Unbundling Package Tours: a Machine Learning Application with the LASSO

Andrea Carboni*, Claudio Doria*, Alessandro Moro†

Abstract

In order to estimate the travel item of the Italian Balance of Payments (BoP), the Bank of Italy carries out an extensive border survey, collecting information about travel expenditures from a sample of resident and foreign travellers. The travel item covers an assortment of goods and services: in particular, according to the international standards, it includes local transport, i.e. transport within the economy being visited, but excludes international transport, reported in a separate BoP item. In the questionnaire of the survey a detailed breakdown of expenditures is asked, allowing the correct split between the travel and international passenger transport components. However, this breakdown is not available if these two items are purchased in a package tour with a single transaction. The unbundling of package tours is therefore needed for the correct compilation of the BoP. The present paper proposes a machine learning algorithm based on LASSO techniques to impute the components of package tours, improving the performance of the current procedure employed by the Bank of Italy.

Keywords: BoP statistics; Travel; Package tours; Machine learning, LASSO

JEL classification: C10, C25, C80

Summary

* Bank of Italy, Statistical Analysis Directorate

† Bank of Italy, International Relations and Economics Directorate

Introduction[1]

The Bank of Italy carries out an extensive border survey, called Survey on International Tourism, designed to elicit the travel expenditures of a sample of resident travellers coming back to Italy from a trip abroad and of foreign travellers leaving Italy after a visit in the country.[2] The main purpose of the survey is the estimation of the travel item of the current account of the Italian Balance of Payments (BoP). Travel is a relevant component of Italian economy: in 2018, foreign travellers' expenditures in Italy were 41.7 billion (2.4 per cent of Italian GDP), while Italian expenditures abroad amounted to 25.5 billion (1.5 per cent relative to the GDP).

Unlike most of the other service categories of the BoP, travel is a transactor-based component that covers an assortment of goods and services. On the one hand, as reported in the Balance of Payments and International Investment Position Manual (IMF, 2009), goods and services provided to visitors during their trips, that would otherwise be classified under another item (such as postal services, telecommunications, local transport, hire of equipment, or gambling), are included under travel. On the other hand, travel excludes goods for resale, which are included in general merchandise, and the acquisition of valuables (such as jewellery), consumer durable goods (such as cars and electric goods) that are included in customs data when in excess of custom thresholds.

Moreover, according to international standards,[3] travel includes local transport (i.e., transport within the economy being visited and provided by a resident of that economy), but excludes international transport, which is included in a specific BoP item. International passenger transport covers all services provided in international transports to non-residents by resident carriers, as a credit, and those provided to residents by non-resident carriers, as a debit.

In the questionnaire of the survey a detailed breakdown is adopted, making a distinction between international and local transport, accommodations, meals, other services (museums, courses, concerts, etc.), and goods (shopping). However, this breakdown is not available for package tours, when two or more items of an international travel are purchased with a single transaction. In fact, with regard to this kind of trips, it is possible to know only the total value of the package and which services are included but the value of each service bought with the package is unknown. The unbundling of package tours, which account for more than 20% of the travel credits and debts in 2018, is therefore needed for the correct compilation of the BoP.

Currently, the donor method is used to unbundle package tours: in fact, the package of a given traveller is broken down in its different components using the proportion of an average "twin" traveller, who has not purchased a package in his travel. In principle, the traveller and his twin should have the same characteristics: country of residence, mean of transport, length of the stay, type of accommodation and reason of the trip. However, it is difficult to find enough twins to have stable estimates according to all these features and, consequently, some constraints must be relaxed with the risk of introducing bias in the estimates.

In order to overcome the limitations of the current procedure, it is necessary to model the relationship between the most important components of a package tour (transportation, accommodation, and other included services) and the characteristics of travellers and those of their trip. Moreover, it is worth to select the relevant features, to be included in the model as explanatory variables, in an efficient way.

This paper proposes a Machine Learning (ML) approach[4] to solve these two issues: firstly, a linear relationship is supposed between the shares of expenditure in the three major components of a package tour and a huge set of explanatory variables derived from the border survey; then, the relevant features are selected using a popular

[1] We thank Matteo Piazza, Alfonso Rosolia and Simonetta Zappa for helpful comments and suggestions. The views expressed in the paper are those of the authors and do not involve the responsibility of the Bank of Italy.

[2] For the sake of brevity, in the rest of the paper we might refer to resident travellers using the adjective Italian and we might use the term foreign for non-resident ones.

[3] In addition to IMF (2009), see also United Nations (2008, 2010).

[4] There is a growing interest on practical applications of ML algorithms in central banks. For reviews of these techniques and central bank applications, see Friedman et al. (2001) and Chakraborty and Joseph (2017).

regularisation method in the ML literature, i.e. the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996).[5]

Our proposed algorithm is trained using the interviews of the travellers without a package tour: in fact, for these travellers we know both their characteristics (and the features of their travel) and the expenditures in the different items. Then, the algorithm is applied to travellers with a package tour in order to impute the value of the unknown components. The comparison of the LASSO and donor method shows that the LASSO approach clearly outperforms the latter method in terms of prediction accuracy. In fact, the LASSO exhibits a lower variance of the forecast error term and eliminates completely the systematic bias that affects the current procedure.

Moreover, the strength of the approach described in this paper is also in its ability to incorporate the potential effects of exogenous variables that may alter the expenditure behaviours of international travellers. This flexibility will allow to take into account the impacts of the recent COVID-19 pandemic in the unbundling procedure applied to the next waves of the tourism survey.

The rest of the paper is organised as follows: the next section briefly describes the Italian frontier survey while Section 2 presents an operational definition of package tours from a BoP perspective and Section 3 describes the major trends in the diffusion of package tours among the Italian and foreign travellers and the composition of a standard package. Section 4 illustrates the current donor procedure while the new ML algorithm is presented in details in Section 5; Section 6 compares the forecast performance of the two approaches. Section 7 concludes.

1. The Bank of Italy border survey on international tourism

The Italian Survey on International Tourism is conducted on a monthly basis by the Bank of Italy since 1996. It is an inbound-outbound frontier survey whose main objective is the collection of information about tourist expenditures in order to estimate the travel item of the BoP according to the international standards (IMF, 2009). From a methodological point of view the survey consists of two operations carried out at each of the selected border points: counting and interviewing.

The counting aims at assessing the number of travellers entering or exiting Italy at the end of the trip, broken down by country of origin: this information is essential to gross up the sample data to the reference population. The interviewing consists in questioning a sample of travellers, after having approached and stopped them, in order to assess a number of basic classification characteristics of travellers, their trip and expenditures. The interviews are of the Computer-Assisted Personal Interviewing (CAPI) type and realised through a structured questionnaire.

Since the primary objective of the survey is the measurement of tourist expenditures, the interviews are carried out at the end of the stay, i.e. when travellers return to their own country of residence. The interviews and the counting operations are performed at the same time, so that the travellers interviewed belong to the population of people crossing the border in a given time frame.

