www.ejthr.com

European Journal of Tourism, Hospitality and Recreation
Vol. 3, Issue 3, pp. 9-19, 2012

Printed in Portugal

DETECTION OF HIGH IMPACT TOURIST EVENTS IN PIEMONTE REGION, ITALY

Roberto Fontana
DISMA Politecnico di Torino, Italy

Giovanni Pistone
Collegio Carlo Alberto, Italy

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 it 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. A fully unsupervised methodology for Piemonte Region (Italy) tourism events detection has been developed and implemented using SAS Forecast Server. The methodology can be transferred, without any modification, to any of the 20 Italian Regions.
Keywords: Tourism management, Time series, Tourism statistics, Statistical computing.

RESUMEN: El turismo es un sector muy competitivo y complejo. En Italia, las instituciones públicas tienen un rol crucial en el apoyo a eventos que puedan aumentar los flujos turísticos. La actual crisis económica mundial vuelve aún más necesaria la adopción de un proceso de tomada de decisiones esclarecidas relativamente a la afectación de recursos. La metodología estadística que es descrita en este artículo científico analiza los flujos turísticos diarios en Piamonte, obtenidos por el Instituto Nacional de Estadística Italiano, ISTAT, durante los censos "Ocupación en estabelecimientos de alojamiento turístico colectivos". Los días del año en que se espera que las dormidas registradas estén fuertemente relacionadas con eventos, tales como feriados nacionales, ferias de negocios y competiciones deportivas, están identificados. Fue desarrollada e implementada una metodología de detección de eventos turísticos no supervisada para la Región de Piamonte (Italia) utilizando el SAS Forecast Server. La metodología puede ser transferida, sin cualquier modificación, para cualquier una de las 20 Regiones Italianas. **Palabras-clave**: gestión del turismo, series cronológicas, estadística turística, computación estadística.

RESUMO: O turismo é um sector altamente competitivo e complexo. Em Itália, as instituições públicas desempenham um papel crucial no apoio a eventos que possam aumentar os fluxos turísticos. A atual crise económica mundial torna ainda mais necessária a adoção de um processo de tomada de decisões elucidativo relativamente à afetação de recursos. A metodologia estatística que é descrita neste artigo científico analisa os fluxos turísticos diários em Piemonte, recolhidos pelo Instituto Nacional de Estatística Italiano, ISTAT, no decorrer do censos "Ocupação em estabelecimentos de alojamento turístico coletivos". Os dias do ano

Authors' address: **Roberto Fontana**: DISMA Politecnico di Torino - Corso Duca degli Abruzzi 24, 10129 Torino, Italy; roberto.fontana@polito.it; **Giovanni Pistone:** Statistics Initiative, Collegio Carlo Alberto - Via Real Collegio 30, 10024, Moncalieri, Italy; giovanni.pistone@carloalberto.org

em que se espera que as dormidas registadas estejam fortemente relacionadas com eventos, tais como feriados nacionais, feiras de negócios e competições desportivas, estão identificados. Foi desenvolvida e implementada uma metodologia de deteção de eventos turísticos não supervisionada para a Região de Piemonte (Itália) utilizando o SAS Forecast Server. A metodologia pode ser transferida, sem qualquer modificação, para qualquer uma das 20 Regiões Italianas. **Palavras-chave**: gestão do turismo, séries cronológicas, estatística turística, computação estatística.

INTRODUCTION

The regions of Italy are the first-level administrative divisions of the state, constituting its first NUTS (Nomenclature of Territorial Units for Statistics) administrative level. There are twenty regions and they have legislative power in tourism. 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 (Fontana & Pistone, (2010), they are the input of a methodology that allows, among others, 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.

FROM MONTHLY TO DAILY TOURISM FLOWS

The Italian National Institute of Statistics, ISTAT, maintains, 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 the 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 summarize 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 significant 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 Piemonte). 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 of 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 online 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.

THE AVAILABLE DATA

We have analyzed 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 the number of structures for which the daily tourism flows are electronically available (shortly referred as TUAP structures) with the total number of structures.

Table 1. Number of accommodation structures vs type (2008)

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

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

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

If we exclude these accommodation categories, we obtain a good agreement between the observed and expected counts (x^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.

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 methodology 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:

- to find the days of the year in which something unusual has happened in terms of bednights;
- 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 (SAS Institute Inc, 2007) 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 , i = 1,…,495; t = 1,…,366.

Briefly, for each time series , i = 1,…,495, the procedure is made by two steps.

