A non-linear approximation to the distribution of total expenditure distribution of cruise tourists in Uruguay

Juan Gabriel Brida*, Bibiana Lanzilotta, Leonardo Moreno, Florencia Santiñaque

Research Group in Economic Dynamics (GIDE), Department of Quantitative Methods, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Uruguay

A B S T R A C T

This study contributes to the literature on the determinants of tourism spending on cruises at a microeconomic level, through the application of innovative methodologies framed within the machine learning literature. The objective is to study the distribution of the total expenditure of cruise passengers in Uruguay, using data of the 2016–2017 cruise season survey (collected by the Ministry of Tourism of Uruguay). Due to the nature of this variable, we implement a two stages modeling strategy. In the first stage, we model the probability of spending, and in the second, the strictly positive spending. The paper analyze the distribution of conditional expenditure to a set of sociodemographic, travel, contextual and satisfaction variables applying non-linear regression techniques with Lasso penalty and nonparametric techniques such as Random Forest. The empirical results show that the key variables that determine the average spending of cruise tourists are their residence and the port of arrival of the cruise. The analysis of the predictive performance of the models (applied through a training sample and a test sample) shows that Random Forest method has the greater predictive capacity. Finally, the importance variable is analyzed by Random Forest.

1. Introduction

The contribution of the tourism sector to the economies is so important that it manifests itself as the main economic sector in many countries Song, Dwyer, Li, and Cao (2012). In addition to this, it has to be noted the importance of the tourism sector as an engine of economic growth for most of the countries. This topic has been extensively studied in the so called Tourism Led Growth Hypothesis (TLGH), as is shown in recent review papers (see Brida, Cortes-Jimenez, and Pulina (2016); Castro-Nuño, Molina-Toucedo, and Pablo-Romero (2013) and Pablo-Romero and Molina (2013).) This phenomenon is the main motivation for the increasing interest from academicians on studying the economic impacts of tourism. From this economic perspective, one of the main topics that have to be analyzed and studied is tourism demand. The analysis and interpretation of elements such as the demand elasticity or the average expenditure of tourists associated with specific characteristics can provide valuable information for the development of the tourism economy. One strand of this research topic, the analysis of aggregated demand, has been extensively analyzed. There are hundreds of empirical studies using different econometric and statistical methodologies and several published review papers on aggregate tourism demand modeling and forecasting (see Dogru et al. (2017), Peng et al.

(2014), Peng et al. (2015), Lin and Song (2015), Dwyer et al. (2012), Divisekera (2013), Song and Li (2008) and Lim (2006)). However, the academic work that analyze the determinants of tourist spending at the individual (microeconomic) level has been relatively less than those studying aggregate (macroeconomic) demand, considering both the number of contributions and the heterogeneity, of statistical methods and econometric models. As highlighted by Wang and Davidson (2010), apart from other published forty years ago (for example see Mak, Moncur, and Yonamine (1977)). Research on this topic starts in the 1990s and has produced no more than 27 studies after 2000.

According to Brida and Scuderi (2013), the interest towards the study of the determinants of individual tourist expenditure is linked to the evolution of the tourist market after the 1990s. Before then tourism was seen as a mass phenomenon, in the fashion of the Fordist economic development framework, when products came with rigidity and high standardization. As pointed in the review papers Mayer and Vogt (2016), Brida and Scuderi (2013) and Xiao and Smith (2006), most of the papers studying individual tourist expenditure have been case studies for specific tourism destinations. Therefore, global conclusions, referring to a larger population cannot be drawn. As it is explained in Brida and Scuderi (2013) different measurements of expenditure, models, dependent variables and regressors are employed in these

---

studies and then particular results cannot be generalized. This last comprehensive review reports that various individual characteristics have been tested in models by the past studies, such as economic attributes (income, preferences, occupation, etc.), sociodemographic characteristics (age, gender, place of residence, education, marital status, etc.), trip-related features (accommodation type, transport, destination, travel information source, etc.) and psychographic variables (opinions, attitudes, motivations, etc.). In addition, the same paper sustains that many different research questions are still open. In particular, the authors show that the studies on the determinants of individual tourist expenditure are not characterized by great heterogeneity in the use of statistical econometric models.

Statistical and econometric models involving tourist expenditure and a set of covariables can be categorized into two main groups, according to the scale of measurement of the response variable: models for metric response (for example, linear regressions and its generalizations) and models considering categorical responses (like logistic regressions). Very recent publications introduce innovative and original new methodologies to study the determinants of tourist spending. Some examples of these alternative methodologies are described below. Marrocu, Paci, and Zara (2015) introduces the quantile regression approach to analyze the main determinants of tourist expenditure of non-resident tourists who spent their holidays in Sardinia during the period April–October 2012. Olya and Mehran (2017) uses complexity theory and fuzzy set qualitative comparative analysis to explore a variety of configurations sufficient for simulation of both high and low scores of outbound tourism expenditures. As an alternative to conventional OLS and Tobit estimators, Melstrom (2017) presents an exponential model of tourist expenditures estimated by a quasi-maximum likelihood (QML) technique. Abbruzzo, Brida, and Scuderi (2014) introduces the use of scad-elastic net in the assessment of the determinants of individual tourist spending.

