Irving Fisher Committee on
Central Bank Statistics

11th Biennial IFC Conference on "Post-pandemic landscape for central bank statistics"

BIS Basel, 25-26 August 2022

How can big data improve the quality of tourism statistics? The Bank of Italy's experience in compiling the "travel" item of the Balance of Payments[1]

Andrea Carboni, Costanza Catalano and Claudio Doria,
Bank of Italy

[1] This presentation was prepared for the conference. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

How can big data improve the quality of tourism statistics?

The Bank of Italy’s experience in compiling the “travel” item of the Balance of Payments

Andrea Carboni, Costanza Catalano, Claudio Doria
Department of Economics, Statistics and Research, Bank of Italy

Abstract

In tourism statistics, the search for timelier and cheaper data sources than the traditional ones, like surveys, is becoming more and more important. In this paper, we investigate how mobile phone data (MPD), electronic payments data and internet search data (Google Trends) can improve the compilation of tourism statistics and the “travel” item of the Balance of Payments (BoP). We find that MPD have the characteristics to improve the estimates on the number of international travelers and to be integrated with the survey, although a constant interaction with the data supplier is required to define the phenomena to be caught. We highlight the limitations and the issues related to the use of electronic payment data for estimating expenditures in tourism statistics and we propose a model for producing timelier preliminary estimates for BoP purposes. Finally, we point out that Google Trends data can be used to complement the sample estimates of international travelers and to improve the quality of provisional data.

Keywords: Big data, tourism statistics, balance of payments, mobile phone data, payment data, Google Trends.

JEL classification: C20, C55, C80

Contents

1. Introduction

The use of big data is rapidly spreading in several fields as economics and statistics. National Central Banks play a role in this growing exploitation, as in general, official statistics follow a pressing and strictly defined calendar.[1] Therefore, timely information, as that coming from big data sources, is very attractive and potentially useful for compilers. Moreover, big data can be of great help as a supplementary source whenever information from traditional sources is difficult to obtain, time demanding and burdensome to acquire. In 2014, the UN Statistical Commission, recognizing the relevance of these new data sources, established the Global Working Group on Big Data for Official Statistics to promote the use of big data for compiling official statistics.[2]

Against this scenario, the Bank of Italy has carried out an in-depth analysis to understand whether, and how, big data can enhance the data production of the BoP "travel" item. This paper illustrates the results of this research, focusing on mobile phone data, electronic payment statistics (credit/debit cards) and web research information (Google trends), and discussing how the traditional approach for the compilation of the "travel" item can be improved by their use.

The paper is structured as follows. The next section describes the Bank of Italy's traditional methodology for the compilation of the "travel" item. Sections 3, 4 and 5 respectively illustrate the three research paths – based on the use of mobile phone data, electronic payments statistics and Google Trends data - developed for improving and validating the compilation approach. The final section summarizes the main findings and conclusions.

2. The estimate of the "travel" item in the Italian Balance of Payments

The Balance of Payments (BoP) is a statement that summarizes the economic transactions between residents and non-residents during a specific period.[3] The "travel" item of the BoP covers the monetary value of goods and services acquired in a country by non-resident travelers[4], in relation to visits to that country (BPM6 10.86),

[1] E.g. EU Member States disseminate provisional statistics of Balance of Payment on a monthly basis at t+45 days.

[2] https://unstats.un.org/unsd/statcom/decisions/, 45th Session (2014).

[3] As defined in the Balance of Payments and International Investment Position Manual, Sixth Edition, International Monetary Fund, (BPM6).

[4] Only transactions between residents and non-residents are relevant. The definition of "residence" is economic and not administrative: the country of residence of an international traveller is where the centre of her economic interest is located. To ease the exposition, we use for travellers the improper terms of "Italian" and "foreigner" when referring to their residence.

with the exception of the expenses for transport incurred to reach it[5], which are instead recorded under the "passenger transport" item. The BoP compilation standards of "travel" require a breakdown by counterpart countries and by purpose of the visit.[6]

Countries can adopt different methodological approaches[7] for compiling this item, based on the relevance of tourism in their economy, the characteristics of the border points, the administrative controls on incoming and outgoing flows and, of course, the budget constraints.

Since 1996, the Bank of Italy has been collecting the relevant information (number of international travelers, expenditures, length of the trips) through a sample survey carried out at border points; the data collected are then integrated, when available, with administrative sources.[8]

From a methodological point of view, the survey consists of two operations, carried out at each of the selected border points: counting and interviewing.

Counting aims at estimating the reference population, i.e. the total number of travelers entering or leaving Italy, broken down by country of residence or destination, on a monthly basis. In a selected interval of time, all the travelers crossing the border are counted and their residence is registered. Since having permanent counting operations on all the borders is not feasible, a grossing up algorithm is necessary for estimating the total amount of the international travelers crossing Italian borders during the reference period. Where administrative data are available, as for airports, they integrate the sample survey.

The second main operation consists in interviewing a sample of the travelers passing through each selected border point. The interviews primarily collect data on the expenditures and other relevant aspects for BoP purposes (e.g. reason, counterpart country), but also gather information that allows a broader analysis on tourism related topics, such as the means of payment and the type of accommodation. The interviews are carried out at the end of the stay, which is when the memories of the traveler about the trip are the most recent and reliable and all the expenses have already been determined so no guess by the traveler is necessary. Interviews are realized through questionnaires: each questionnaire refers to the group

[5] E.g. flight tickets, international train tickets, tolls etc. The expenses for transports within the visited economy are instead included in the travel item.

[6] Personal vs. business travels.

[7] In Europe, the main data sources for compiling the "travel" item are (frontier and households) sample surveys; some countries integrate these information with payment statistics as an additional source or for control purposes.

[8] E.g. the number of international travellers published by airports, ports authorities and railroads companies.

of people, if any, that shares the expenses of the trip with the interviewed traveler (e.g. a family).[9]

At each border point, interviewing and counting are carried out, as much as possible, at the same time, so that the characteristics of the interviewed traveler are coherent with those of the counted sample.[10] The information acquired with the questionnaires is then grossed up to the reference population, by taking into account the stratification variables listed in Table 1. Annually about 100,000 interviews and 1,000,000 counting operations are realized in more than 60 frontier points. This size ensures that the sampling error of the total international travel expenditure estimates is small and the statistics for the main partner countries are reliable.

Table 1 - Stratification variables and corresponding levels

VARIABLE	LEVELS
1. Direction	2 (inbound, outbound)
2. Type of carrier	4 (road, rail, airport and seaport)
3. Frontier point	62 (22 roads, 4 rails, 25 airports, 11 seaports)
4. Day of data collection	number of days in the month (e.g. 31)
5. Time of the day	3 (first shift, second shift, third shift)

3. The mobile phone data experiment

Mobile phone data (MPD) are one of the most promising big data sources for the study of many social and economic phenomena and behaviors. Several pilot studies in the literature analyze the potentiality of MPD, e.g. for computing the population of an area (Deville et al., 2014), for estimating the population density (Ricciato et al., 2015), for traffic statistics (Janecek et al., 2015) for transport and urban planning (Lokanathan et al., 2016) and, finally, for travel statistics (Ahas et al. 2007; Ahas et al. 2008; Ahas et al., 2014). In this regard, the contributions of the Estonia Central Bank[11] and the Banque de France[12] have also to be mentioned.

While MPD data can provide a great amount of information (number of international travelers, proxy of the country of residence, locations visited, length of stay, etc.) they say nothing about the expenses, the main variable to be estimated in

9 For example, if the interviewed traveler shares the expenses with another traveler and they spend a total of 100 euros, we count two trips, each with a total expense of 50 euros.

10 An interviewed traveler is not necessarily part of the counted ones. For example, on road borders it is very difficult to interview the passengers of the vehicles counted while passing through the frontier; it becomes possible only when there are checkpoints and a collaboration with frontier authorities is arranged.