- We search for the best model in a wide class of model types M. The class M include ARIMA, exponential smoothing models and Intermittent Demand Models, see SAS (2007). The best model is one of the models of M for which the Mean Average Percentage Error (MAPE) is minimum; if we denote by the values of the bednights predicted by a model mM, the best model is such that the value , is the number of days for which bednights are non missing for structure *i* and the summation is extended to all the non missing values. In principle, there could be more than one model that minimizes the Mean Average Percentage Error (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 the Mean Average Percentage Error (MAPE) because this criterion helps in finding models that provide a good representation of the observed data.

Figure 1 shows the original time series (circles) and the predicted values (continuous line) for one of the accommodation structures of the sample.

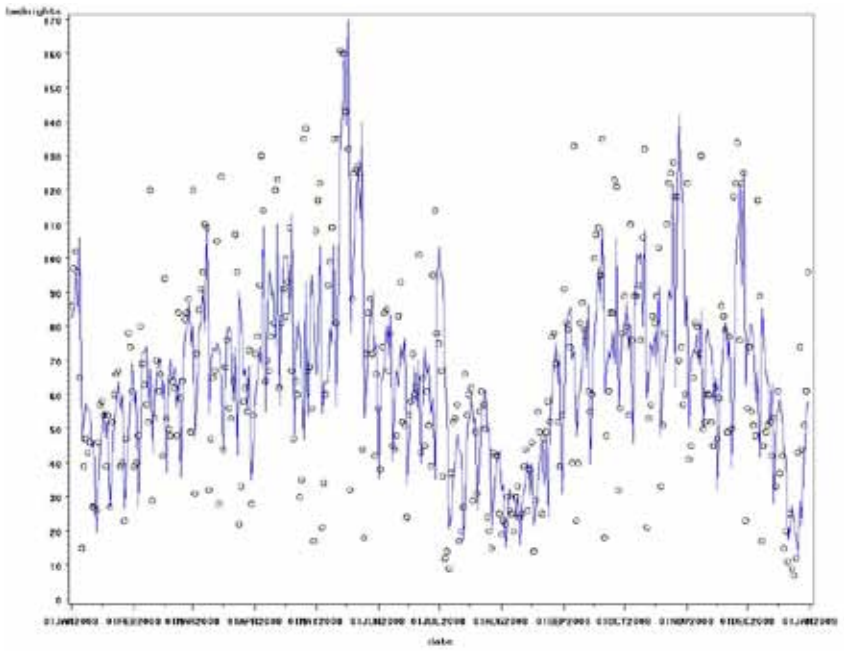

Figure 1. Comparison between observed and predicted values for a single accommodation structure

• Using , we store in a dataset all the days for which the difference between the observed value and the predicted value lies outside the 95% individual prediction interval. Indeed these large residuals can represent days that the model cannot properly explain because something of unusual has happened. Figure 2 shows the lower and upper 95% individual prediction interval (dashed lines) for the same accommodation structure of Figure 1.

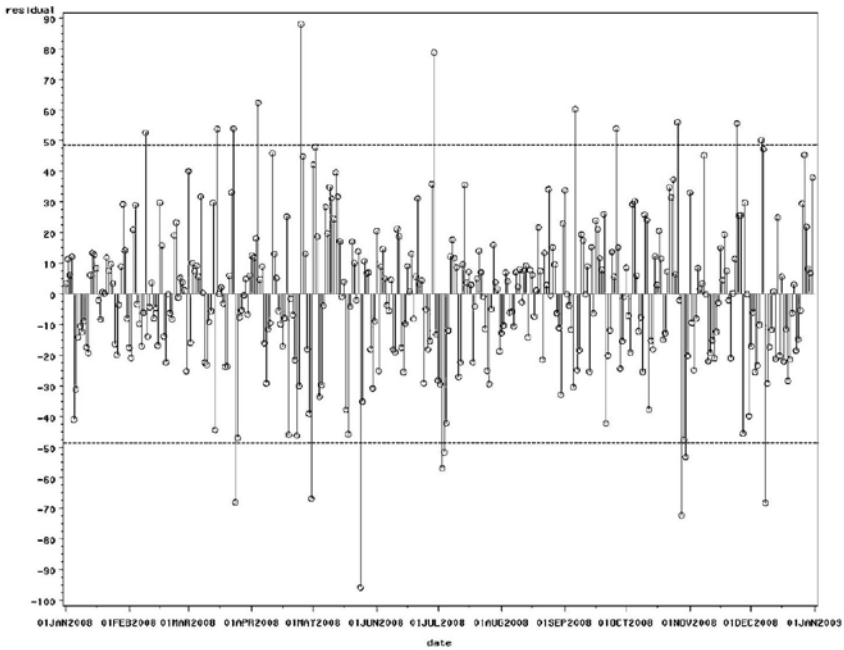

Figure 2: Residuals and 95% individual prediction interval for the accommodation structure of Figure 1

