Unbundling Package Tours: a Machine Learning Application with the LASSO

Andrea Carboni* [1]; Claudio Doria[2]; Alessandro Moro[3]

[1] Banca d’Italia – Rome, Italy – andrea.carboni@bancaditalia.it
[2] Banca d’Italia – Rome, Italy – claudio.doria@bancaditalia.it
[3] Banca d’Italia – Rome, Italy – alessandro.moro2@bancaditalia.it

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:
Balance of payments; Travel; International transport; Linear regression; Donor method.

1. Introduction[1]:

The Bank of Italy carries out an extensive border 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 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 IMF Balance of Payments and International Investment Position Manual (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.

[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.

Moreover, according to international standards, 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, accommodation, 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 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 regularisation method in the ML literature, i.e. the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996).

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), Friedman et al. (2001), Donoho (2006), Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Bunea et al. (2007), Meinshausen (2007) and, more recently, Lee et al. (2016), Plan and Vershynin (2016), Dalalyan et al. (2017).

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.

2. Methodology:

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.[3] A bias in the estimation of the first component has also an effect in the correct allocation of the monetary flows between two BoP components, i.e. travel and international transport: from the compiler perspective, the priority in unbundling 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 accommodation in package tours, as almost all packages contain this component.

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 shares of “similar” travellers, the so-called twins, who have purchased the same services without buying a package (for details see Introduction).

In order to overcome the limitations of the donor method, a ML 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, it is possible to estimate the following relationship in which the share of expenditure of traveller *i* in item *j* (international transport, accommodation, other services) is explained by a set of features:[4]

$$(1)\;\; Q_{i,t}^{j} = \beta_0^{j} + \beta_{TD}^{j} TD_t + \beta_{CO}^{j} CO_{i,t} + \beta_{SD}^{j} SD_{i,t} + \beta_{TC}^{j} TC_{i,t} + \beta_{AC}^{j} AC_{i,t} + \varepsilon_{i,t}^{j}$$

where TD_t are 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 (1) can be estimated separately for Italian and foreign travellers without a package tour.

Then, the model can be applied to travellers that have bought a package tour in order to impute the unknown expenditure shares from their characteristics. In fact, for this latter kind of travellers, we know the input variables and the total value of the package, but we ignore the expenditures for the different items. The imputed expenditure shares $\hat{Q}_{i,t}^{j}$ are calculated as the predicted values of equation (1), rescaled in order to guarantee that $\sum_{j=1}^{3} \hat{Q}_{i,t}^{j} = 1$ (ruling out the few cases of negative predicted values).

[3] 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).

[4] 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.

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 (1). 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 (1) 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 ML 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:

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

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.

3. Result:

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 out-of-sample forecast errors with method *m* (donor or LASSO) are defined as:

$$(3)\ e_{i,t}^{j,m} = \left(Exp_{i,t}^j - TOT_{i,t}\hat{Q}_{i,t}^{j,m}\right) \cdot w_{i,t}$$

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 (3). Table 1 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.

Table 1: 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%

Notes: Bias, STD and RMSE are in euro.

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., 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 analysis: in fact, the improvements gained with the new algorithm, in terms of bias and variance reduction, are confirmed. It also suggests that the new approach will probably be 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.

4. Discussion and Conclusion:

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. 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:

1. Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1, 169-194.
2. Dalalyan, A. S., Hebiri, M., and Lederer, J. (2017). On the prediction performance of the lasso. *Bernoulli*, 23(1), 552-581.
3. Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ1-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.
4. Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, pp. 337-387). New York: Springer Series in Statistics.
5. IMF (2009), *Balance of Payments and International Investment Position Manual*, Sixth Edition (BPM6), Washington, D.C.: IMF
6. Knight, K., and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5), 1356-1378.
7. 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.
8. Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, 52(1), 374-393.
9. Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3), 1436-1462.
10. Plan, Y., and Vershynin, R. (2016). The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3), 1528-1537.
11. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (methodological),* 58 (1), 267–88.
12. Zhao, P., and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(Nov), 2541-2563.