11 https://statistika.eestipank.ee/failid/mbo/valisreisid_eng.html
https://statistika.eestipank.ee/failid/mbo/kv_mb2_eng.html

12 https://blocnotesdeleco.banque-france.fr/en/blog-entry/measuring-international-travel-during-covid-19-pandemic,

https://unstats.un.org/bigdata/task-teams/mobile-phone/MPD%20Handbook%2020191004.pdf the BoP “travel” item. MPD can thus only be considered as a complementary source of information, useful to estimate the dimension and some characteristics of the reference population.

In 2018, the Bank of Italy started a test phase with the purpose of integrating MPD into the international frontier survey, in order to gradually replace the counting operations. Counting is in fact a costly and demanding activity, and this is particularly true for road borders, given the high number of this type of frontier points in Italy and the scarcity of administrative data, and for seaports, due to restricted access zones as the ones reserved to cruise ships. These problems might affect the quality of the grossing-up factors and hence of the estimated values.

MPD may represent an alternative, efficient and less costly data source to count travelers crossing the frontiers. The arrival of a foreign traveler at the Italian border is signaled by the connection of a mobile phone, with a SIM card[13] issued by a non-resident phone operator, to the cells controlled by an Italian phone-operator. Likewise, the disappearance for some period of time of the signal of an Italian SIM card near the border would indicate that this traveler has gone abroad.

The Bank of Italy collaborated with one of the major Italian Mobile Network Operator[14] (MNO) to develop an algorithm for the estimate of travelers’ inflows and outflows through each border point by exploiting the MPD. These data are not “ready to use” for BoP purposes and a close, constant cooperation between the Bank of Italy and the MNO has been necessary to define the best metrics to elaborate the raw data and achieve measures compatible with the BoP standards. For example, it is necessary to define the minimum docking time of a foreign SIM card to a cell located in Italy for it to be considered associated to an international traveler present in Italy. This problem is very relevant near the road borders due to handover effects.

Since each frontier point has specific features that should be incorporated in the final algorithm, a test phase was developed for two important Italian border points: the main airport of Rome (Fiumicino), which is the largest in Italy in terms of international traffic, and the highway frontier point of Tarvisio, one of the most relevant in the North-East of Italy.

For the Fiumicino airport, the traditional survey is supported by data provided by the company that manages the airport, Aeroporti di Roma (ADR). This source is used for correcting, by means of calibrated estimators (see Deville and Särndal 1992), the estimate of the total international flows derived from the counting operations, although it does not provide information on the residence of the passengers.

Tables 2 and 3 compare the MPD, the ADR[15] statistics and the Bank of Italy’s official statistics on the Fiumicino airport for the period August 2018 - June 2019. The MPD

[13] Subscriber Identity Module.

[14] 31% of the market share in 2018.

[15] Differences between the official data grand total and ADR statistics are due to minor adjustment in the Bank of Italy estimation process.

and the ADR grand totals are broadly aligned, with the MPD always larger than the official data (BI). As for the breakdown residents/non-residents, the number of Italian travelers estimated by the MPD is in line with the one estimated by the Bank of Italy (BI). On the other side, the number of foreign travelers estimated by the MNO is always greater than the one estimated by the BI.

Table 2 – Fiumicino airport: comparison between MPD and ADR statistics on the number of international passengers

	MPD(1)	ADR(2)	MPD/ADR %
Aug-18	1,802,051	1,679,511	7.3
Sep-18	1,723,145	1,521,956	13.2
Oct-18	1,590,179	1,437,316	10.6
Nov-18	1,220,903	1,083,621	12.7
Dec-18	1,045,675	1,066,898	-2.0
Jan-19	1,113,629	989,903	12.5
Total	8,495,582	7,779,205	9.2

(1) Estimates based on mobile phone data.

(2) Aeroporti di Roma administrative data on passenger transits at Fiumicino airport.

Table 3 – Fiumicino airport: comparison between MPD and BI statistics on the number of international passengers

	TOTAL			ITALIANS			FOREIGNERS		
	BI(1)	MPD(2)	MPD/BI%	BI(1)	MPD(2)	MPD/BI%	BI(1)	MPD(2)	MPD/BI%
Aug-18	1,717,076	1,802,051	4.9	640,288	621,419	-2.9	1,076,788	1,180,632	9.6
Sep-18	1,574,571	1,723,145	9.4	446,884	516,638	15.6	1,127,687	1,206,507	7.0
Oct-18	1,380,639	1,590,179	15.2	423,402	449,204	6.1	957,237	1,140,975	19.2
Nov-18	1,053,956	1,220,903	15.8	392,909	466,087	18.6	661,047	754,816	14.2
Dec-18	1,037,503	1,045,675	0.8	506,530	417,820	-17.5	530,973	627,855	18.2
Jan-19	831,120	1,113,629	34.0	344,529	457,947	32.9	486,591	655,682	34.8
Total	7,594,865	8,495,582	11.9	2,754,542	2,929,115	6.3	4,840,323	5,566,467	15.0

(1) Bank of Italy official statistics.

(2) Estimates based on mobile phone data.

The estimate of the number of international travelers crossing Tarvisio border only relies on counting operations, due to the lack of complementary administrative sources.

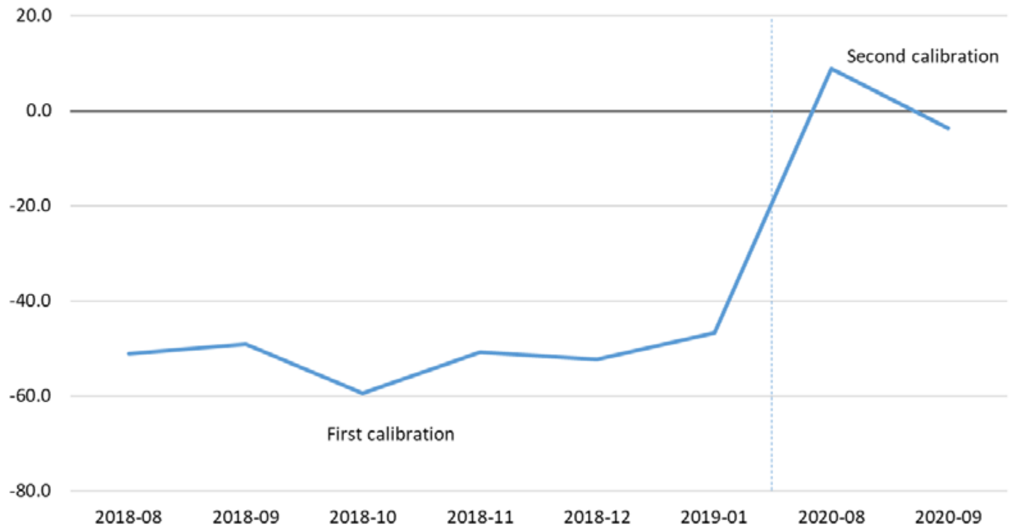
Table 4 compares the Bank of Italy's and the MPD statistics in this road border point on the same period: the differences are very large, and they are sharper for Italian travelers than for foreigners.

Table 4 – Tarvisio: comparison between Bank of Italy and MPD statistics on the number of international passengers

	TOTAL			ITALIANS		FOREIGNERS	
	BI	MPD	MPD/BI%	BI	MPD	BI	MPD
Aug-18	2,005,595	980,066	-51.1	662,710	115,841	1,342,885	864,225
Sep-18	1,544,727	785,843	-49.1	408,999	78,895	1,135,728	706,948
Oct-18	1,026,265	416,988	-59.4	261,135	64,948	765,130	352,040
Nov-18	691,340	340,325	-50.8	212,436	80,532	478,904	259,793
Dec-18	600,309	285,953	-52.4	272,065	103,197	328,244	182,756
Jan-19	686,784	366,037	-46.7	219,408	128,927	467,376	237,111
Total	6,555,020	3,175,212	-51.6	2,036,753	572,340	4,518,267	2,602,873

Further interactions with the mobile network operator led to a shortening (from four hours to 30 minutes) of the minimum time a foreign/Italian SIM card has to spend on the national/foreign territory in order to be considered an international traveler, and thus to a recalibration of the algorithm. This resulted in a new test, only involving the months of August and September 2020: the result showed a good convergence on the grand total between the two data sources, with a great improvement compared to the first release (Figure 1). On the other hand, the distribution between resident and non-resident travelers was still quite different, suggesting the need to continue investigating the causes underneath different estimates.