Whenever possible, the counting procedures are integrated with administrative data, in particular those provided by airport and port authorities. The administrative data are useful to estimate the total number of cross-border movements. For instance, an airport authority registers the number of passengers (but not their nationality) who is arrived and departed with an international flight from the airport layover in each month. By combining the counting results of the survey with the official data of the airport authority, it is possible to obtain an estimation of the international travellers by country of origin.

[5] A number of authors have studied the ability of the LASSO and related procedures to select the relevant features and recover the correct model. Examples of this kind of literature include Knight and Fu (2000), Greenshtein and Ritov (2004), Tropp (2004), Donoho (2006), Meinshausen and Buhlmann (2006), Tropp (2006), Wainwright (2006), Zhao and Yu (2006), Buneaet al. (2007), and Meinshausen (2007) and, more recently, Lee et al. (2016), Plan and Vershynin (2016), Dalalyan et al. (2017).

Annually about 120.000 face-to-face interviews and 1.200.000 counting operations are realised. These numbers assure a modest sampling error of the total expenditure estimates abroad, which are the credits and debits of the travel item published in the Italian BoP. The sample is stratified according to six variables, indicated in table 1 along with their respective levels.

Table 1: *Stratification variables and corresponding levels*

VARIABLE	LEVELS
1. Direction	2 (inbound, outbound)
2. Type of carrier	4 (road, rail, airports and seaports)
3. Frontier point	62 (22 road, 4 rail, 25 airports, 11 seaports)
4. Day of data collection	day in the month (e.g., 31)
5. Time of the day	3 (first shift, second shift, third shift)
6. Type of day (only road frontiers)	2 (working, holiday)

The selection of levels, for each of the six variables, is realised aiming at the optimisation of the resources allocated to the data collection. For each level two domains are defined: one for the Italian travellers, the other for foreign travellers. The need to select the direction as a stratification variable is clearly in relation to the objectives of the project: as the survey is both outbound and inbound, a sample of Italian travellers coming back to Italy after a journey abroad and a sample of foreign visitors leaving Italy has to be selected.

The selection of four types of carriers answers the requirement of detecting the flow of travellers considering all types of means of transport and all types of border crossings usable to reach or to leave the Italian territory: road, rail, airports and seaports. Actually, the investigation can be considered as composed by four separate surveys, as the sampling is independently carried out at each type of frontier.

The selection of the individual frontier points is based on the results obtained with the survey itself and on the information available from administrative data. For instance, some airports were included or excluded from the survey considering their volume of international flights and with regard to their agreements with the major airline companies. Overall, in the survey about 60 frontier points (25 airports, 22 roads, 12 ports and 4 railways) are considered.

At each border point, a given number of interviews is assigned to the interviewers in proportion to the flow of passengers, residents and non-residents, passing through them. Still in relation to the flow of passengers, for each frontier point, the days and the time periods for the investigations are also decided. All the data are recorded in electronic form with a tablet.

2. Package tours: an operative definition from the BoP perspective

As mentioned in the Introduction, from the BoP compilers' perspective it is relevant to know the value of the international transport in a package tour in order to allocate this expenditure correctly in the international transport item, and not in travel. More in general, it is crucial to define a "package" and verify in which cases an unbundling procedure is needed to allocate the total value of the package among the different components.

Trying to give an operational definition from the compilers' perspective, a package is the purchase of two or more services, of which at least one to be recorded in the travel item of the BoP, when it is unknown the partition of the value of the transaction among the different components.[6]

[6] The Directive EU 2015/2302 gives an official definition of package tours for the consumer protection as a combination of at least two different types of travel services for the purpose of the same trip (or holiday).

Broadly speaking, services that are included in the travel item are: accommodation services (Hotels, B&B, houses, camping), food-serving services (restaurants, bars), local transport services (taxies, buses, undergrounds, national air/train/coaches tickets), other services not included elsewhere (guided tours, museums, sports events, concerts, courses). However, international transport services, such as the ticket of the flight that a foreign traveller pays to reach Italy, have to be excluded from the travel item and considered in the international passenger transport item.

With regard to international travels, the unbundling of a package is the procedure to allocate the total expenditures of the package to its different components: in other words, knowing the total value of the package tour, unbundle it means to estimate the value of the different services that make it up.

In unbundling a package tour, BoP compilers could incur in three kinds of mistakes:

1. wrong estimation of the travel item and of the total amount of the goods and services account;
2. misallocation of the amounts between the two categories of services, i.e. travel and international transport;
3. misallocation of the amounts within the components of the travel item.

The first mistake happens when the international carriage is provided by an operator resident in the same economy of the traveller: in fact, in this case, the transaction should not be recorded in the BoP and a wrong estimation of the package components leads to an overestimation/underestimation of the travel item and thereby of the current account.

When the international carriage is provided by an operator who is resident in a different country with respect to that of the traveller, an international transaction occurs, and the value of the services should be reported in the BoP under international transport. In these circumstances, a wrong estimation of the international transport service in the package generates a misclassification between travel and international transport.

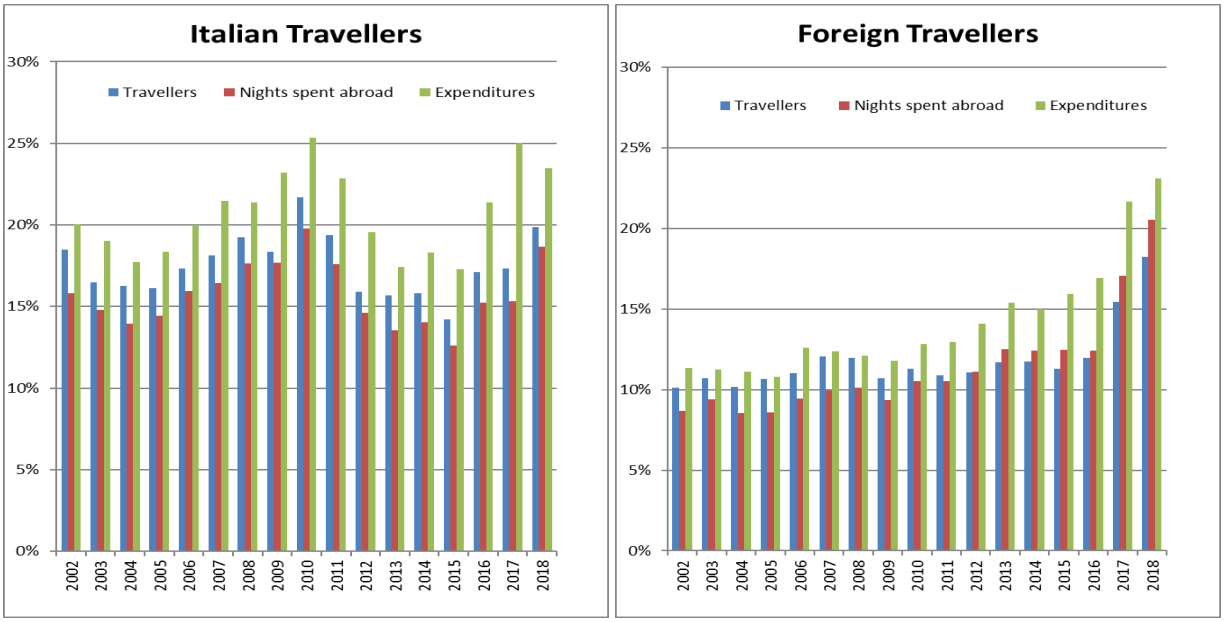
The misclassification within the travel item may occur when all the components of the package are elements recorded in the BoP under the travel item: i.e., if the package does not include the carriage of passengers. For instance, a package comprehensive of a hotel half board and a guided tour of the town includes three kinds of services recorded under the travel item (accommodation, food-serving and other services). In this case a mistake in the unbundling has only an effect on the value of the different components of the travel item.

