occupancy data

An application to Piemonte, Italy

Roberto Fontana · **Giovanni Pistone**

Abstract Tourism is a complex and highly competitive sector. In Italy, public institutions play a crucial role in supporting events that can increase tourism flows. The current world economic crisis makes even more necessary than in the past to adopt an informed decision making process for resource allocation. The statistical methodology that is described in this paper analyses daily tourism flows in Piemonte as collected by the Italian National Institute of Statistics, ISTAT, under the 'Occupancy in collective accommodation establishments" census. The days of the year in which the registered bednights are expected to be strongly correlated with events like public holidays, commercial fairs and sport competitions are identified. The methodology has been implemented using SAS Forecast Server.

Keywords Time series · Tourism statistics · Statistical computing

1 Introduction

Regione Piemonte officials wish to exploit the considerable amount of data on tourism flows that are currently available to support their own decision process. In this context, tourism flows that are collected under the "Occupancy in collective accommodation establishments - Movimento dei clienti negli esercizi ricettivi" census are extremely important. As described in [4], they are the input of a methodology that allows, among the other, to estimate the total tourism bednights while minimizing the effect of non-respondent accommodation structures. In this paper, daily tourism flows are used to identify the days of the year that are expected to be connected with high impact events. The work was carried on under a project funded by Sviluppo Piemonte Turismo, Turin, Italy and has been implemented using SAS Forecast Server.

R. Fontana
DIMAT Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino Italy,
Tel.: +39-090-7504
E-mail: roberto.fontana@polito.it

G.Pistone
DIMAT Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino Italy and
Statistics Initiative, Collegio Carlo Alberto, Via Real Collegio 30, 10024, Moncalieri Italy,
Tel.: +39-090-7505
E-mail: giovanni.pistone@gmail.com

2 From Monthly to Daily Tourism Flows

The Italian National Institute of Statistics, ISTAT, mantains, among data banks, a system of demographic, social, environmental and economic indicators referring to geographical areas, regions, provinces and municipalities (Territorial indicators). Indicators are grouped into 15 information areas, including transportation and tourism. With respect to tourism sector, the data on tourism flows come from the "Occupancy in collective accommodation establishments - Movimento dei clienti negli esercizi ricettivi" census. Briefly, each accommodation structure registers on a predefined questionnaire its own data on arrival and bednights and makes them available to ISTAT through local tourism agencies. It is important to point out that data are registered by the accommodation structure on a *daily* basis (questionnaire "ISTAT C/59" or "Tavole di spoglio A1 e A2") but local statistical offices summaryze them on a *monthly* basis (form "MOV/C") before the transmission to ISTAT. The main reason is that, apart from some exceptions, data from accommodation structures are written on paper and so a signicant work should be done to type them into a dataset for the statistical analysis.

We now describe in more detail how the data are collected in Piemonte. The collection process involves all the accommodation structures (as a reference there were 4.719 structures at the end of 2007) and is carried on under the supervision of the Assessorato al Turismo della Regione Piemonte with the support of all the provincial statistics offices (there are 8 provincial statistics offices in Piedmont). Every month each accommodation structure has to send to its provincial office the total daily tourism flows (arrivals and bednights), classified according to the country of origin of tourist. Total means that the individual data (i.e. the data referring to a single tourist) are summed up over every day of all the month for privacy preservation. Individual data are made available only to police for security reasons. Every year, in March, the provincial statistics offices certify and make final the data that have collected for the previous year. After that, the Assessorato al Turismo della Regione Piemonte publishes a report that summarizes the main trends that have been registered in Piemonte in the previous year. This report provides the official figures of tourism in Piemonte. Finally the data are transmitted to ISTAT.

In Piemonte, since the end of 2007, each structure can transmit its own data or, as usual, by surface mail or using an on-line web based service (TUAP). TUAP makes a significant improvement of the collection process. In particular daily tourism flows are typed into the database by each accommodation structure that uses this service and so they become immediately available for the statistical analysis.

