Domestic tourism in Italy: a Beta regression model

LUISA STRACQUALURSI, PATRIZIA AGATI
Department of Statistics "P. Fortunati"
University of Bologna
via Belle Arti, 42, 40126 Bologna
ITALY
luisa.stracqualursi@unibo.it, patrizia.agati@unibo.it

Abstract:
Planning and implementing strategies and programs to develop tourism has long been a priority issue for policy makers and decision makers. Hence, the interest in quantitative methods for assessing the extent to which possible modifications of some determinants can affect the tourism flows of the destination country. Besides, over the last decades, in Italy as well as in many other destinations worldwide, most of policy makers have shifted their priority from the promotion of inbound tourism to the promotion of domestic tourism. In this framework, the present study builds on a regional dataset and implements a beta regression model with a twofold aim: to investigate the relevance and the statistical significance of some determinants of Italian domestic tourism, and the extent to which possible modifications in such determinants can affect the tourism flows across the Italian regions.Among the several factors, the model showed the number of accommodation facilities, the number of great shopping centres and the hospital density as statistically significant, showing a good explained variation and a reasonable root mean square error. On this basis, simulation studies were performed to assess the impact of new investments, such as new accommodation facilities and new congress centres on local tourism flows.

Key-Words: -Domestic tourism, Beta model, Tourism flow, Italian regions

1 Introduction

Tourismhasexperiencedacross the time a continued upward trend and increased competition among destinations, as each touristregionseeds to attract a maximum share from the total stock of tourists.Consequently, planning and implementing strategies and programs to develop tourism has become a major concern for policy makers and decisionmakers. Besides, over the last decades, in Italy as well as in many other destinations worldwide, most of policymakers have shifted their priority from the promotion of inbound tourism to the promotion of "domestic tourism"[7],defined as tourism involving residents of a given country travelling only within the country itself [10]. Domestic tourism accounts by far for most of this activity: it is estimated that worldwide, out of the 4.8 billion tourist arrivals per year, 4 billion (83 per cent) correspond to domestic tourism [6].
From decision makers' point of view, some relevant issues stem from this framework: why do people prefer a destination to another one? That is, in a "domestic" context, what are the factors that cause people to prefer a region to another one? What modifications in the tourism policy, what regional planning, what new investments, if any, could persuade people to change their preferences about the destination region?
The present study builds on an Italian regional dataset and implements a beta regression model with a twofold aim: to investigate the relevance and the statistical significance of some determinants of Italian domestic tourism, and assess the extent to which possible modifications in such determinants can affect people's destination preferences and, consequently, the local tourism flowsacross the Italian regions.
The paper is organized as follows. The following Sectionfocuses on some key concepts and outlines a synthetic review of beta regression model. The case-study is described and the model is fitted in Section 3. Finally, Section 4 presents a simulation study and draws some concluding remarks.

2 Conceptual Model

In this paper, domestic tourism flow in Italy is analyzed in terms of interregional tourism flows. It can be thought and graphically represented by a network, where the nodes are the regions (the first-level administrative divisions in Italy: sub-national territorial decision areas where tourism management and planning can develop) and the links are directed

connections between two nodes. By way of example, Figure 1 shows a map chart where the colors of Italian regions change gradually between the darkblue and the light blue according with the number of arrivals in the reference time period (calendar year 2012): the darker the blue color, the greater the number of tourists that visited the region in the period.

Indeed, if the relevant question is why do domestic tourists prefer a region rather than another one, the core measure we propose to analyze data is a normalized indicator of the domestic tourism inflow into each region, named *travel-in rate* (TIR): for each region, it is computed as the ratio of the number of visitor arrivals in the region to the total number of national arrivals in Italywithin aperiodof time. This measure ranges from 0 (no arrival in the region) to 1 (all arrivals concentrated in a single region): the larger the TIR, the larger the attractiveness of the region in terms of domestic tourism inflow. On this basis, it is straightforward for any decision maker to distinguish "critical" regions, characterized by low attractiveness rates, from "successful" regions, characterized by travel-in rates close to one. Then, a beta regression model can be used to investigate the relation between the regional travel-in rate and some explanatory variables (such as the number of accommodation facilities, congress centers, theme parks, etc.), in such a way to identify the statistically significant determinants that cause people to prefer a region to another one. Finally, since the final aim of the study is prediction, a cross validation procedure can be performed to validate the model: it allows to assess how the results can be generalized to an independent dataset or, in other words, how accurately the model will perform in practice.