From the accounting perspective, the most relevant issue is the estimation of the carriage of passengers in a package tour. In fact, only an imprecise estimation of the carriage of passengers in the unbundling procedure of a package may generate a mistake in the estimation of the travel item (and, as a result, of the total value of the services), or in the allocation of the amounts between travel and transport in the account of goods and services. In the other cases, also in presence of an incorrect splitting process, there are no mistakes in the compilation of the BoP, but at most an inappropriate breakdown within services included in the travel item.

3. Package tours: some evidence from the Italian border survey

Information about package tours are available in the Survey on International Tourism of the Bank of Italy, starting from 2002. Since then, the number of travellers with a package tour, their nights spent abroad and their expenditures have shown an increasing trend, both from the inward and the outward side. However, since all tourism indicators have grown in Italy in the last years, it seems appropriate to analyse the evolution in the use of package tours in comparison to the general trends of tourism, i.e., in relative terms (see fig. 1).

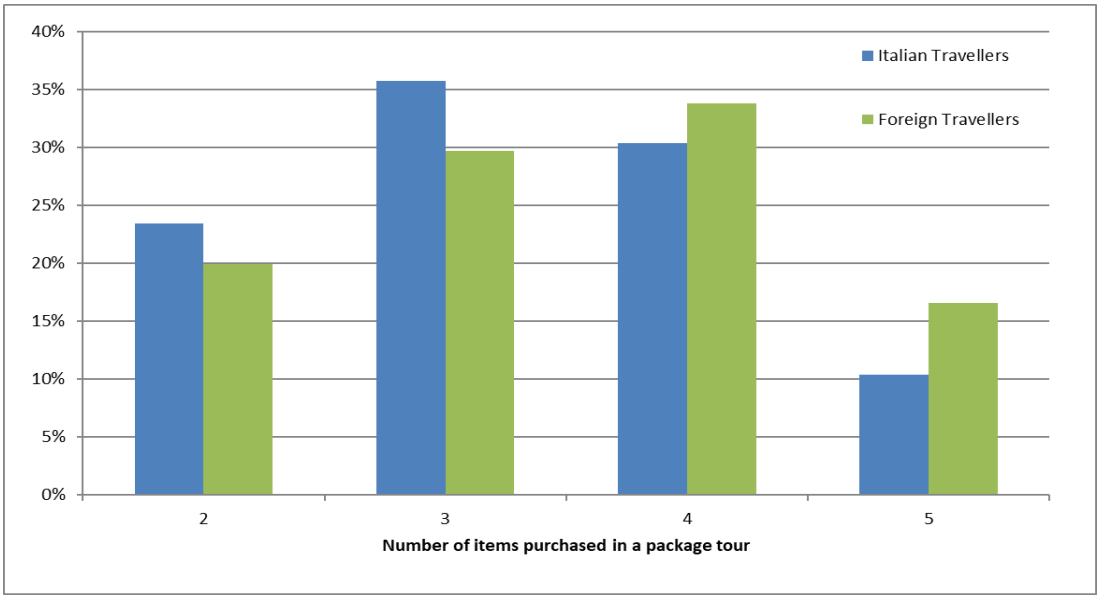
Fig. 1: *Packages tours - share of travellers, nights and expenditures (only overnight stays)*



In relative terms, the share of package tours bought by Italian travellers presents some oscillations over the period analysed with an increase in the last three years. On the other hand, an increasing trend can be observed for foreign travellers in the last decade.

In the questionnaire both the total value of the package, i.e. the total expenditure of the traveller for the package, and the types of services included (accommodation, flight, etc.) in the package are requested (but the amount spent in each service is unknown). The classification of services in a package is the following one: international transport[7]; transport abroad (internal to the hosting country); accommodation; food-serving services; other services not included elsewhere. The majority of package tours (see fig. 2) are composed by a combination of three or four of these categories of services: this behaviour could be indicative of a preference for all-inclusive trips.

Fig. 2: *Package tours - composition by number of services included (data refer to 2018)*



[7] The international transport could include also a domestic component, given by the physical movements of the people in the country of origin before the international travel (for instance a domestic flight, before the international flight).

In 2018, 37 per cent of Italian international travellers have purchased package tours with three categories of services and 30 per cent with four types of services. While 30 per cent of the foreign visitors who have reached Italy have bought a package tour with three different services and 34 per cent with four categories. In 2018, less than a quarter of package tours (23.6 per cent for Italian and 19.9 for foreign travellers) was composed by two categories of services.

The large majority of package tours include the accommodation and the international carriage of passenger. In 2018 Italian international travellers, who have purchased a package tour, have asked for the accommodation in the 99 per cent of the cases and for the international carriage in the 95 per cent of these transactions. The proportions are similar for foreign travellers who has reached Italy purchasing a package tour: the 95 per cent has bought the accommodation and 90 the international transport. In the same year, the package tours that included simultaneously international carriage and accommodation abroad were about 94 per cent for the Italian international travellers and 88 per cent for foreign visitors.

Moreover, considering the travellers who have bought a package tour with accommodation, in about 85 per cent of the cases they have booked an hotel; analysing the travellers who have acquired a package with international transport included, in more than 90 per cent an international flight was used as mean of transport.

In 2018, the 60 per cent of the buyers of a package tour have purchased a local transport abroad: the percentage is the same for Italian international travellers and for foreign ones. The purchase of a food-service is rather rare in a package tour: in 2018, it was present in the 26 per cent of the packages bought by Italian travellers and in the 28 percent of the packages purchased by foreign visitors in Italy.

Other kind of services of heterogeneous nature (guided tours, museums, theatres, sport and music events, courses, etc.) were present in about 40 per cent of the package tours purchased by Italian international travel and in a half of the package tours bought by non-residents travellers.

In conclusion, looking at the results of the Bank of Italy's border survey, in the last three/four years there is a growing trend in the use of package tours, especially for foreign travellers who have visited Italy. Less than a quarter of package tours are a combination of only two services, while the majority includes 3 or 4 categories of different services related to travel. Lastly, looking to the services purchased with a package tour, the large majority of package tours includes the accommodation (generally, a hotel) and the international carriage of passengers (almost ever a flight).