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 $nt^{(+)}$, t=0,…,366 and $nt^{(-)}$, t=0,…,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 the accommodation structures stay open for all the days of the year. We have therefore normalized $nt^{(+)}$ and $nt^{(-)}$, dividing them by the number r_t of the accommodation structures for which bednights are not missing for day t. Being $0 \leq nt^{(+)}, nt^{(-)} \leq r_t$, we obtain the time series of the proportions of accommodation structures for which the day is positive or negative

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

For a given day t , $pt^{(+)}$ (resp. $pt^{(-)}$) is an unbiased estimate of the proportion of accommodation structures for which the day is positive (resp. negative) on the entire population of accommodation structures that are open on day . The population is made by the structures that used TUAP plus the structures that used the traditional paper form. We denote by Nt the size of the population, $0 \leq r_t \leq Nt \leq 4{,}666$ and by $\pi t^{(+)}$ (resp. $\pi t^{(-)}$) the unknown proportions of accommodation structures for which the day is positive (resp. negative) defined over all the population. We know from the literature (see Valliant et al., 2000; Johnson & Wichern, 2007) that an estimate of the variance $vt^{(+)}$ of the estimator of the proportion of accommodation structures for which the day t is positive is

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

where $f = \frac{r_t}{Nt}$ is the fraction of sampling. An analogous expression holds for the variance of $vt^{(-)}$. This allows to build confidence interval for the unknown proportions $\pi t^{(+)}$ and $\pi t^{(-)}$. An approximate $(1 - a)$ confidence interval for $\pi t^{(+)}$ is

$$(lt, ut) = \left(pt^{(+)} - z_{1-\frac{a}{2}} \sqrt{vt^{(+)}} + \frac{1}{2r_t}, pt^{(+)} + z_{1-\frac{a}{2}} \sqrt{vt^{(+)}} + \frac{1}{2r_t}\right)$$

where $z_{1-\frac{a}{2}}$ is the $(1-\frac{a}{2})$ 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 $z_{0,975} = 1,96$.

RESULTS

The maximum value of $pt^{(+)}$ is 0,26 and has been obtained for the 25th of April, the Anniversary of Liberation. It is well known that this day is important for tourism because it is close to another public holiday, Labour Day (the 1st of May) and is in late spring, usually a good time from the point of view of weather conditions. 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.

Table 2. The top 12 positive days in Piemonte

Date	Week Day	Description	Event(s)	$pt^{(+)}$	lt	ut
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		78th 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	78th International white truffle fair of Alba	0.11	0.08	0.15

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 sub-regions, (see Fontana, 2008).

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 definitions does not apply.

Table 3 reports, for each sub-region, the day that appears to be the most positive.

Table 3. The most positive days for each subregion of Piemonte

Sub-region	Date	Week Day	Description	Event(s)	$pt^{(+)}$
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		78th International white truffle fair of Alba	0.22

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 dynamics of the observed flows are 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 to bad weather conditions.

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 the Mean Average Percentage Error (MAPE). SAS system is currently used by the *Osservatorio Turistico della Regione Piemonte (OTRP)* to do statistical analysis. The result of this work is now part of the statistical software tools of OTRP.

Further research developments are also connected to consider the values 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 bed-nights into turnover, because actual prices are not known.

REFERENCES

Fontana, R. (2008). The use of correspondence analysis to study daily tourism flows. *Statistica Applicata*, *20* (2), 93-101.

Fontana, R., & Pistone, G. (2010). Anticipating Italian census tourism data before their official release: a first solution and implementation to Piemonte, Italy. *The International Journal of Tourism Research, 12* (5), 472–480.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis.* (Sixth ed.). Pearson Prentice Hall, Upper Saddle River, NJ.

SAS Institute Inc (2007). *SAS Forecast Studio User's Guide, Version 1.4*. Cary, NC.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference. A prediction approach*. Wiley-Interscience, New York.

Submitted:19th January, 2012 *Accepted:30th April, 2012*
Final version:28th March, 2012 *Refereed anonymously*