2.1 Some issuesabout beta regression model

Beta regression model is tailored for situations where the response variable y takes on values within the real open interval (0, 1). For such variables, that tipically stem from rate and proportions, the normality assumption underlying the linear regression model is not supported: bounded range continuous variables usually display heteroscedasticity (the variance is smaller near the extremes) and asymmetry; linear fitted values could exceed the lower and upper bounds of y, resulting in invalid and misleading outcomes.

Fig.1: map of the regions of Italy – year 2012
To simplify the structure,links smaller than 300000 units are not displayed.

Instead, the beta distribution is a very flexible model for variables within the standard unit interval: its density, given by

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} I_{\{0,1\}}(y) \qquad (1)$$

can have quite different shapes depending on the values of the two parameters $p>0$, $q>0$, and can accommodate skew and asymmetry. The expected value and the variance of y are $E(y) = p/(p+q)$ and $V(y) = pq/[(p+q)^2(p+q+1)]$. For modelling purposes, a different parameterization of the beta density was proposed by Ferrari and Cribari-Neto [4] by setting $\mu = p/(p+q)$ and $\varphi = p+q$, i.e. $p = \mu\varphi$ and $q = (1-\mu)\varphi$:

$$f(y;\mu,\varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((1-\mu)\varphi)} y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1} I_{\{0,1\}}(y) \qquad (2)$$

where $0 < \mu < 1$ and $\varphi > 0$. The expected value and the variance of y, in the new parameterization, are $E(y) = \mu$ and $V(y) = \mu(1-\mu)/(1+\varphi)$, so that μ is the mean of y and φ can be regarded as a precision parameter: for fixed μ, the larger the value of φ, the smaller the variance of y.

Let $y_1, \ldots, y_j, \ldots, y_n$ be independent random variables, where each y_j, $j = 1,\ldots, n$ is Beta distributed with mean μ_j and unknown precision φ, and $(x_{j1}, \ldots, x_{ji}, \ldots, x_{jk})$ be observations on k

covariates, which are assumed as fixed and known. The beta regression model can be written as

$$g(\mu_i) = \sum_{i=1}^{k} x_{ji}\beta_i \qquad (3)$$

Where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_i \ldots, \beta_k)' \in \Re^k$ is a vector of unknown regression parameters and $g(\cdot)$ is a strictly monotonic and twice differentiable link function that maps (0, 1) to $\Re$. Several choices are possible for the link function$g(\mu)$, such as the logit function $\log[\mu/(1-\mu)]$ (the inverse cumulative distribution function of the logistic) or the probit function $\Phi^{-1}(\mu)$ (the inverse cumulative distribution function of the standard normal variable). Extensions of the beta regression model outlined above were proposed by Smithson and Verkuilen [9], Simas, Barreto-Souza and Rocha [8], Cook, Kieshnik, McCullogh [2] among others.

3 Case study and results

The analysed datacomes from various sources:

- *GeoWebStarter* - a web platform for analysis of socio-economic phenomena of territory, provided by GuglielmoTagliacarne Institute of Unioncamere(*i.e.* the Union of Italian Chambers of Commerce)
- *I.Stat* - a warehouse of statistics currently produced by Italian National Institute of Statistics.
- *MOVIMPRESE* - the quarterly statistical analysis of the birth/death rate for businesses, run by InfoCamere

Dataset consists of the national arrivals in the twenty-one Italian regions in 2012 (regions are twenty-one instead of twenty because the "South Tyrol Trentino" was split in two autonomous provinces: Bolzano and Trento). In addition, the value of several potential predictors of TIRwas observed or computed for each region:

- ***district****:* number of towns in the region;
- ***pilgrimage****:* number of pilgrimage areas;
- ***outlet****:* number of great shopping centres (we consider just centres having more then 10 stores);
- ***parks****:* number of theme parks (we consider just parks having more then>85'000 square meters);
- ***hospitals****:* number of public and private hospitals divided by the n° of districts;
- ***facilities1****:* number of hotel facilities
- ***facilities2****:* number of others accommodation facilities;
- ***congress***: number of exhibitions and conference centres;
- ***food***: number of food services.