4. The current procedure to unbundle package tours

According to the descriptive evidence of the Bank of Italy's border survey, the main issue related to the unbundling of package tours is the estimation of the value of three components: the carriage of international transport, the expenditures for accommodation, and the residual component.[8] A bias in the estimation of the first component has also an effect in the correct allocation of the monetary flows between the two BoP components, i.e. travel and international transport: from the compiler perspective, the priority in unbundling the package tours is therefore the correct estimation of the international carriage of passengers. According to the results of the border survey, the second important aspect is the correct estimation of the accommodation in the package tours, as almost all packages contain this component.

[8] In the rest of the analysis, we decide to aggregate in this residual component the other services different from international transport and accommodation (i.e., food-serving services, local transport, other services not included elsewhere).

The current procedure adopted in the Bank of Italy is the donor method (or nearest neighbour approach). The value of the package of a traveller is split in its components, using the proportions of "similar" travellers, the so-called twins, who have purchased the same services without buying a package. The twins should have similar features (country of residence, length of stay, reason of the trip, type of accommodation, etc.) to those of the package purchaser. The idea is that similar travellers should have similar behaviours in their expenditures.

In order to clarify this approach, let suppose that an international traveller has purchased, spending 1.200 Euro, a package tour with three components: an international air ticket, a hotel abroad, and a car rental (always abroad). Let assume that in the data we find three "twin" travellers with the same characteristics that have purchased the same services without buying a package. Applying the donor method, the mean of the twins' proportions of expenditure is calculated in order to split the original package (see table 2).

Table 2: *Example of an application of the donor method*

	International Transport	Accommodation	Car Rental	Total
Twin 1	500	350	150	1,000
Twin 2	400	550	50	1,000
Twin 3	600	1,200	200	2,000
Average share (%)	0.4	0.5	0.1	
Package	480	600	120	1,200

In order to compute the weight of a specific component, it is necessary to estimate its share of expenditure for each twin and then calculate the average. For instance, the weight of the international transport in the example in table 2 is given by $\frac{1}{3} \cdot \left(\frac{500}{1000} + \frac{400}{1000} + \frac{600}{2000}\right) = 40\%$.

The idea that similar travellers should have analogous behaviours in their expenditure seems reasonable. Anyway, the donor method have a limit in the trade-off between the similarity of the twins and the sample size. In fact, on the one hand, it is desirable that the twins have the highest possible number of features identical to those of the traveller who has bought the package. On the other hand, it is necessary to have enough twins to produce robust estimates: this aspect is crucial since the donor approach can be seen as a non-parametric estimation method, which does not impose a parametric model to the data and hence requires more observations.

Looking at the survey results, for the main partner countries of Italy, it is usually possible to find enough donors; however, for others countries, with a limited number of international travellers, it is necessary to relax some constrains and, accordingly, reduce the similarity between twins and package purchasers, with the risk of introducing bias in the estimates.

5. A new methodology: the LASSO approach

In order to overcome the limitations of the donor method, a machine learning algorithm is proposed which should be able to improve the results of the unbundling procedure by exploiting in a more effective way all the information contained in the international tourism survey. In fact, on one side, a parametric structure is imposed to the relationship between the shares of expenditure in the different package components and the characteristics of travellers and their trips; on the other, the most useful features for the estimation of these shares are automatically selected using regularisation techniques.

The basic idea is to find a predictive model relating the shares of expenditure for international transport, accommodation and remaining services to the other variables collected in the survey, such as the travellers' socio-demographic characteristics, the country of origin/destination, the type of transportation and accommodation, the number of nights, and so on. Since these shares of expenditure are unobservable for package tours, this model must be estimated using the travellers who have not purchased a package tour: in fact, for this kind of travellers we can observe both the target variables (international transport, accommodation and other services) and the input variables (i.e., the characteristics of the travel and of the travellers). Then, the model can be applied to the travellers with a package tour in order to infer from their features the value of the different components of the package.

More precisely, denoting with $Exp_{i,t}^{j}$ the expenditure of traveller *i* in item *j* (international transport, accommodation, other services) at time *t*, $TOT_{i,t}$ the sum of the expenditures in the three items, the share of expenditure for item j can be defined as:

$$(1)\ \ Q_{i,t}^{j} = \frac{Exp_{i,t}^{j}}{TOT_{i,t}}$$

Then, it is possible to estimate the following relationship in which the share of expenditure in item *j* is explained by a set of features:[9]

$$(2)\ \ Q_{i,t}^{j} = \beta_0^j + \beta_{TD}^j TD_t + \beta_{CO}^j CO_{i,t} + \beta_{SD}^j SD_{i,t} + \beta_{TC}^j TC_{i,t} + \beta_{AC}^j AC_{i,t} + \varepsilon_{i,t}^j$$

where TD_t is a set of time dummies; $CO_{i,t}$ is the country of origin (destination) of the foreign (Italian) traveller; $SD_{i,t}$ is a vector of socio-demographic characteristics, such as the number of travellers, distinguished by sex and age, the job of the interviewed, the reason of the journey (work, pleasure, other); $TC_{i,t}$ are the transportation features, like the mode of transport (car, train, boat and plane), the transportation company, the class of the flight/boat; finally, $AC_{i,t}$ indicates a vector of accommodation variables, such as the number of nights distinguished by type of accommodation.

Equation (2) can be estimated separately for Italian and foreign travellers without a package tour and the model can be applied to travellers that have bought a package tour for the imputation of the unknown expenditures. In fact, for this latter kind of travellers, we know the total value of the package $TOT_{i,t}$, but we ignore the expenditures for the different items. The estimation of these components are therefore given by:

$$(3)\ \ \widehat{Exp}_{i,t}^{j} = \hat{Q}_{i,t}^{j} TOT_{i,t}$$

in which the shares $\hat{Q}_{i,t}^{j}$ are calculated as the predicted values of equation (2), using the features of the travellers with a package tour, rescaled in order to guarantee that $\sum_{j=1}^{3} \hat{Q}_{i,t}^{j} = 1$. More precisely, denoting with $\tilde{Q}_{i,t}^{j(1)}$ the rough predicted shares from equation (2), the final shares are obtained as:

[9] We have also tested a model in which the logit of the expenditure shares are regressed on the explanatory variables. However, this specification exhibits worse forecasting performance than the linear model presented in this section.

$$(4)\quad \hat{Q}_{i,t}^{j} = \frac{\tilde{Q}_{i,t}^{j(2)}}{\sum_{j=1}^{3} \tilde{Q}_{i,t}^{j(2)}}$$

where $\tilde{Q}_{i,t}^{j(2)} = \max\left(0, \tilde{Q}_{i,t}^{j(1)}\right)$ in order to rule out the few cases of negative predicted values.