3 The available data

We have analysed the daily tourism flows for the year 2008. The data are to be considered provisional even if they are very close to their final release. They refer to 495 accommodation structures out of a total of $4,666$, around 10%. For each type of accommodation, Table 1 compares 1 the number of structures for which the daily tourism flows are electronically available (shortly referred as TUAP structures) with the total number of structures.

We point out that we are considering a self-determined sample. Indeed, as we said, it is up to each accommodation structure to decide to use or not to use the on-line service TUAP. Anyhow it appears that the sample has a distribution among types of accommodation close to that of the population of all the accommodation structures. Standard chi square goodness

Type of accommodation	TUAP structures	Total structures	TUAP/Total [%]
Hotel - Albergo	155	1450	10.7
Hotel - Albergo Residenziale	11	73	15.1
Camping - Campeggio	14	164	8.5
Holiday village - Villaggio Turistico	1	5	20.0
Holiday home - Casa per Ferie	37	209	17.7
Youth hostel - Ostello per la gioventú	1	26	3.8
Mountain dew - Rifugio Alpino	7	157	4.5
Mountain dew - Rifugio Escursionistico	3	50	6.0
Mountain dew - Bivacco Fisso	0	34	0.0
Farm holidays - Alloggio Agriturismo	62	691	9.0
Room rental - Affittacamere	35	409	8.6
Room rental with restaurant - Affittacamere con Ristorante	13	182	7.1
Holiday home - Casa o Appartamento per Vacanze	44	230	19.1
Bed & Breakfast - Alloggio in Locazione - Bed & Breakfast	112	982	11.4
Holiday home - Alloggi Vacanze	0	4	0.0
TOTAL	495	4,666	10.6%

Table 1 Number of accommodation structures vs type

of fit analysis, that has been performed without “Holiday village - Villaggio Turistico” and “Holiday home - Alloggi Vacanze” for which the expected counts are less than one, points out that:

- “Mountain dew - Rifugio Alpino” and “Mountain dew - Bivacco Fisso” are under sampled;
- “Holiday home - Casa per Ferie” and “Holiday home - Casa o Appartamento per Vacanze” are over sampled.

If we exclude these accommodation categories, we obtain a good agreement between the observed and expected counts ($\chi^2 = 9,60$ with 8 degrees of freedom and p-value=0.294). Besides that, the mean value of bednights computed using the daily structures is close to that computed using all the structures of the population.

4 The analysis

The main goal of the work is to use the available daily tourism flows to point out major events, where major means with a significant impacts on bednights. Moreover the methodolgy should be easily computable and usable for a larger number of structures (some thousands) because the users of TUAP are quickly increasing.

We have translated this goal into two related but different objectives:

1. to find the days of the year in which something unusual has happened in terms of bednights;
2. to associate these days with events.

We have developed a statistical methodology to reach the first objective, as we describe in the next sections. The second step is based on the association between the days that have been found in step 1 and a calendar of events, that has been made available by Regione Piemonte. We point out that, in this work, *events* have a wide meaning because they include public holidays, commercial fairs and sport competitions. We use SAS Forecast Server - SAS Forecast Studio 1.4 to do statistical computing.

The methodology considers all the available daily bednights time series in the sample. As we said in the previous Section 3 there are 495 time series, each one with a maximum of 366 values because the 2008 was a leap year. Missing values correspond to days in which the accommodation structure was closed. We denote these time series by $Y_t^{(i)}$, $i = 1, \ldots, 495$; $t = 1, \ldots, 366$.

Briefly, for each time series $Y_t^{(i)}, i = 1, \ldots, 495$, the procedure is made by two steps.