Figure 1 - % differences between MPD and BI statistics on total travelers' flows



4. The payment statistics analysis

Similarly to mobile phone data, electronic payment data are a promising source for the measurement and study of social and economic phenomena, including the production of statistics on national and international expenditure (Dubreuil 2017, Demunter 2017). Recently, they have started to be used for tourism statistics (Li et.

Al. 2018), in particular by international institutes and national central banks such as the Banco de Portugal (Coelho et. Al. 2011), the Banque de France[16] and the Central Bank of Armenia (Yezekyan. 2018). Moreover, the European Central Bank recently approved a regulation[17] on payment statistics also with the purpose of gathering data that the Eurozone countries could use for the compilation of their external statistics.[18]

Electronic payment data are attractive because of their timeliness, relative ease in collection and processing and moderate costs; moreover, their availability is not subject to high-impact perturbative phenomena like the Covid-19 pandemic. The steady increase of the share of electronic payments on total expenditure (Ardizzi et. Al. 2021) will keep strengthening the informative power of this source, although all the other possible means of payment such as transactions made by cash, bank transfers, etc. have to be estimated with other sources.

Against this background, the Bank of Italy conducted an explorative analysis in order to assess if and to what extent electronic payment data can contribute to the production of the "travel" item of the BoP and/or can be used for checking the consistency with the tourism statistics.

For this purpose, two databases were considered, provided by one of the main paytech companies operating in Italy, with data spanning from May 2014 to August 2021. The market share of this company was unknown, making impossible the grossing-up of raw data. One database contains all the electronic payments made by credit and debits cards on POS[19] (physical database), while the other includes the on-line (e-commerce) transactions. Both databases are divided into acquiring and issuing: the first one contains the transactions made by foreigners cards on Italian POS and websites (potentially contributing to estimate the foreigners' expenditures in Italy), while the latter includes the transactions on foreign POS and websites made by Italian cards (potentially contributing to estimate the Italians' expenditures abroad).

Every record of the databases is made up of five variables: the date (day-month-year) of the payment, the Merchant Category Code (MCC) identifying the type of purchase[20], the nationality of the bank emitting the payment card, the country of the POS/website where the payment has been made and the transactions amount in euro.[21]

There are major limitations in electronic payments data for the compilation of official statistics on travel:

1. The nationality of the bank issuing the card is just a proxy of the residence of the traveler;

16 https://blocnotesdeleco.banque-france.fr/en/blog-entry/measuring-international-travel-during-covid-19-pandemic

17 Regulation ECB/2020/59.

18 https://www.ecb.europa.eu/press/govcdec/otherdec/2020/html/ecb.gc201211~ef1f417b08.en.html

19 Point Of Sale.

20 The available MCC are the following: clothing, hotels and restaurants, groceries, home, cash advance, work, retail, services, mobile web, travels and transports.

21 The amount is the aggregation of all the transactions sharing the same values of the first four variables.

2. Confidentiality issues allow the use of only aggregated data, which may increase the difficulties in discerning the transactions that are related to tourism from the ones that are not;
3. There is no information on the reason of the trip (business/personal), which is a mandatory BoP requirement;
4. There are difficulties in registering and correctly classifying the Digital International Platforms (DIP) transactions, in terms of misallocation issues for the counterpart country and of failure in recording some transactions. Three main examples could help understand the matter:
   I. The payment of a stay in Paris made by an Italian tourist on the Booking.com platform[22] is recorded as a transaction from Italy to The Netherlands and not to France, as it should be recorded in the BoP;
   II. The payment on Airbnb[23] of an accommodation in Rome by a French traveler is recorded as a transaction from France to Ireland, thus not appearing in our database, although it should be recorded in the BoP;[24]
   III. The payment on Airbnb of an accommodation in Rome by an Italian traveler is recorded as a transaction from Italy to Ireland, although it refers to a domestic trip and thus should not be recorded in the BoP.

Figure 2 compares the official BoP data on foreigner travelers' expenditure in Italy (only by means of electronic cards) with the grand totals of, respectively, the electronic payments recorded in the physical acquiring database, in the e-commerce acquiring database, and the sum of the two. The level of the e-commerce transactions is much lower than the other time series. Indeed, it does not cover the transactions of item 4-II: the large use of these platforms, which mostly have foreign headquarters, can explain its negligible levels.

Figure 3 compares the official BoP data on Italian travelers' expenditure abroad (only by means of electronic cards) with the grand totals of, respectively, the electronic payments from the physical issuing database, the e-commerce issuing database and the sum of the two. The e-commerce time series shows higher levels than for the acquiring side, although it does not show the typical seasonality of the tourism phenomena, which has peaks in the summer. This is probably because such database contains large amounts of on-line transactions that are not connected to tourism, as the purchases of goods on Amazon, Ebay, etc. The low granularity of the database does not always allow to distinguish them, as some categories contain both transactions that are related to travel and transactions that are not.

[22] Whose legal headquarter is in The Netherlands.

[23] Whose legal headquarter is in Ireland.

[24] The digital platform can carry out a further transaction with an Italian counterpart, but not necessarily using a credit card.

Figure 2 - Foreigner travelers’ expenditure in Italy

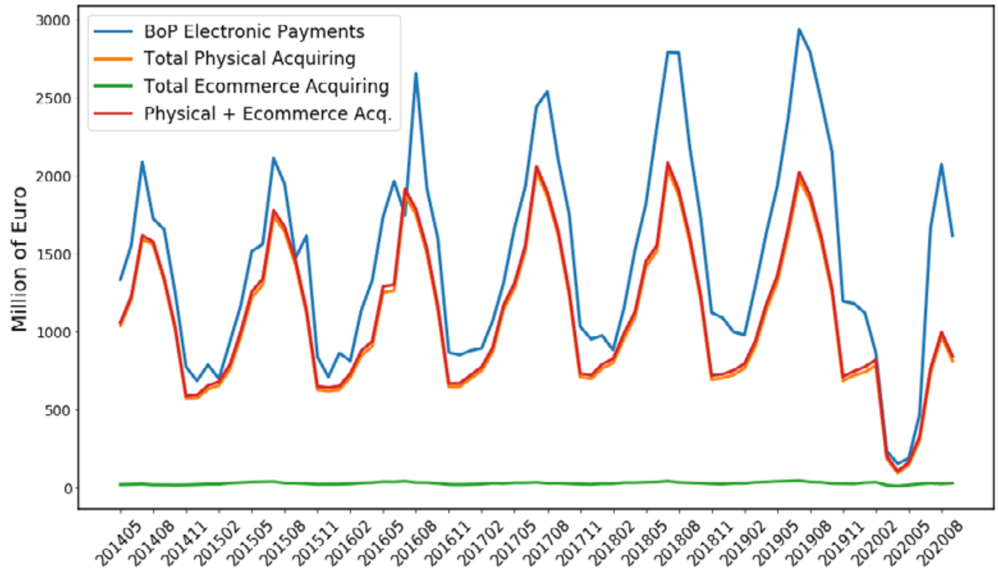
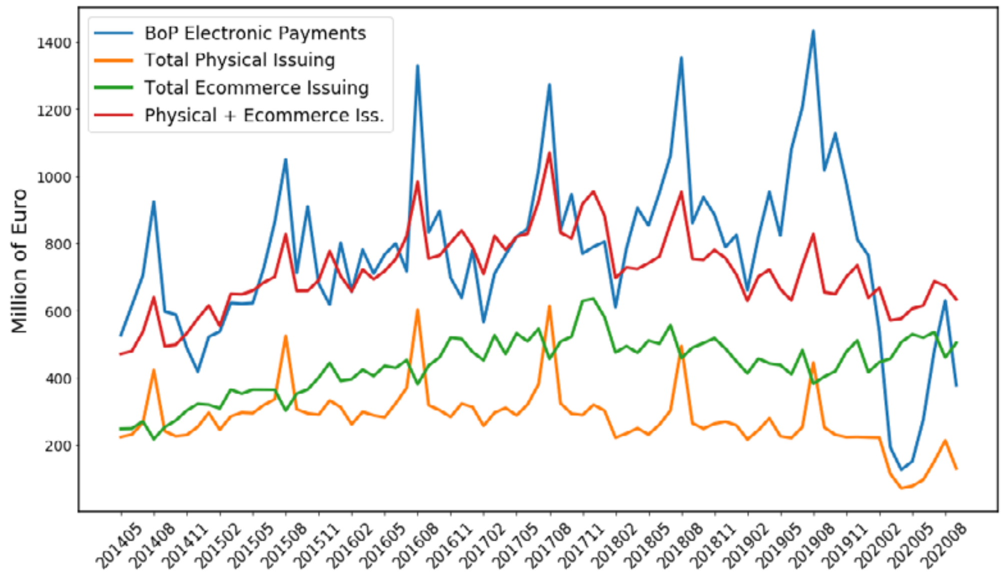