The underlying assumption of the procedure is that there are no systematic differences in the expenditure shares between travellers with a package tour and travellers without a package, once controlling for the observed characteristics included in equation (2). Unfortunately, this hypothesis cannot be tested directly with the available data. However, it is important to stress that this assumption does not impose the equality between the total value of a package and the sum of the values of the different components if purchased separately: in fact, these two values are likely to be different due to agency costs or discount strategies. The assumption is violated only if the expenditure shares in the different items are different between package and standard tours, which is a far less restrictive hypothesis.

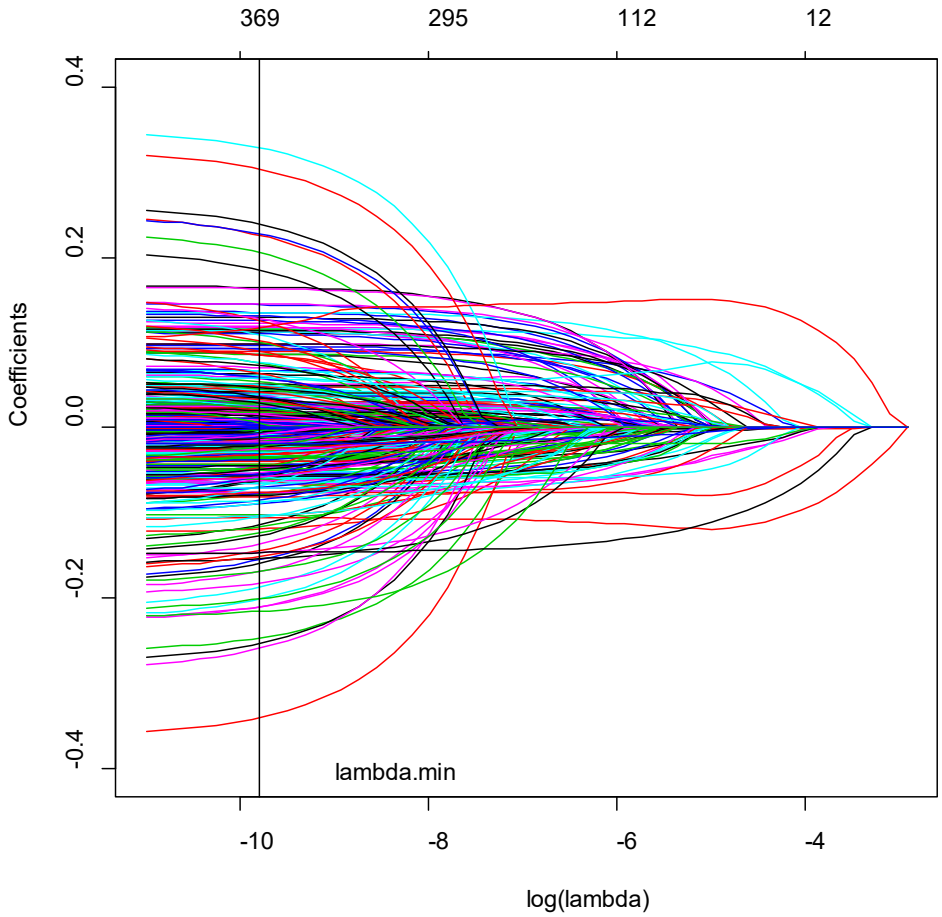
For the training and validation of the proposed algorithm, the Bank of Italy's data of the International Tourism Survey are employed. In particular, it is worth to consider the interviews of the Italian and foreign travellers without a package tour that have sustained all the three types of expenditures (international transport, accommodation and the residual component) during their journey. The interviews carried out in the 2011-2018 period are used ending up with a repeated cross-section database: the total number of observations are 216,974 for Italian and 294,636 for foreign travellers. The 80% of the sample is used for the training of the algorithm, i.e. for the estimation of the model (hyper-)parameters, and the remaining 20% for its validation, comparing the observed expenditures with the ones predicted by the model.

Since the right-hand side of equation (2) includes many variables, especially dummies (e.g., one dummy variable for each month and year of the interview, country of origin/destination, transportation company, etc.), it is useful to employ a regularisation method to automatically select the relevant features. One of the most common methods used in the machine learning literature is the Least Absolute Shrinkage and Selection Operator (LASSO). This approach adds the sum of the absolute values of the model coefficients to the sum of squared residuals to be minimised, forcing the coefficients of the irrelevant variables to zero. In formula, the coefficients are estimated in this way:

$$(5)\quad \min_{\beta_0^j,\beta_{TD}^j,\beta_{CO}^j,\beta_{SD}^j,\beta_{TC}^j,\beta_{AC}^j} \sum_{i,t} \left(Q_{i,t}^{j} - \beta_0^j - \beta_{TD}^j TD_t - \beta_{CO}^j CO_{i,t} - \beta_{SD}^j SD_{i,t} - \beta_{TC}^j TC_{i,t} - \beta_{AC}^j AC_{i,t}\right)^2 + \lambda^j \left(\left\|\beta_0^j\right\|_1 + \left\|\beta_{TD}^j\right\|_1 + \left\|\beta_{CO}^j\right\|_1 + \left\|\beta_{SD}^j\right\|_1 + \left\|\beta_{TC}^j\right\|_1 + \left\|\beta_{AC}^j\right\|_1\right)$$

For larger values of λ more coefficients are forced to zero: the choice of the value for this hyper-parameter becomes therefore crucial. Following the literature, λ is chosen by minimising the out-of-sample Mean Squared Error (MSE) in a cross-validation exercise in which the training sample is divided in five subsets. With the data considered the optimal λ is very small: this implies that many variables are relevant. For example, figure 3 reports the values of the estimated coefficients (on the y-axis) and the number of parameters set to zero (x-axis above) in correspondence to different values of $\log\lambda$ (x-axis below) in the case of international transportation expenditures of Italian travellers: it is possible to see that around 369 parameters are relevant to predict accurately the target variable (we start with 388 regressors).

Fig. 3: *Estimates of model coefficients and parameters set to zero for different values of λ in the specification with transport expenditures of Italian travellers as target variable. The vertical line represents the selected λ*



As pointed out in Friedman (2011, p. 91), the LASSO shrinkage causes the estimates of the non-zero coefficients to be biased towards zero. One of the suggested approaches for reducing this bias is to run the LASSO to identify the set of non-zero coefficients, and then fit an unrestricted linear model with OLS to the selected set of features. Since this is not always feasible in the case in which the selected set is large, it is alternatively possible to use the LASSO to select the set of non-zero predictors, and then apply the LASSO again, but using only the selected predictors from the first step: this is known as the relaxed LASSO (Meinshausen, 2007). In this paper we follow the first approach given that the OLS estimation in the second step is feasible.

6. Results of the algorithm

It is worth to compare the predictive performance of the proposed approach with the current donor method in order to understand if and how the new methodology can improve the unbundling of package tours.