1. We search for the *best model* in a wide class of model types $\mathscr{M}$. The class $\mathscr{M}$ include ARIMA, exponential smoothing models and Intermittent Demand Models, see [1]. The *best model* is one of the models of $\mathscr{M}$ for which the Mean Average Percentage Error (MAPE) is minimum; if we denote by $\hat{Y}_t^{(i,m)}$ the values of the bednights predicted by a model $m \in \mathscr{M}$, the best model $m_*^{(i)}$ is such that

$$m_*^{(i)} \in \underset{m \in \mathscr{M}}{\operatorname{argmin}} \left(\frac{1}{n_i} \sum_{i=1}^{n_i} \left| \frac{Y_t^{(i)} - \hat{Y}_t^{(i,m)}}{Y_t^{(i)}} \right| \right)$$

   The value n_i, $n_i \leq 366$, is the number of days for which bednights are non missing for structure i and the summation is extended to all the n_i non missing values. In principle there could be more than one model $m_*^{(i)}$ that minimizes the MAPE for a given time series. Even if this case is unlikely from a practical point of view, if it should happen, we simply pick up one of them randomly. We choose to minimize MAPE because this criterion helps in finding models that provide a good representation of the observed data. Figure 4 shows the original time series (circles) and the predicted values (continous line) for one of the accommodation structures of the sample.
2. Using $m_*^{(i)}$, we store in a dataset all the days for which the difference between the observed value $Y_t^{(i)}$ and the predicted value $\hat{Y}_t^{(i,m_*^{(i)})}$ lies outside the 95% individual prediction interval. Indeed these large residuals can represent days that the model can not properly explain because something of unusual has happened. Figure 4 shows the lower and upper 95% individual prediction interval (red lines) for the same accommodation structure of Figure 4.

Then we analyse the dataset of these large residuals. The basic idea is that the dataset contains, for each accommodation structure, the days which might be considered *anomalous* by the structure itself. It should be noted that this kind of data can be easily discussed with the management of each structure. It is evident that there are two different types of large residuals.

– The residuals that are greater than the upper bound of the prediction interval. They are connected to days for which bednights have been superior to the standard performance; we briefly refer to these days as *positive* days.
– The residuals that are smaller than the lower bound of the prediction interval. They are connected to days for which bednights have been inferior to the standard performance; we briefly refer to these days as *negative* days.

Now we count, for each day t, the number of structures for which the day t is positive or negative; we obtain $n_t^{(+)}$, $t = 0, \ldots, 366$ and $n_t^{(-)}$, $t = 0, \ldots, 366$, respectively. We observe that the number of structures for which the value of bednights is not missing is not constantly equal to the size of the sample all over the year. The main reason is, as we said, that not all

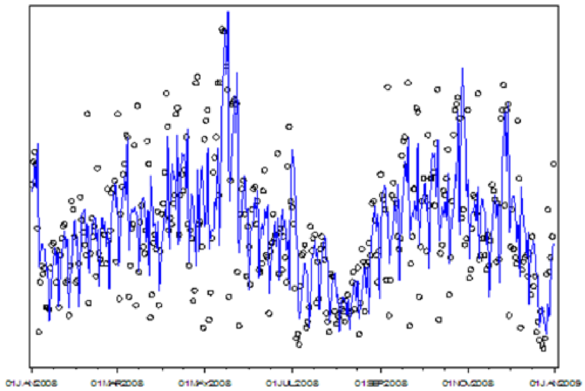

Fig. 1 Comparison between observed and predicted values for a single accommodation structure

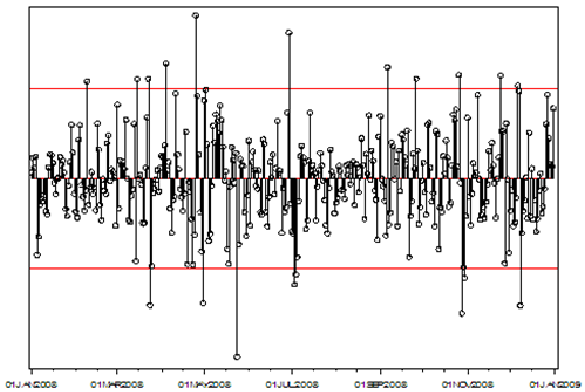

Fig. 2 Residuals and 95% individual prediction interval for the accommodation structure of 4

the accommodation structures stay open for all the days of the year. We have therefore normalized $n_t^{(+)}$ and $n_t^{(-)}$, dividing them by the number r_t of the accommodation structures for which bednights are not missing for day t. Being $0 \leq n_t^{(+)}, n_t^{(-)} \leq r_t$, we obtain the time series of the proportions of accommodation structures for which the day is positive or negative

$$p_t^{(+)} = \frac{n_t^{(+)}}{r_t} \text{ and } p_t^{(-)} = \frac{n_t^{(-)}}{r_t}$$