Figure 3 - Italian travelers’ expenditure abroad



In an attempt to push further the exploratory analysis, the following Machine Learning models have been tested: Ridge and Lasso models,[25] regression trees and boosted regression trees. Due to the relative short length of the whole payment data time series,[26] the years 2015-2017 have been used as training set, the year 2018 as validation set[27] and the year 2019 as test set, in order to verify the model out-of-

[25] With and without enforcing positive coefficients.

[26] 60 months from January 2015 to December 2019, plus 20 months (January 2020-August 2021) during the Covid-19 pandemic.

[27] It was used to select the best parameter λ for the Lasso and Ridge models minimizing the MSE on the validation set.

sample using the MSE index.[28] Moreover, the Covid-19 years 2020-2021[29] were used as supplementary test set for verifying the robustness of the model to external shocks. In each model, the dependent variable is the BoP travel item total, while the independent variables are all the physical MCC data (at lag 0) plus all the e-commerce MCC data for all lags from 0 to -4.

Table 5 reports the performance of such models in terms of the MSE index[30] on the validation, test and Covid set. On the acquiring side, almost all the models show a quite good performance; in particular, the LASSO model with positive coefficients and the regression tree shows the smallest MSE on the test set and perform quite well in the Covid one. On the issuing side, the performance of each model is worse than in the acquiring case, as expected. The best results are obtained by the LASSO models, which have an unexpected good performance in the Covid set.

Table 5- Model performance in predicting the BoP 'travel' item grand total on the different sets

	ACQUIRING			ISSUING		
	MSE val	MSE test	MSE covid	MSE val	MSE test	MSE covid
Ridge	0,09	0,76	3,82	0,13	0,39	0,53
Lasso	0,04	0,29	1,74	0,08	0,57	0,15
Lasso pos. coeff.	0,04	0,3	1,79	0,15	1,29	0,16
Regression tree	0,2	0,28	1,15	0,7	1,93	2,33
Boosted tree	0,1	0,32	1,23	0,48	2,04	2,6

Note: val= validation set, test=test set, covid=Covid years 2020-2021. Acquiring: foreigner travelers' expenditures in Italy; Issuing: Italian travelers' expenditures abroad.

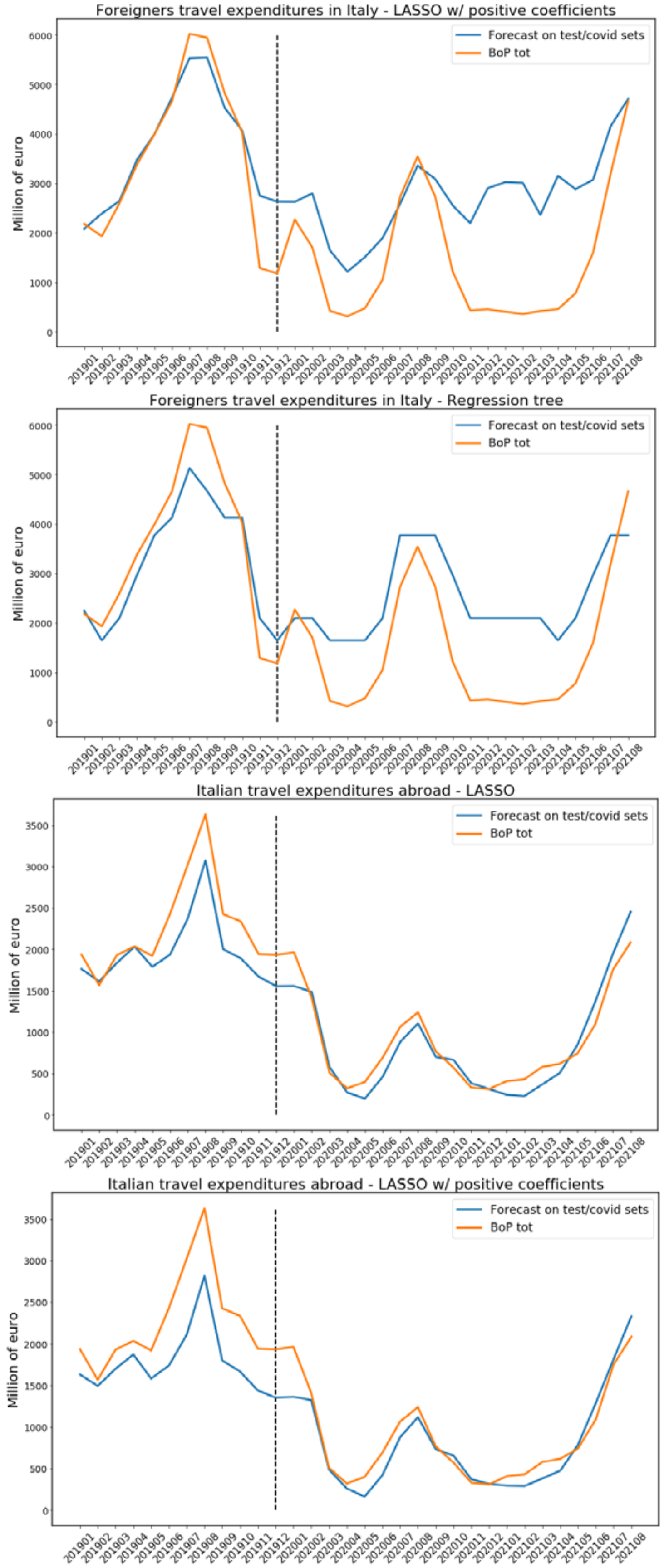
Figure 4 shows the plots of the forecasts in the test and Covid sets compared with the official BoP figures. The graphs confirm what was already pointed out: in forecasting foreigners' travel expenses in Italy, we obtain a good performance on the test set, while the forecast behaves quite poorly in the subsequent years affected by the Covid pandemic. On the other hand, the forecast of the Italians' travel expenses abroad is worse in the test set, but surprisingly good in the Covid one, as the trend is fully captured.

28 For the regression tree model, both 2018 and 2019 are used as test sets, fixing the maximum high of the tree to 4.

29 We will call it the 'Covid set'.

30 Mean Squared Error. The smaller the MSE is, the better the model prediction is.

Figure 4 – Forecasts on test set and Covid set for selected models



Note: Test set on the left of the black dashed vertical line, Covid set on the right.

5. The Google Trends experiments

The third experimentation relies on the use of Google Trends as a complementary source to estimate the "travel" item in the BoP compilation process, in particular to assess the number of international travelers in Italy.