Both methods are trained using the 80% of the sample and the remaining 20% is employed to compare the accuracy of predictions measured in terms of forecast bias, variance of the prediction errors and, combining these two dimensions, with the MSE. In particular, the forecast errors with method *m* (donor or LASSO) considered in the comparison are defined as:

$$(6)\quad e_{i,t}^{j,m} = \left(Exp_{i,t}^{j} - TOT_{i,t}\hat{Q}_{i,t}^{j,m}\right) \cdot w_{i,t}$$

where $w_{i,t}$ are the survey grossing-up factors. The bias, standard deviation (STD) and the Root Mean Squared Error (RMSE) of the forecasts are calculated using the error terms in expression (6).

Table 3 shows the results of this comparison, distinguishing between Italian and foreign travellers, as well as different package components. The analysis is conducted on the overall time period, i.e., the years from 2011 to 2018, and by focusing on the more recent four-year period 2015-2018, when there has been a significant growth of package tours, especially among foreign travellers (see fig. 1).

Table 3: *Comparison of the out-of-sample forecasting performance of the LASSO and donor methods*

		Italian Travellers			Foreign Travellers		
		Donor	LASSO	Diff (%)	Donor	LASSO	Diff (%)
Overall validation set (2011-2018)							
International Transport							
	Bias	-20.468	-12.490	-39%	-11.261	-5.452	-52%
	STD	262.110	203.690	-22%	270.976	222.304	-18%
	RMSE	262.905	204.070	-22%	271.207	222.369	-18%
Accommodation							
	Bias	14.921	6.389	-57%	9.189	2.248	-76%
	STD	233.524	207.552	-11%	230.049	210.824	-8%
	RMSE	233.998	207.648	-11%	230.231	210.835	-8%
Other Expenditures							
	Bias	5.547	6.100	10%	2.072	3.204	55%
	STD	192.472	177.094	-8%	236.475	228.858	-3%
	RMSE	192.550	177.197	-8%	236.482	228.879	-3%
Sub-sample (2015-2018)							
International Transport							
	Bias	-18.591	-13.978	-25%	-15.031	-6.640	-56%
	STD	290.378	221.787	-24%	323.495	263.541	-19%
	RMSE	290.966	222.222	-24%	323.838	263.619	-19%
Accommodation							
	Bias	13.513	5.198	-62%	12.841	3.832	-70%
	STD	261.990	226.663	-13%	279.434	272.820	-2%
	RMSE	262.332	226.717	-14%	279.724	272.842	-2%
Other Expenditures							
	Bias	5.078	8.780	73%	2.190	2.808	28%
	STD	219.074	206.611	-6%	276.205	272.130	-1%
	RMSE	219.127	206.792	-6%	276.209	272.139	-1%

It is possible to observe that the proposed approach clearly outperforms the donor method in the forecast of the most relevant components of package tours, i.e., the international transport and accommodation. In fact, the LASSO method exhibits systematically lower values of bias and standard deviation, both for Italian and foreign travellers: considering the RMSE, the reduction in percentage terms is around 20 per cent for international transport and 10 per cent for accommodation. Looking at the residual component, the new method shows an increase of the bias with respect the donor approach; however, this increase is more than compensated by the reduction of the forecast error variability: in fact, the RMSE of the LASSO approach is still lower than the one obtained with the donor method.

The reduction of the bias for the international transport and accommodation components means that the model imposed to the data by the LASSO approach seems quite reasonable. Moreover, the variability of the imputation errors is lower in the case of LASSO given that in this method we need to estimate a vector of parameters, while the donor approach is fully non-parametric. These considerations explain the reason why our proposed approach outperforms the existing one.

The comparison of the forecasts for the 2015-2018 period proves the robustness of the main results of the analysis: in fact, the improvements gained with the new algorithm, in terms of bias and variance reductions, are confirmed. It also means that the new approach is capable to learn quickly possible changes in the structure of travellers' expenditures, which might have happened after the COVID-19 pandemic. On the contrary, the donor method, using only partially the information in the interviews, might require a longer time and many waves to identify enough twins to produce unbiased estimates after the pandemic outbreak.

Figure 4 decomposes the RMSE in the different years considered in our analysis, i.e. the period 2011-2018, for the three components and distinguishing between Italian and foreign travellers. It is possible to observe that in each period the LASSO approach exhibits more accurate forecasts than the donor method, with only few exceptions related to the residual component of the other expenditures.

Furthermore, in figure 5 the RMSE is evaluated for the counterpart countries with the most relevant tourism flows from/to Italy, i.e. Switzerland, Germany, Spain, France, United Kingdom and United States, which account for around 50 per cent of both credits and debits of the travel item. It is interesting to notice that RMSE obtained with the LASSO approach is, again with few exceptions, lower than the one realised with the donor method, especially for the accommodation and, above all, for the international transport item. The two approaches perform in a similar way for the residual component, where the additional variables selected by the LASSO have less predictive power.

7. Concluding remarks

The increasing diffusion of package tours, especially among foreign travellers, observed in the data of the Bank of Italy's International Tourism Survey, motivates the development of more sophisticated unbundling methodologies in order to impute accurately the different components of package tours and estimate consequently the international transport and travel items of the BoP.

In fact, the currently employed donor method has some limitations. The most important of these drawbacks is the requirement that the traveller and the associated twins should have the same characteristics: country of residence, mean of transport, length of the stay, type of accommodation, and reason of the trip. In fact, it is difficult to find enough twins to have reliable and stable estimates according to all these features and, consequently, some constraints must be relaxed with the risk of introducing bias in the estimates. This negative aspect is exacerbated since the donor approach is a fully non-parametric method which requires more observations than a parametric alternative.

In this paper, a ML approach is proposed with the aim of overcoming the limitations of the current donor method. The new approach improves the existing one in two directions: firstly, it models explicitly the relationship between the different components of a package and the characteristics of travellers and their trips in a parametric framework; secondly, it adopts a regularisation method, i.e. the LASSO, to automatically select the relevant features for the estimation of the package components.

The comparison of the out-of-sample forecasting performance of the two methods reveals that the ML algorithm generally outperforms the donor method in terms of more precise and, above all, less biased predictions. This is true considering the overall validation set as well as comparing the forecasts in each year selected for the analysis or the predictions for the major counterpart countries in terms of tourism flows from/to Italy.

The robustness of the ML approach, tested with a more recent sub-sample, is a further advantage in the production of reliable estimations in the presence of behavioural changes in travellers' expenditures, which might have occurred after the COVID-19 pandemic.

It is important to stress that in the analysis carried out in this paper we have made some minor simplifications, such as considering the interviews with strictly positive expenditure shares in all the three components, i.e., international transport, accommodation, and other expenditures. In the (few) cases in which a package does not include all the three items, but only two of them, the observed expenditure will be used for the service excluded from the package, while the model equations for the other two components will be employed to obtain the predicted shares in order to impute the unobserved expenditures. Moreover, the residual component called "other services" in this paper includes different services, like local transport, food-serving services, other services not included elsewhere, that will require ad-hoc models in the practical implementation of the proposed approach.