For a given day t, $p_t^{(+)}$ (resp. $p_t^{(-)}$) is an unbiased estimate of the proportion of accommodation structures for which the day t is positive (resp. negative) on the entire population of accommodation structures that are open on day t. The population is made by the structures that used TUAP plus the structures that used the traditional paper form. We denote by N_t the size of the population, $0 \leq r_t \leq N_t \leq 4,666$ and by $\pi_t^{(+)}$ (resp. $\pi_t^{(-)}$) the unknown proportions of accommodation structures for which the day t is positive (resp. negative) defined over all the population. We know from the literature, see [5] and [2], that an estimate of the variance $v_t^{(+)}$ of the estimator of the proportion of accommodation structures for which the day t is positive is

$$v_t^{(+)} = \frac{1-f}{r_t-1} p_t^{(+)}(1-p_t^{(+)})$$

where $f = \frac{r_t}{N_t}$ is the fraction of sampling. An analogous expression holds for the variance of $v_t^{(-)}$. This allows to build confidence interval for the unknown proportions $\pi_t^{(+)}$ and $\pi_t^{(-)}$. An approximate $(1-\alpha)\%$ confidence interval for $\pi_t^{(+)}$ is

$$(l_t, u_t) = (p_t^{(+)} - z_{1-\frac{\alpha}{2}}\sqrt{v_t^{(+)}} + \frac{1}{2r_t}, p_t^{(+)} + z_{1-\frac{\alpha}{2}}\sqrt{v_t^{(+)}} + \frac{1}{2r_t})$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})\%$ percentile of the standardized normal random variable. When $r_t p_t^{(+)}(1-p_t^{(+)}) > 10$, as it often happens in our case, the approximation is usually quite good. We compute 95% approximate confidence intervals, taking $z_{0.975} = 1.96$.

5 Results

The maximum value of $p_t^{(+)}$ is $0,26$ and has been obtained for the 25th of April, the Anniversary of Liberation. It is well known this day is important for tourism because it is close to another public holiday, the Labour Day (the 1st of May) and is in late spring, usually a good time from the point of view of weather conditions. The following Table 2 reports the top 12 positive days in Piemonte. It is worthwhile to point out that a completely automatic processing has been able to identify days that are connected to the most relevant public holidays as well as to some very well known international fairs. This has increased the confidence of the political decision makers in the methodology.

Date	Week Day	Description	Event(s)	$p_t^{(+)}$	l_t	u_t
April, 25	Friday	Anniversary of Liberation		0,26	0,21	0,30
December, 31	Wednesday	New Year's Eve		0,21	0,17	0,26
May, 1	Thursday	Labour Day		0,17	0,13	0,21
March, 22	Saturday	Easter Holidays		0,17	0,13	0,21
December, 6	Saturday	Immaculate Conception, extended holiday		0,15	0,11	0,19
October, 4	Saturday		78^{th} International white truffle fair of Alba	0,14	0,10	0,17
March, 21	Friday	Easter Holidays		0,13	0,09	0,17
March, 23	Sunday	Easter Sunday		0,13	0,09	0,17
May, 2	Friday	Labour Day, extended holiday		0,13	0,09	0,16
November, 1	Saturday	All Saints	78^{th} International white truffle fair of Alba	0,11	0,08	0,15
May, 31	Saturday	Republic Day, extended holiday		0,11	0,08	0,15
October, 25	Saturday		Salone del Gusto - Terra Madre	0,11	0,08	0,14

Table 2 The top 12 positive days in Piemonte

The analysis can be easily repeated limiting to the accommodation structures of a certain geographical area. From the point of view of tourism, Piemonte can be partitioned into five subregions, see [3].