Google Trends (GT) is a website[31] provided by Google that reports the popularity of search queries in the Google search engine over time and across various regions of the world. The popularity of a given query is measured by an index between 0 and 100 (the maximum frequency). The data are collected and aggregated continuously on a daily, weekly or monthly basis. One can visualize the popularity of the selected query by specifying the state or region, the category they belong to[32], and the time frame of interest. Timeliness is one of the main advantages of this website, as data are updated almost in real time.

In order to understand if this kind of data can be usefully employed, a specific predictive exercise was developed with the aim of forecasting the number of travelers visiting Italy in the period from January 2006 to May 2019. In particular, the tourist flows from the most important counterpart countries in terms of arrivals - France, Germany, United Kingdom, United States and Spain - were considered.

The predictive variable of GT was defined by considering the frequency of the search queries performed in the aforementioned countries that contain the word "Italy" in the category "Travel". For each of these selected countries, a seasonal AR(1) process was used for modelling the number of travelers $N_{c,t}$ arrived in Italy during month t from country c according to the Bank of Italy's tourism survey, where the l-period lagged Google Trends index $GT_{c,t-l}$ is included as an exogenous regressor:

$$N_{c,t} = \phi_0 + \phi_1 N_{c,t-1} + \phi_{12} N_{c,t-12} + \beta GT_{c,t-l} + \varepsilon_{c,t} \quad (4) .$$

The most suitable lag of the GT index is chosen by minimizing errors of the out-of-sample forecasting performance, measured in terms of Mean Squared Error (MSE) reduction. In particular, the months between September 2012 and May 2019 have been considered to compare the observed value and the one-step-ahead forecasts, with an expanding windows approach[33]. In all cases, except for France where the coefficient β is not statistically different from zero, the GT index increased the performance of the predictive model.

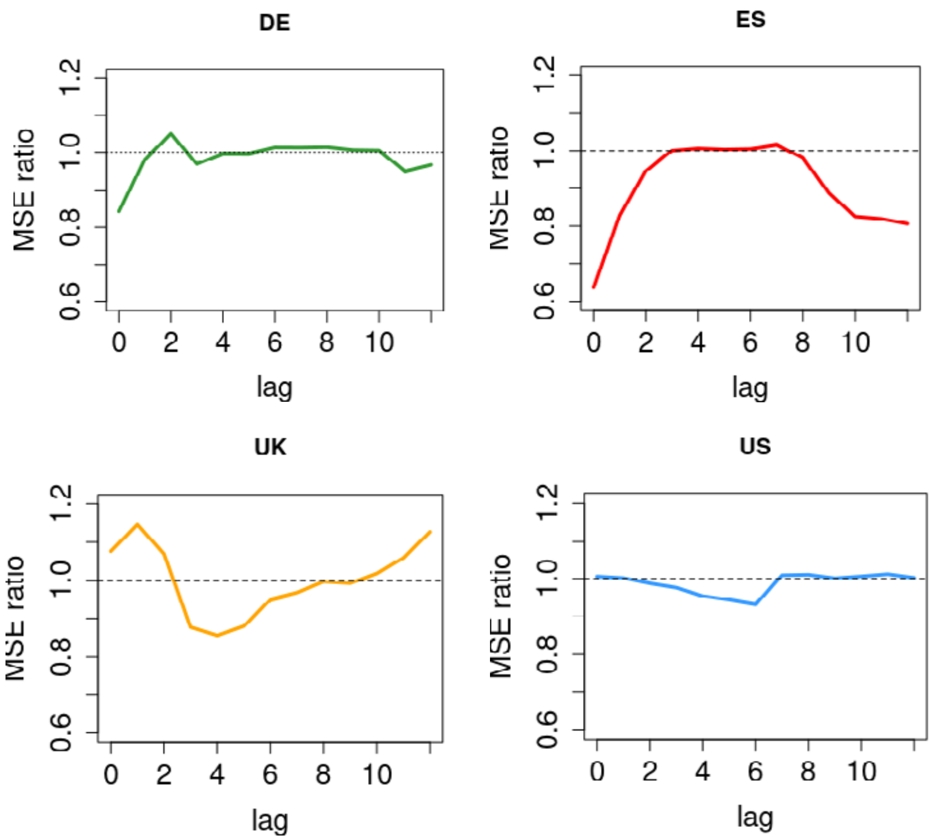
Figure 5 shows how the ratio between the MSE of specification (4) and the one obtained using the model without $GT_{c,t-l}$ ($\beta = 0$) depends on the lag for the different countries considered.

[31] https://trends.google.com/trends/?geo=IT

[32] More than 20 categories of search, which helps avoiding multiple meanings for the chosen query.

[33] Adding an observation at each step.

Figure 5 - Out of sample normalized MSE for different lags of the GT index



The contemporaneous variable GT_t (l=0) is the best predictor for Germany and Spain, while the variable at lag l=4 and l=6 minimizes the MSE for UK and US respectively. These last results seem only partially reasonable: US travelers may have to organize their trips towards Italy more in advance than German and Spanish travellers; moreover, it is possible that Google Trends classifies in the Travel category web searches performed by tourists during their travel; this should increase the weight for lag l=0 in the model; less clear is the situation for UK, where there is not an intuitive explanation for a better performance of lag l=4 in comparison to smaller lags.

The estimates of specification (4) for each country involved in the exercise are shown in Table 6: the GT index is always highly significant and the model indicates a good in-sample fit, measured by a high value of the R^2. However, each time series presents a strong auto-regressive component[34] and the marginal contribution of the GT index is significant only for Spain[35]. The negative sign of GT coefficient in the UK and US regressions means that the variable is not robust enough for these two countries, confirming the doubts in the interpretation of the optimal lag.

[34] The number of travellers at lag 1 and 12 are significant at 95% for all the considered countries.

[35] For Spain, the R^2 adjusted is 0.67 in the model without GT, and 0.77 in the one with the variable included. For the other countries, the R^2 is near to 0.9 in the model with only the AR component and the addition of the GT index only increases it of around 0.01.

Table 6 – Estimates of the model for different countries

	(1) DE	(2) ES	(3) UK	(4) US
N_{t-1}	0.20***	0.34***	0.33***	0.13***
	(0.04)	(0.05)	(0.05)	(0.04)
N_{t-12}	0.73***	0.46***	0.76***	0.89***
	(0.05)	(0.05)	(0.05)	(0.04)
GT_{t-1}	6.01***	2.11***	-1.50***	-0.91***
	(1.17)	(0.29)	(0.44)	(0.24)
Const	-182.49***	-22.78*	29.87*	38.84***
	(43.34)	(12.07)	(15.52)	(13.93)
R^2	0.92	0.77	0.89	0.92

* p<0.10, ** p<0.05, *** p<0.01

In order to examine the predictive performance of the model Figure 6 compares, for each selected country, the observed value of the number of travellers to Italy and the one-step-ahead predicted levels in the out-of-sample period September 2012 - May 2019. The model seems to capture well the fluctuations of the phenomenon and the main turning points.

All in all, although this data source proved to be interesting during the analysed period, the Covid-19 pandemic pointed out its limits. Indeed, in March 2020 we witness a peak of search queries including the word "Italy" (see Figure 7 for selected countries), while in that month Italy was blocking the tourist inflow because of the pandemic; such peaks very likely reflect the interest by the users in understanding the developing of the pandemic or in checking whether travelling to Italy was still doable or safe, even if were not necessarily followed by actual travels.[36] Indeed, in presence of extraordinary events, the Google classification seems to be less effective and the risk of outliers, given by false positive searches not related to tourism, increases significantly.

Moreover, since the use of Google Trends strictly depends on the keywords included in the analysis, the use of other words as search queries, for example referring to specific Italian locations, could generate more accurate results.

Figure 6 – Observed (solid line) and predicted (dashed line) number of travelers to Italy from Germany, Spain, United Kingdom and United States

[36] This might explain why the peak in search queries appears also by considering only the GT category "Travel".