Despite these minor considerations, the evidence produced in this work should be enough to convince BoP compilers on the usefulness of ML methods to improve the unbundling of package tours.

References

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, *1*, 169-194.

Chakraborty, C., and Joseph, A. (2017). Machine Learning at Central Banks. *Bank of England Staff Working Paper*, 674.

Dalalyan, A. S., Hebiri, M., and Lederer, J. (2017). On the prediction performance of the lasso. *Bernoulli*, *23*(1), 552-581.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal $\ell 1$-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, *59*(6), 797-829.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, pp. 337-387). New York: Springer Series in Statistics.

Greenshtein, E., and Ritov, Y. A. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, *10*(6), 971-988.

IMF (2009), *Balance of Payments and International Investment Position Manual*, Sixth Edition (BPM6), Washington, D.C.: IMF.
https://www.imf.org/external/pubs/ft/bop/2007/pdf/bpm6.pdf

Knight, K., and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*, *28*(5), 1356-1378.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907-927.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, *52*(1), 374-393.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, *34*(3), 1436-1462.

Plan, Y., and Vershynin, R. (2016). The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, *62*(3), 1528-1537.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (methodological)*, 58 (1), 267–88.

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, *50*(10), 2231-2242.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, *52*(3), 1030-1051.

United Nations (2008). *International recommendations for tourism statistics 2008*. United Nations Publications.

United Nations (2010). *Manual on Statistics of International Trade in Services 2010*. United Nations Publications.

Wainwright, M. J. (2006). Estimating the ``Wrong'' Graphical Model: Benefits in the Computation-Limited Setting. *Journal of Machine Learning Research*, *7*(Sep), 1829-1859.

Zhao, P., and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, *7*(Nov), 2541-2563.

Fig. 4a: *Comparison of the RMSE of the LASSO and donor methods for Italian and foreign travellers, distinguishing between international transport, accommodation and other expenditures – selected years*

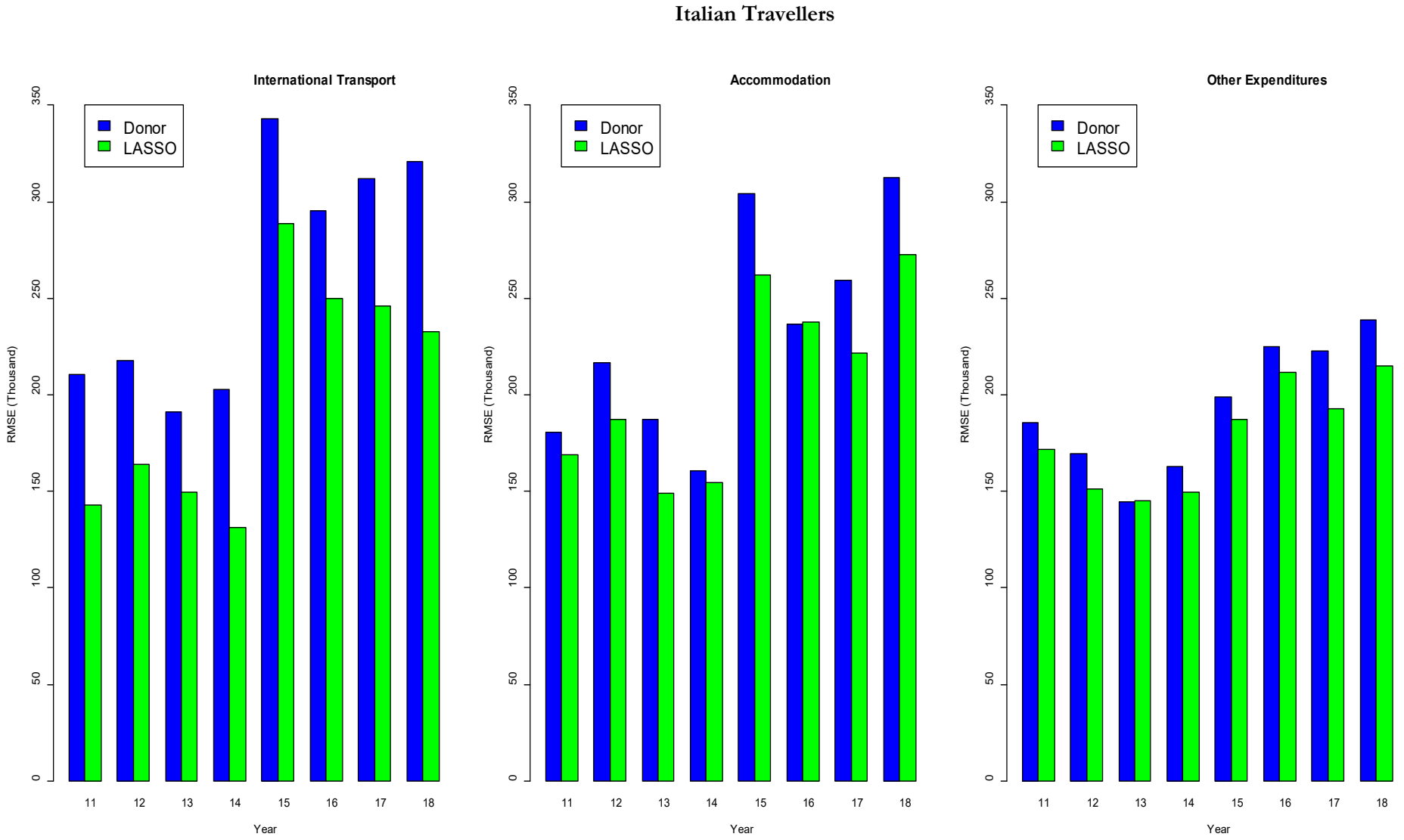


Fig. 4b: *Comparison of the RMSE of the LASSO and donor methods for Italian and foreign travellers, distinguishing between international transport, accommodation and other expenditures – selected years*