Table 1: Critical regions within the Italian tourism network (total arrivals in 2012: 54'994'582)

Regions	*arrivals*	*TIR (%)*
Aosta Valley	672'268	1.22%
Friuli-Venezia Giulia	1'088'400	1.98%
Umbria	1'561'746	2.84%
Abruzzo	1'386'602	2.52%
Molise	164'923	0.30%
Basilicata	457'302	0.83%
Calabria	1'264'836	2.30%
Sardinia	1'247'003	2.27%

Table 1 shows travel-in rates for some critical regions, where we defined as "critical" a region characterized by a *travel-in* rate lower than 3 per cent.

The R package `betareg` [3] was used to fit a Beta regression model with a probit link, where *travel-in rate* depends on all variables listed above. Bias corrected ML estimates [5] of the β_i parameters are numerically obtained by using the BFGS method and are shown in Table 2, together with their own statistical significance: the number of outlets, hotels and other accommodation facilities as well as the number of hospitals are significant predictors of TIR. The *estimated*travel-*in* rate $\hat{y}_j, j = 1,\ldots,n$, can be written as

$$\hat{y}_j = g\left(\hat{\mu}_j\right) = \sum_{i=1}^{k} x_{ji}\hat{\beta}_i \qquad (4)$$

The explained variation is 86% of the total variation (pseudo R-squared = 0.8579). A ten-fold cross-validation yielded a reasonable RMSE, equal to 0.045.

Table 2: Beta regression model: bias corrected ML estimates and their significance.

	β_i *estimates*	*Significance*
(Intercept)	-2.34900	<0.000***
district	0.00018	0.0791 ·
pilgrimage	0.02590	0.638
outlet	0.08584	0.0329 *
parks	0.00482	0.8945
facilities1	0.00011	0.0133 *
facilities2	0.00001	0.0136 *
hospitals	1.37300	0.0243 *
congress	-0.00016	0.1278
food	0.00000	0.6842

4 Simulation study and conclusions

A simulation study was performed with the aim of investigating the effect of modifications of the initial conditions on the *travel-in*rates $\hat{y}_j$.
Table 3 shows the values of the covariates that enter the model (4) and the estimates$\hat{y}_j$ for the critical nodes displayed in Table 1, while Table 4 shows the new estimates$\hat{y}_j$after generating the following (quite hypothetical) modifications in the covariates:

- for all the critical region except Friuli, the determinant"outlet" was modified by promoting at least one newcommercial centre;
- for all the regions, the covariate"facilities2" was modified by promoting the launch of new bed and breakfast, that increases the estimated values of 5 per cent.

In all "critical" regions this perturbation successfully increases the estimated *travel-in* rate.

Table 3: Critical nodes: values of the covariates and model estimates$\hat{y}_j$

Critical nodes	outlet	facilities 1	facilities 2	hospital	$\hat{y}_j$	*arrivals*
Aosta Valley	0	482	576	0.0270	1.26%	672'268
Friuli	2	742	4347	0.0833	2.83%	1'088'400
Umbria	0	554	3324	0.1630	2.40%	1'561'746
Abruzzo	1	800	1580	0.0951	2.33%	1'386'602
Molise	0	108	429	0.1103	1.35%	164'923
Basilicata	0	238	567	0.1221	1.42%	457'302
Calabria	0	840	1900	0.1418	2.21%	1'264'836
Sardinia	0	913	3191	0.1087	2.07%	1'247'003

Table 4: Simulation studies: perturbed values (in bold) and model estimates $\hat{y}_j$

Critical nodes	outlet	facilities 1	facilities 2	hospital	$\hat{y}_j$	*Arrivals* in_{est}
Aosta Valley	**1**	482	**605**	0.0270	**1.56%**	**859'723**
Friuli	2	742	**4564**	0.0833	**2.84%**	**1'562'891**
Umbria	**1**	554	**3490**	0.1630	**2.94%**	**1'617'154**
Abruzzo	1	800	**1659**	0.0951	**2.34%**	**1'284'732**
Molise	**1**	108	**345**	0.1103	**1.67%**	**920'917**
Basilicata	**1**	238	**490**	0.1221	**1.76%**	**966'307**
Calabria	**1**	840	**1995**	0.1418	**2.70%**	**1'486'847**
Sardinia	**1**	913	**3351**	0.1087	**2.55%**	**1'402'427**

From a graphical point of view, the modifications in the determinants yield changes in the weights of the linkage structure of a network.