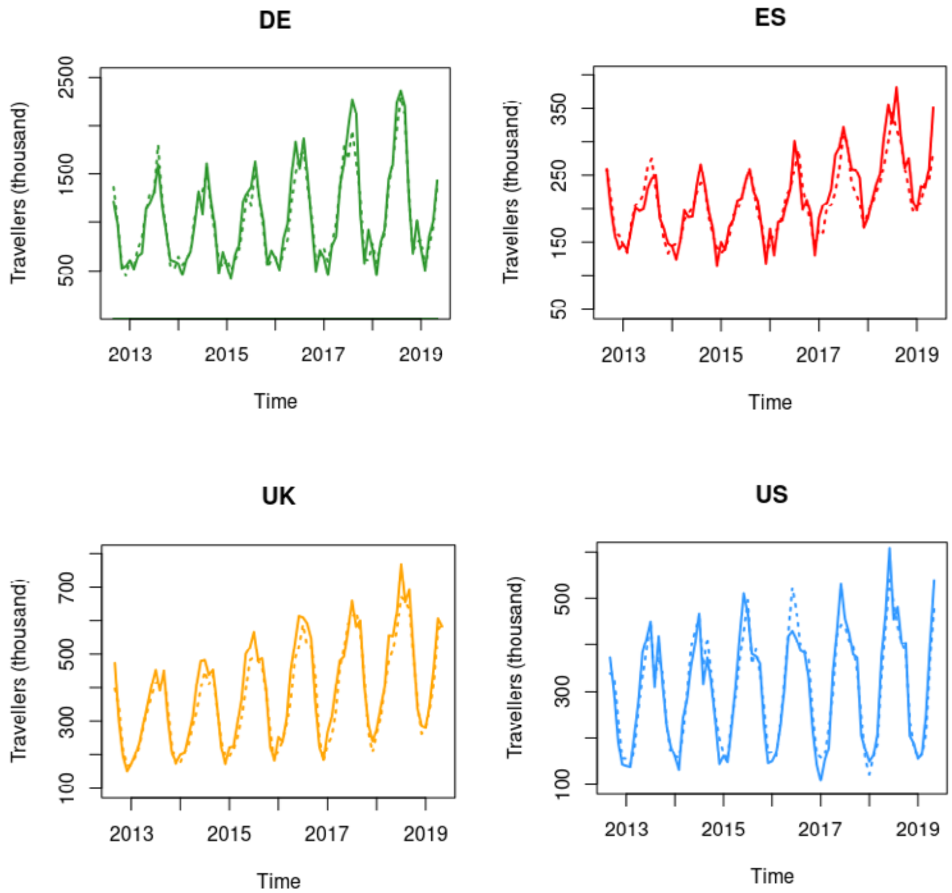
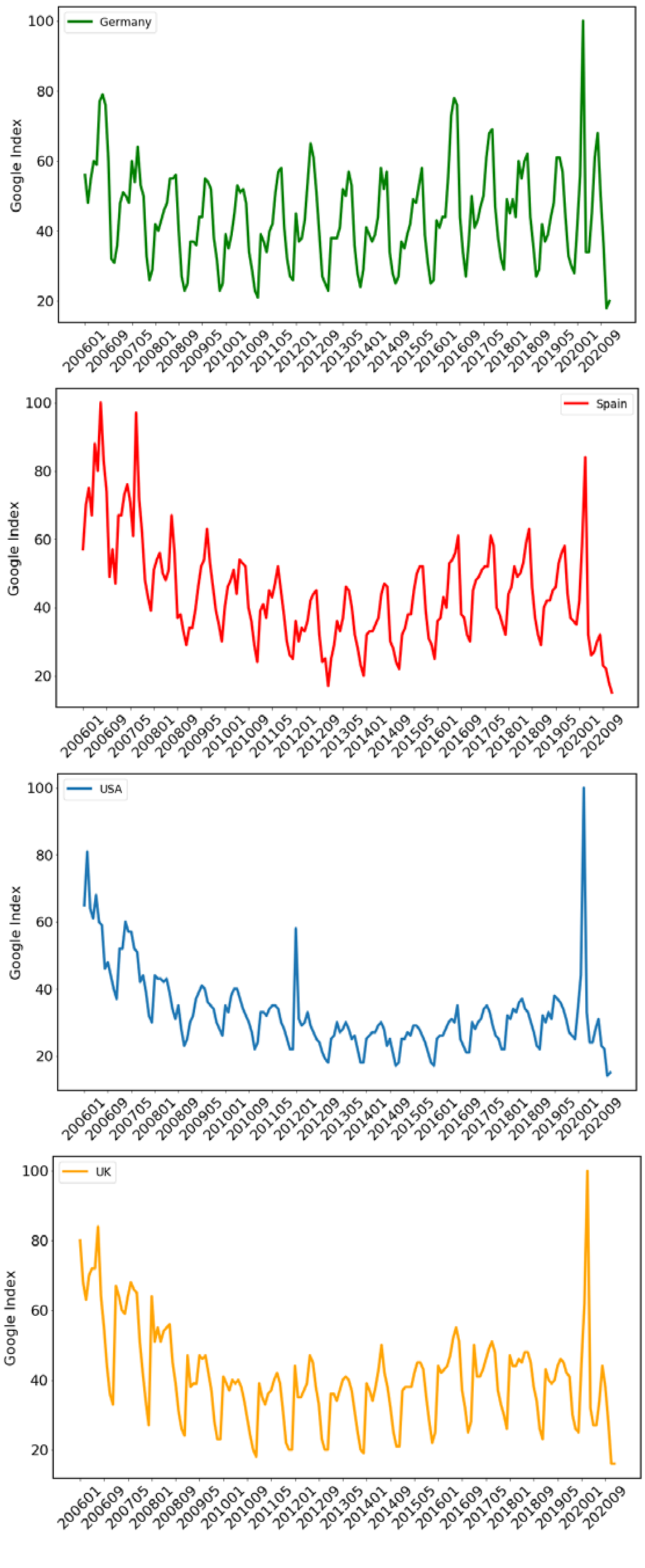

DE
Travellers (thousand)
500
1500
2500
2013
2015
2017
2019
Time
ES
Travellers (thousand)
50
150
250
350
2013
2015
2017
2019
Time
UK
Travellers (thousand)
100
300
500
700
2013
2015
2017
2019
Time
US
Travellers (thousand)
100
300
500
2013
2015
2017
2019
Time

Figure 7 - Search queries on Google including the word “Italy” from German, Spanish, UK and USA users



6. Concluding Remarks

In the recent years, the Bank of Italy has carried out several experimental analyses in order to explore the possibility of integrating big data in the production of official statistics, in particular for compiling the “travel” item of the Balance of Payment.

The data that have been tested are appealing for their extraordinary timeliness and the amount of information offered, although they are very far away from being ready to use. Indeed, they need adjustments in order to define metrics that are coherent with the standards and the official definitions. The experiments often required adopting a trial and error approach to align these metrics to the prefixed standards, and making strong assumptions that could potentially affect the results.

According to our tests, mobile phone data seem to be the most suitable ones to be integrated with the frontier survey, as they are able to produce a broadly reliable estimate of the number of international travelers crossing the Italian borders, thus potentially replacing, at least partially, the counting procedures in the Bank of Italy frontier survey. The Bank of Italy is already moving in this direction.

The other big data sources analysed, electronic payments data and Google Trends data, showed more limitations and drawbacks.

Electronic payment data proved useful for achieving a preliminary estimate of total expenditures related to travel, as they are timelier than survey data. However, at this stage they can be used, at least for BoP, only for checking purposes. On the other hand, considering the informative potential of this source, we will continue exploring how to overcome the main problems by identifying the features that the data should have to be fully usable.

The Google Trends index proved to be useful for estimating the number of international travelers. But the sensitivity of the index to extraordinary circumstances like the Covid-19 pandemic needs to be further investigated before considering the integration of such index in the compilation process.

References

Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2007). Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia. *Information and communication technologies in tourism 2007*, 119-128.

Ahas, R., Aasa, A., Roose, A., Mark, Ü., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469-486.