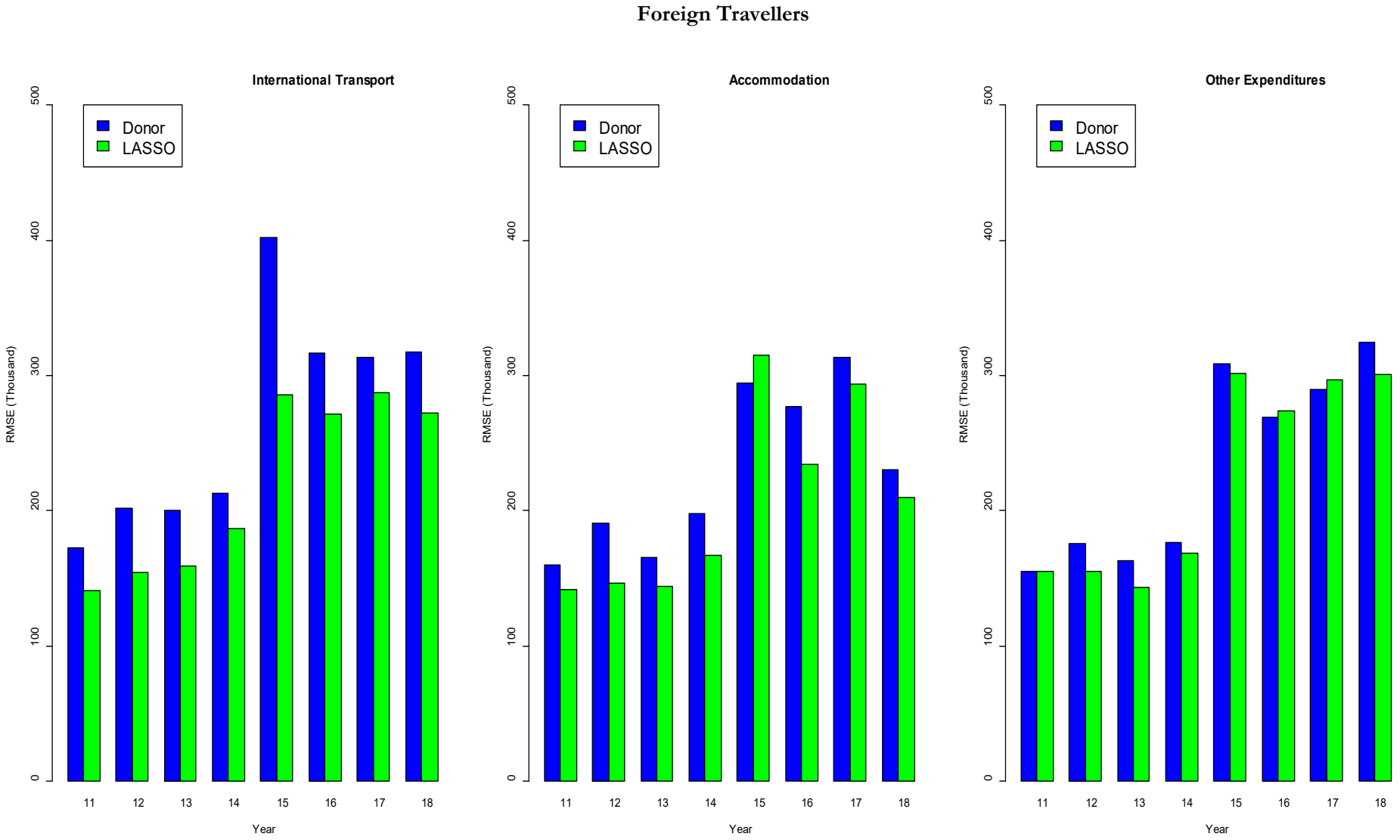


Fig. 5a: *Comparison of the RMSE of the LASSO and donor methods for Italian and foreign travellers, distinguishing between international transport, accommodation and other expenditures – main counterpart countries*

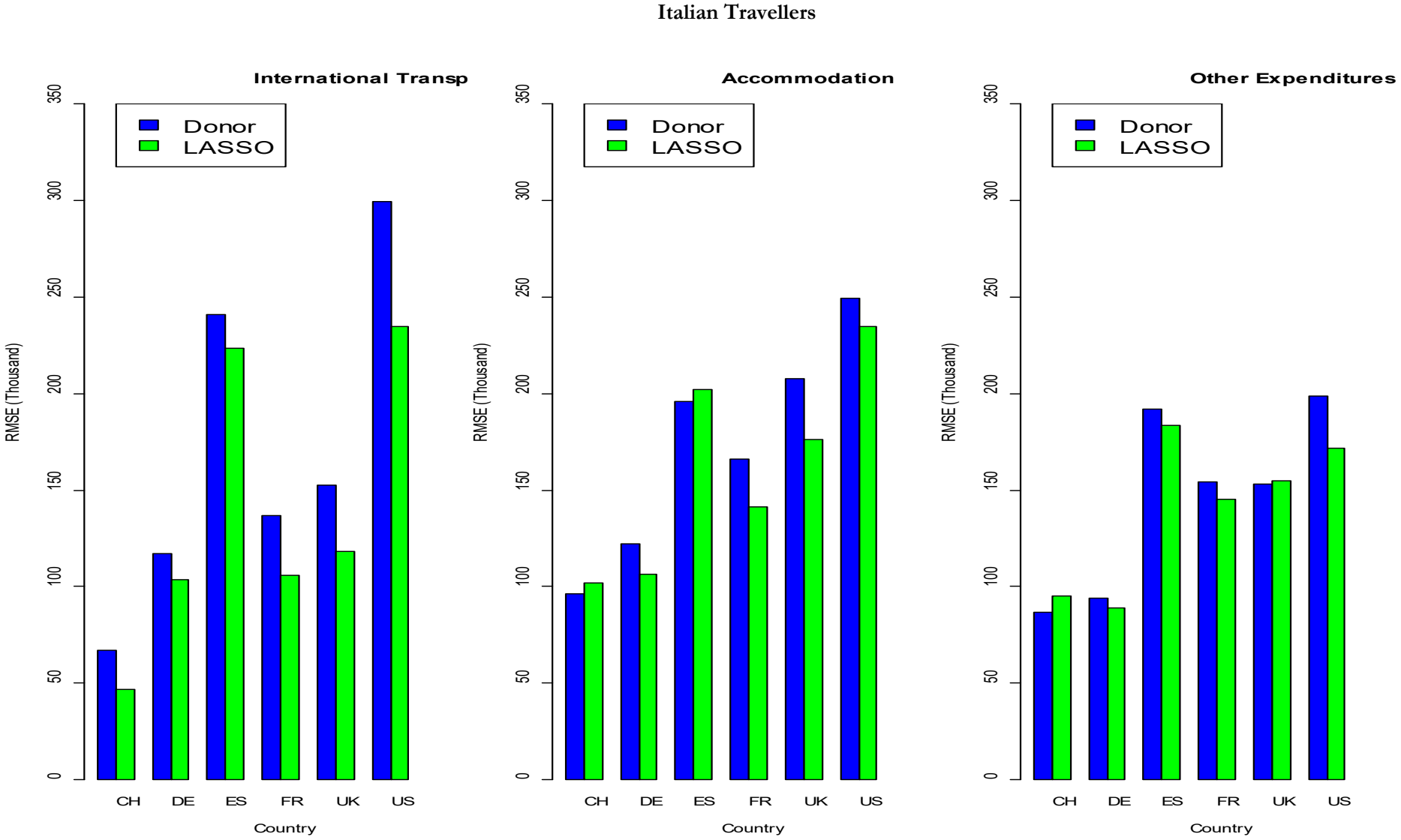


Fig. 5b: *Comparison of the RMSE of the LASSO and donor methods for Italian and foreign travellers, distinguishing between international transport, accommodation and other expenditures – main counterpart countries*