With reference to an *ego network*, consisting of a focal node ("ego") and the nodes whom ego is directly connected to (these are called "others") plus the ties, if any, among the others. Fig.2(a) displays the ego network of Calabria region before the simulation, while Fig.2(b) after the simulation.For both networks, 50'000 units is the cut-off below which links are not included in the network. The number of tourist arrivals (called weight of the link) from a given regionis shown foreach entering link of Calabria. The sum of the weight of all entering links is called *in*-degree of the node or shortly "I".

The graph shows that "**Campania**", "**Apulia**", "**Lombardy**", "**Lazio**" and "**Sicily**" are the nodes from which touristsarrive in Calabria.Hence, given the estimated *travel-in* rate $\hat{y}_j$=0.27 (see table 4), the new *in*-degree of Calabria $\hat{I}_j$ can be assessed as:

$$\hat{I}_j = \frac{\hat{y}_j \cdot I_j}{(TIR)_j} \tag{5}$$

where "j" indicatesCalabria region.

The weight of its links increased from 1'264'836 to 1'486'847: the level of criticality of Calabria decreases.
In conclusion, in the context of tourism strategies planning, this study set out to provide some technical tools to investigate the relevance and the statistical significance of some determinants of Italian domestic tourism, as well as to assess the extent to which possible modifications in such determinants can affect people's destination preferences.
Beta regression model appears to be a valid technique to predict the evolution of a system whose critical points are characterized in terms of

anormalized indicator, ranging from zero to one. It also providesa natural guide to future research: for example, it could be interesting to study whether and how the technique can be extended to apply to international tourism inflows.

a)

b)

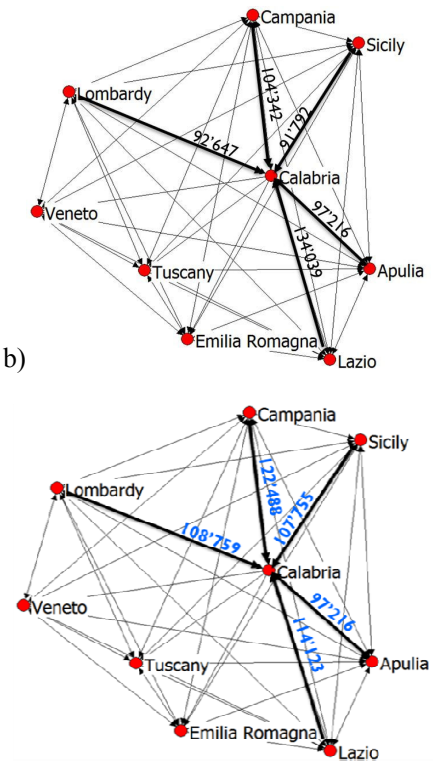

Fig.2 - ego network of Calabria, before (a) and after simulation studies (b)

References:

[1] Bonetti E., Cercola R., Izzo F., Masiello B., *Eventi e strategie di marketing territoriale. Gli attori, i processi e la creazione di valore*. Franco Angeli Edizioni, 2017.

[2] Cook, D.O., Kieschnick, R., McCullough, B.D., Regression analysis of proportions in finance with self selection, *Journal of Empirical Finance*, 15, 2008, pp. 860–867.

[3] Cribari-Neto,F., Zeileis,A., Beta Regression in R., Journal of Statistical Software, **34**,1,2010, pp. 1–24.

[4] Ferrari, S.L.P. and Cribari-Neto, Beta regression for modeling rates and proportions, *Journal of Applied Statistics*, 31, 2004, pp. 799-815

[5] Grun, B., Kosmidis I. and Zileis A., Extended beta regression in R: shaken, stirred, mixed, and partitioned, *Journal of Statistical Software*, 48, 11, 2012, pp. 1-25.

[6] Pierret F. (2011) Some points on Domestic Tourism. http://unwto.org/en/opinion/some-pointsdomestic-tourism

[7] Patuelli R., Mussoni M., Candela G. (2013) *The effect of World Heritage Sites on Domestic Tourism: a Spatial Interaction Model for Italy*. RCEA Working paper.

[8] Simas, A.B., Barreto-Souza, W., and Rocha, A.V., Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics and Data Analysis*, 54, 2, 2010, pp. 348-366.

[9] Smithson, M. and Verkuilen, J., A better lemon squeezer? Maximum likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11, 2006, pp. 54-71.

[10] UNWTO (1994) Recommendations on Tourism Statistics. Statistical Papers. UNWTO, New York