Ahas R., Armoogum J., Esko S., Ilves M., Karus E., Madre J.L., Nurmi O., Potier F., Schmücker D., Sonntag U., and Tiru M. (2014) *Feasibility study on the use of mobile positioning data for tourism statistics*. Consolidated report Eurostat contract No 3051.2012.001-202.452
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

Artola, C., Pinto, F., and de Pedraza García, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower*, 36(1), 103-116.

Askitas, N., and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. Technical Report, SSRN 899

Bangwayo-Skeet, P. F., and Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, 46, 454-464.

Carrière-Swallow, Y., and Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. Journal of Forecasting, 32(4), 289-298.

Choi, H., and Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88, 2-9.

D'Amuri, F., and Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. International Journal of Forecasting, 33(4), 801-816.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, *111*(45), 15888-15893.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376-382.

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

Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F., and Hlavacs, H. (2015). The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE transactions on intelligent transportation systems*, *16*(5), 2551-2572.

Lokanathan, S., Kreindler, G. E., de Silva, N. N., Miyauchi, Y., Dhananjaya, D., and Samarajiva, R. (2016). The potential of mobile network big data as a tool in Colombo's transportation and urban planning. *Information Technologies and International Development*, *12*(2), pp-63.

McLaren, N., and Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, (2011), Q2.

Ricciato, F., Widhalm, P., Craglia, M., and Pantisano, F. (2015). Estimating population density distribution from network-based mobile phone data. *Publications Office of the European Union*.

Suhoy, T. (2009). Query indices and a 2008 downturn: Israeli data (No. 2009.06). Bank of Israel.

UN (2010), Manual on Statistics of International Trade in Services 2010*, Economic and Social affairs, Statistics division, Geneva, Luxembourg, Madrid, New York, Paris and Washington D.C.*
https://unstats.un.org/unsd/publication/seriesm/seriesm_86rev1e.pdf

UNWTO (2008), International Recommendations for Tourism Statistics 2008 (IRTS 2008), *New York, Economic and Social Affairs, United Nations.* https://unstats.un.org/unsd/tourism/methodology.asp

Vosen, S., and Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*(6), 565-578.

How can big data improve the quality of tourism statistics?

The Bank of Italy's experience in compiling the "travel" item of the Balance of Payments

Andrea Carboni, Costanza Catalano, Claudio Doria
Department of Economics, Statistics and Research – Bank of Italy

11th IFC Biennal conference
Basel, 26/08/2022

Big data for tourism statistics. The Italian case

Tourism statistics: number, expenditures and nights spent of

- Foreigners travelers visiting Italy (the reporting country)
- Italian travelers visiting abroad

BoP standards: expenditures by counterpart countries, business vs. personal trips, border/seasonal workers, international transports…

Sources: sample survey at border points (since 1996)

Drawbacks: costly, time-demanding, subjected to external factors (e.g. the covid-19 pandemic)

Big data: timelier, cheaper, less impacted by external shocks

Experiments on:

- Mobile Phone data
- Electronic payment data
- Internet search queries (Google Trends)

Mobile Phone data

May represent an alternative data source to count travelers crossing the border points.

➡ Only **complementary** source, no info on expenditures

- Arrival of a foreign traveler: signaled by the connection of foreign SIM cards to the cells controlled by an Italian network operator;
- Departure of an Italian traveler abroad: disappearance of the signal of an Italian SIM card near the border.

Nationality of the company issuing the SIM card = proxy for the traveler's country of origin

Collaboration with one of the major Italian Mobile Network Operator:

- Algorithm for estimating travelers inflows and outflows
- Constant cooperation necessary to achieve BoP standards (ex. minimum docking time due to handover effect)
- Tests on two main Italian border points (Fiumicino airport and Tarvisio highway)

Mobile Phone data

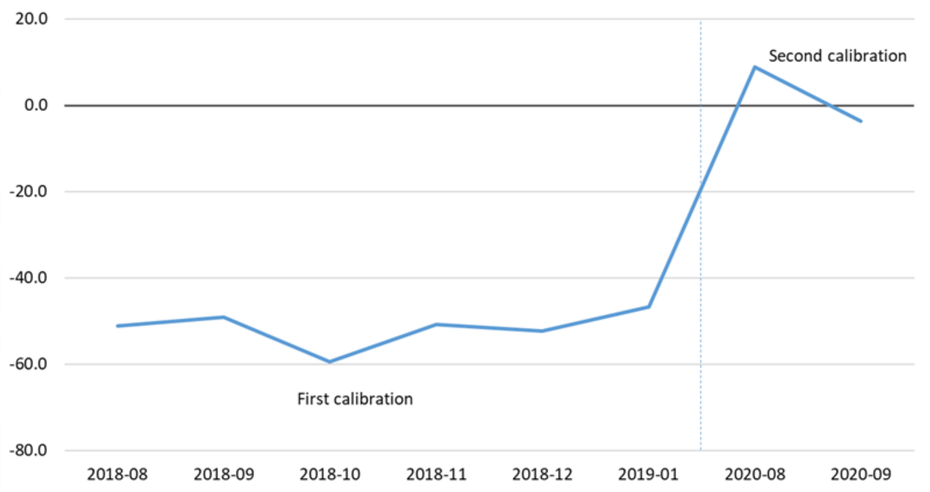
BANCA D'ITALIA
EUROSISTEMA

Fiumicino airport

	TOTAL			ITALIANS			FOREIGNERS		
	BI(1)	MPD(2)	MPD/BI%	BI(1)	MPD(2)	MPD/BI%	BI(1)	MPD(2)	MPD/BI%
Aug-18	1,717,076	1,802,051	4.9	640,288	621,419	-2.9	1,076,788	1,180,632	9.6
Sep-18	1,574,571	1,723,145	9.4	446,884	516,638	15.6	1,127,687	1,206,507	7.0
Oct-18	1,380,639	1,590,179	15.2	423,402	449,204	6.1	957,237	1,140,975	19.2
Nov-18	1,053,956	1,220,903	15.8	392,909	466,087	18.6	661,047	754,816	14.2
Dec-18	1,037,503	1,045,675	0.8	506,530	417,820	-17.5	530,973	627,855	18.2
Jan-19	831,120	1,113,629	34.0	344,529	457,947	32.9	486,591	655,682	34.8
Total	7,594,865	8,495,582	11.9	2,754,542	2,929,115	6.3	4,840,323	5,566,467	15.0

Tarvisio highway: huge discrepancies (order of 50%), needed a second test where the docking time was shortened.

Nowadays: The Bank of Italy has partially replaced the counting procedures in the frontier survey by MPD



Electronic payment data

Database: aggregated by date, nationality of bank emitting the card, type of purchase (10 categories), country of the POS/website

Foreign card & Italian POS/website ➡ Foreigners expenditure in Italy
Italian card & foreign POS/website ➡ Italian expenditure abroad

Main drawbacks in using payment data for BoP statistics:

- The nationality of the card is a <u>proxy of the traveler's residence</u>
- <u>Confidentiality issues</u> allow only aggregated data
- <u>No info on the reason</u> of the trip (business/personal)
- <u>Difficult to correctly classify the Digital Platform transactions:</u>
    - Payment of a stay in Paris by an Italian on Airbnb is recorded as from Italy to Ireland and not to France;
    - Payment of a stay in Rome by a French on Airbnb is recorded as from France to Ireland, thus not appearing in the database;
    - Payment of a stay in Rome by an Italian on Airbnb is recorded as from Italy to Ireland, despite it is a domestic trip.