1. Metropolitan Areas, including Turin and medium sized towns,
2. Lakes, a beautiful natural district in north-east of Piedmont, close to Switzerland and Milan,
3. Mountains, including the famous skiing resorts of the XX Winter Olympic Games,
4. Hills, where food and wine are the key feature of the offer,
5. Other, a relatively small category that contains all the remaining areas for which one of the previous definition does not apply.

The following Table 3 reports, for each subregion, the day that appears to be the most positive

Sub-region	Date	Week Day	Description	Event(s)	$p_t^{(+)}$
Metropolitan Areas	April, 25	Friday	Anniversary of Liberation		0,32
Lakes	May, 1	Thursday	Labour Day		0,48
Mountains	December, 31	Wednesday	New Year's Eve		0,34
Hills	May, 1	Thursday	Labour Day		0,25
Other	October, 4	Saturday		78^{th} International white truffle fair of Alba	0,22

Table 3 The most positive days for each subregion of Piemonte

We now look at the negative days. It is clear that from a methodological point of view there is no difference with positive days. We find that the most "voted" days are those that come immediately after a positive day. We explain this phenomenon observing that, when an event occurs, the dynamic of the observed flows is faster than that of the predicted flows because of the smoothing effect of the model. Apart from this *day-after effect* some negative days result in correspondence of bad weather conditions.

6 Conclusions

The methodology provided very good results and can support the discussion with the operators. It has been implemented using SAS Forecast server and this makes it easily scalable to a larger number of TUAP structures, to a wider class of time series models, including user defined and also to fitting criteria different from MAPE. SAS system is currently used by the Osservatorio Turistico della Regione Piemonte to do statistical analysis. The result of this work is now part of the statistical software tools of the Osservatorio Turistico della Regione Piemonte.

Further research developments are also connected to consider the *v*alues of the residuals and not only if they are outside the prediction interval or not. Some preliminary analysis shows that this approach could lead to a more accurate system of events evaluation. Finally it should be pointed out that it is not possible to directly transform bednights into turnover, because actual prices are not known.

7 Acknowledgements

The present paper is an output of the ”Analisi statistica dei flussi turistici” project, carried on by the Department of Mathematics of Politecnico di Torino and funded by Sviluppo Piemonte Turismo (Torino). The authors thank Marzia Baracchino (Direzione Cultura, Turismo e Sport della Regione Piemonte), Cristina Bergonzo, Livia Falomo e Maria Elena.Rossi (Sviluppo Piemonte Turismo) for their continuous cooperation and Valeria Varetti (Politecnico di Torino) for her support in software development. Finally the author thank Daniela Formento, director of the Direzione Cultura, Turismo e Sport della Regione Piemonte for having provided to the Politecnico research group the direct access to the database (individual data).

References

1. SAS Forecast Studio User's Guide, Version 1.4. Cary, NC: SAS Institute, 1 edn. (2007)
2. Frosini, B.V., Montinaro, M., Nicolini, G.: Il campionamento da popolazioni finite. Statistica e metodi quantitativi. UTET Libreria (1994)
3. R., F.: The use of correspondence analysis to study daily tourism flows. STATISTICA APPLICATA **20 / 2** (2008)
4. R., F., G, P.: Anticipating italian census tourism data before their official release: a first solution and implementation to piemonte, italy. THE INTERNATIONAL JOURNAL OF TOURISM RESEARCH DOI 10.1002/jtr.766. URL http://dx.medra.org/10.1002/jtr.766. In press
5. Valliant, R., Dorfman, A.H., Royall, R.M.: Finite population sampling and inference. Wiley Series in Probability and Statistics: Survey Methodology Section. Wiley-Interscience, New York (2000). A prediction approach