This paper contributes to the strand of the empirical literature on the determinants of individual tourist expenditure. The main research question deals with the study of the impact of a set of socioeconomic and behavioural variables on individual spending, applying an appropriate model in order to overcome multicollinearity and perform selection of regressors. Our analysis runs on the behavior of cruise tourists in Uruguay during the cruise season 2016–2017, by using data from the official survey conducted in Uruguay by Ministry of Tourism (Ministerio de Turismo (2017)). To our best knowledge, this is the first attempt in the microeconomic tourism literature to use this technique. As shown by the comprehensive review of Brida and Scuderi (2013), previous papers apply classic econometric models. Cruise tourism has experienced a very important expansion over the past twenty years, with an average annual growth rate of around 8% in the number of worldwide cruise passengers during the period 1995–2010 (see Paoli et al. (2017); Wang, Wang, Zhen, and Qu (2016); Paoli et al. (2017) and Wang et al. (2016)). In Uruguay, the number of cruise tourists has passed from around 56 thousand (2004–2005) to more than 400 thousand tourists-cruisers in 2011–2012 and 2012–2013. (see Brida, Bukstein, and Tealde (2015); Brida et al. (2015)). The privileged position of Montevideo and Punta del Este ports, located between Buenos Aires and Rio de Janeiro, is a determinant variable in the number of cruises that have as stopping point such ports, even if it should be mentioned that in the last seasons visitors declined but continued to outperform the 300 thousand.

The paper is organized as follows. Section 2 describes the database and the main variables generated from the survey questionnaire. section 3 presents the methodology and section 4 shows the empirical results and discusses them. Finally, section 5 concludes. The codes in the R language (TeamRCore (2017)) are available upon request to the authors.
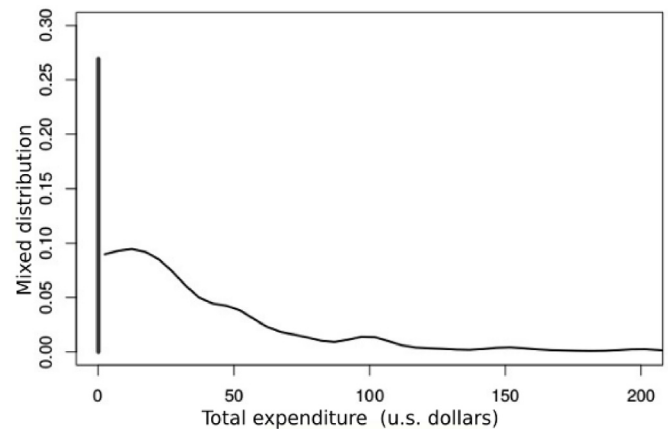

Fig. 1. Representation of the probability function of the mixed variable Total Expenditure. Source: own elaboration with the R language.

2. Dataset description

Individual data came from the 2016–2017 season survey of cruisers, collected by the Ministry of Tourism of Uruguay. The arrivals are distributed between November 2016 and April 2017. Data is open and available at Ministerio de Turismo (2017). The sample was obtained through a systematic sampling on the expected arrivals of the cruisers; in each of them, a simple random sample of tourists is taken. The number of registered surveys was 3117, out of a total of 22 cruise ship arrivals (in Montevideo and Punta del Este ports). The questionnaire includes sociodemographic, context and satisfaction inquiries.

The model is constituted by a nonnegative dependent variable *Total per capita expenditure*. This is a mixed type variable, which distribution is partially discrete and partially continuous, with a positive probability at zero, see Fig. 1.

Mean value of tourist expenditure in American dollars is $ 31, and its standard deviation is $ 35. Nevertheless, excluding null expenditure −27% of the passengers has zero expenditure-the mean value became $ 42.38 with a standard deviation of $ 36.

The explanatory variables can be classified in four different groups: satisfaction variables (variables of pleasure and displeasure), socio-economic variables (residence, age, occupation), travel variables (month, number of previous visits, descending port, sites visited) and context variables (number of group members -one, two and more than two- and group type -mixed, only women or only men-). The satisfaction variables are dichotomous variables, where the tourist determines which were his main pleasures and dislikes during her/his visit at the destination. A summary of those 10 variables with the highest response rate is presented in Fig. 2. Note that tourists appreciate the gentle treatment (54% of agree), and have a low displeasure for the level of prices of the country (12%).

Fig. 3 shows the distribution of the total expenditure of visitors according to the expense of the cruisers of the sample discriminated according to their nationality and port of landing. Note that Brazilian tourists are those who present a higher level of expenditure, particularly in Montevideo. Also note that a significant proportion of passengers from Argentina and Brazil tend to spend less when the port of landing is Punta del Este. This behavior could be a consequence of the availability of free activities at that destination (beautiful beaches) and the high level of prices. (according to Fig. 2, this is the main source of displeasure mentioned by cruise passengers).

In Fig. 4 it can observed the evolution of Total average expenditure longitudinally, i.e. as a function of the months of the season. Note that lowest average expenditure in the port of Montevideo (less than 25 dollars) was in December 2016, which was also a low-spending month for Punta del Este. Anyway, in the latter, the worst performance was in