Electronic payment data

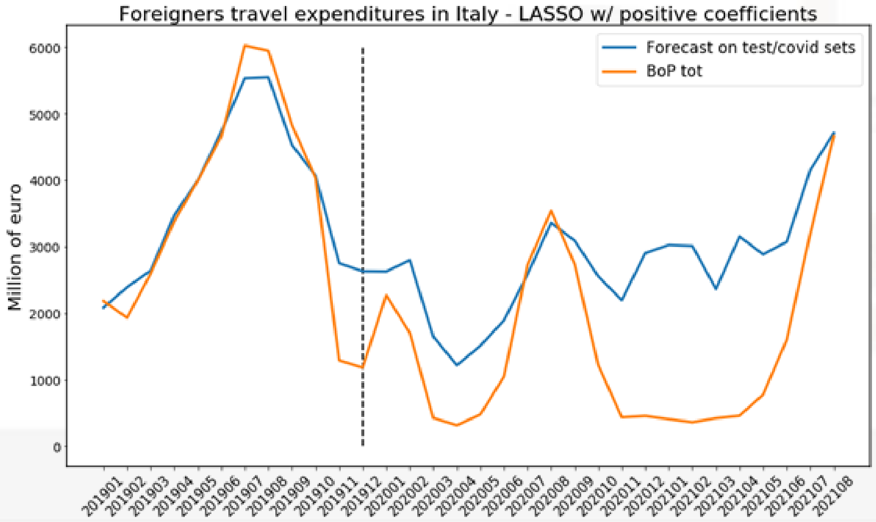
Market-share unknown ➡ Impossible the grossing-up of raw data

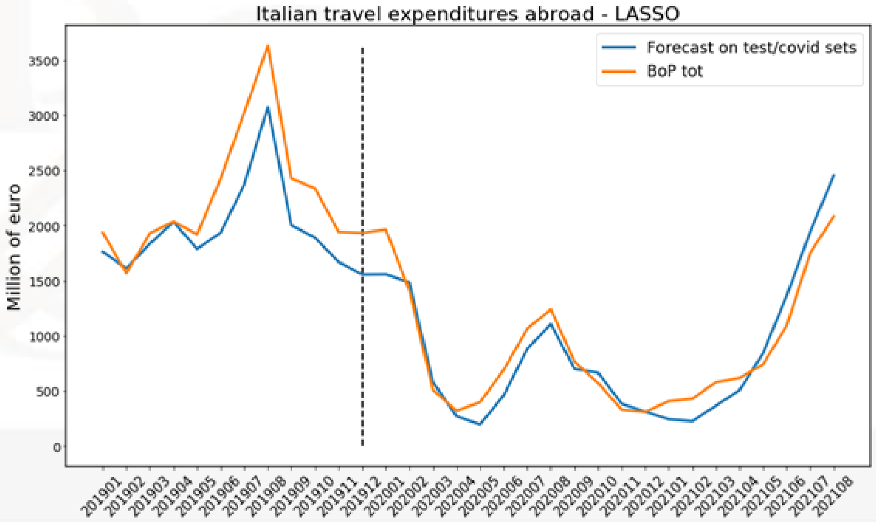
Tested some forecast models:

Ridge, Lasso, regression trees, boosted regression trees

<u>Training set:</u> years 2015-2017, <u>Validation set:</u> year 2018, <u>Test set:</u> year 2019, <u>Supplem. test set:</u> covid years.

Best performance in terms of MSE on the test set:





Internet search data: Google Trends

Google Trends index: reports the popularity of a given query in a given time period, country and category. It spans from 0 to 100.

Can the GT index be used to improve the provisional estimates on the number of travelers?

The model: seasonal AR(1)

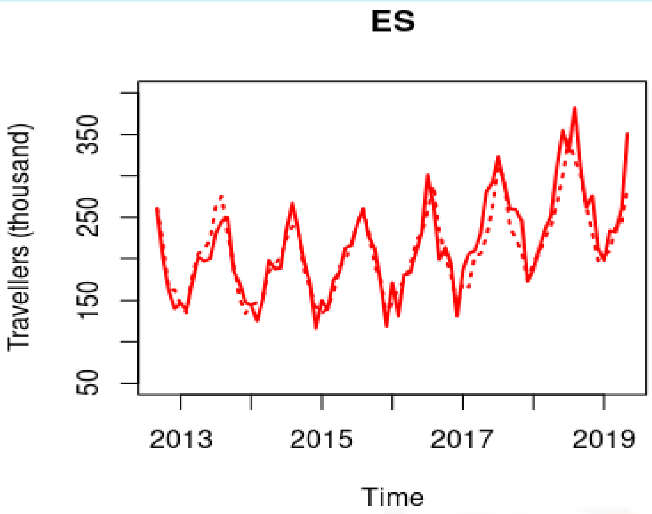
$$N_{c,t} = \phi_0 + \phi_1 N_{c,t-1} + \phi_{12} N_{c,t-12} + \beta GT_{c,t-l} + \varepsilon_{c,t}$$

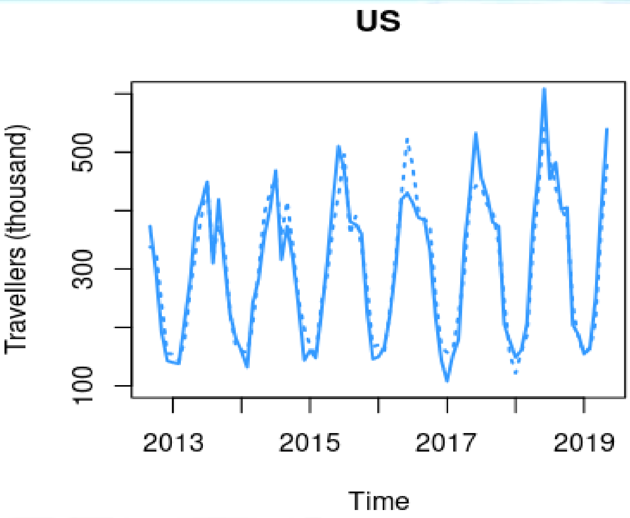
- queries including the word 'Italy', category= 'Travel'
- $N_{c,t}$ = number of travelers from country c in month t
- one-step ahead forecast with expanding windows approach
- lag for GT index chosen by minimizing the out-of-sample MSE
- five countries: France, Germany, Spain, UK and USA

Google Trends

Results: In all cases the GT index increased the performance of the predictive model, except for France where β was not statistically different from zero.

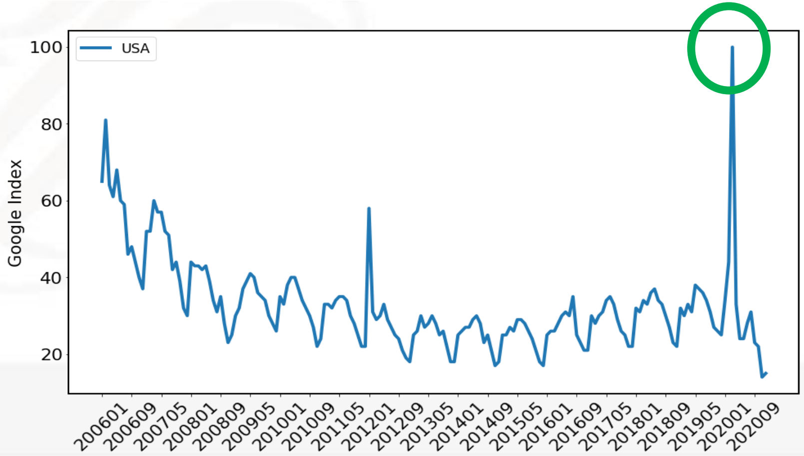
Internet search data: Google Trends





Limits: peak of search queries in March 2020 while Italy was blocking the tourist inflow.

In presence of <u>extraordinary events</u> the Google classification seems to be <u>less effective with high risk of outliers.</u>



Take-away

All the data sources needed adjustments in order to define metrics that are coherent with the BoP standards.

Mobile phone data:

- the most suitable ones to be integrated with the frontier survey in the estimate of the number of international travelers
- Bank of Italy uses MPD for tourism statistics since the end of 2020

Electronic payment data:

- useful to achieve a preliminary timelier estimation of the total expenditure of the "travel" item
- For now, the mentioned relevant issues make it usable only for checking purposes

Google trends:

- useful as explanatory variable for estimating the number of international travelers
- possible noise of the index could be misleading. Use of other/more words as search queries could generate more accurate results.

Thank you for your attention!

...Questions?