

Usage of big data to improve the quality of tourism statistics

The Bank of Italy's experience

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

- Tourism statistics (at Bank of Italy) 
- Mobile Phone data 
- Electronic payment data 
- Internet search queries (Google Trends) 
- Conclusions

Focus: pros and cons on using these data sources for tourism statistics

Tourism statistics (at Bank of Italy)



Balance of Payments (BoP) perspective:

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

No domestic tourism

Interested in:

- Number of travelers
- Expenditure (accommodation, transports, restaurants, shopping,...)
- Nights spent

But also (BoP standards): type of the trip (business vs. personal), border/seasonal workers, trip for medical care, students studying abroad, residence issues (foreigners permanently living in Italy/Italians permanently living abroad), international transports...





sample survey at selected border points (since 1996)

1. **Counting operations** for estimating the number of travelers, broken down by their residence
2. **Interviews** face-to-face for assessing the tourist expenditures and other variables



Operations conducted each month to meet ECB data requirements
Annually ~ 1.1 mln of counting operations and ~ 100k of interviews



Source: 

sample survey at selected border points (since 1996)

Drawbacks:

- Expensive
- Time-demanding (slow collection of data)
- Subjected to sudden interruption due to external factors (e.g. the covid-19 pandemic)

Big data:

- Timelier
- Cheaper
- Less impacted by external shocks



Mobile Phone Data



May represent an alternative data source to count the number of travelers

[only complementary data source if interested also in tourist 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 residence



May represent an alternative data source to count the number of travelers

[only complementary data source if interested also in tourist expenditures]

Arrival of a foreign traveler:

SIM followed along all its permanence in Italy until it disappears

➡ info on number of night and cities/provinces visited

Departure of an Italian traveler abroad:

Wait until it appears again in Italy. “*Welcoming SMS*” as proxy of the country visited abroad

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



Data provided by one of the major Italian Mobile Network Operator (MNO).

The MNO:

- Filters and process the data (~ 20 billion of records per day)
- Develops an algorithm for estimating the travelers inflows and outflows:
 - Per each border point
 - Per each nationality/visited country
 - Takes into account its *share* among the other MNOs and does the grossing-up (expand to the reference population)



Main challenges to face:

- Identify and filter the Non-humans (M2M) SIM cards
- Presence of dual SIMs
- Excluding foreign SIM cards that appear too often in Italy (foreigners permanently living in Italy)
- Handover effect between phone cells near the border
 - Careful choice of the minimum docking time of a SIM card in the cells in order to be considered a traveller



Main challenges to face: [solutions]

- Identify and filter the Non-humans (M2M) SIM cards [updated list of M2M SIM identification number by the MNO]
 - Presence of dual SIMs [MNO estimates]
 - Excluding foreign SIM cards that appear too often in Italy (foreigners permanently living in Italy)
 - Handover effect between phone cells near the border
 - Careful choice of the minimum docking time of a SIM card in the cells in order to be considered a traveller
- Careful calibration of the algorithm
- Trial and error approach through comparison with:
- Places/events where the number of people is known (concerts, fairs, shows,..)
 - Other statistics (Bank of Italy, official airport data,...)

Continuous cooperation between MNO and Bank of Italy
necessary to achieve the desired standards

Preliminary tests on two main Italian border points.

Fiumicino airport (Rome)

BI= Bank of Italy statistics

MPD= mobile phone data statistics

ADR= official airport data statistics

	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

	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

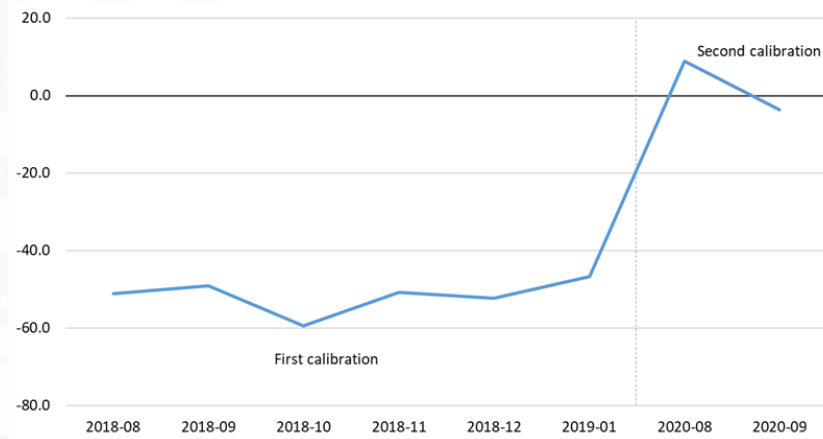
Tarvisio highway (main border with Austria)

Huge discrepancies
(order of 50%)



A second test was needed where
the docking time had been
shortened

	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



The Bank of Italy has replaced all the counting operations (but few for checking purposes) with the statistics coming from MPD since the end of 2020

Take away:



- MPD are suitable to be integrated with the frontier survey in the estimate of the number of international travelers
- Careful when dealing with M2M SIM cards, handover effects, dual sims...
- A constant interaction with the MNO is necessary to define metrics that are coherent with tourism and obtain reliable statistics

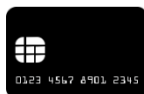
The Bank of Italy has replaced all the counting operations (but few for checking purposes) with the statistics coming from MPD since the end of 2020

Take away:



Questions?

- MPD are suitable to be integrated with the frontier survey in the estimate of the number of international travelers
- Careful when dealing with M2M SIM cards, handover effects, dual sims...
- A constant interaction with the MNO is necessary to define metrics that are coherent with tourism and obtain reliable statistics



Electronic payment data

Copyright 2007 John Crowther



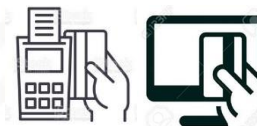
"Hey everyone, check it out, I got one over here actually paying cash money."

Database provided by one of the main Italian paytech operator



Variables:

- Date (day/month/year)
- Nationality of the bank issuing the payment card
- Type of purchase (Merchant Category Code - 10 categories)
- Country where the **POS** or **website** is located
- Amount



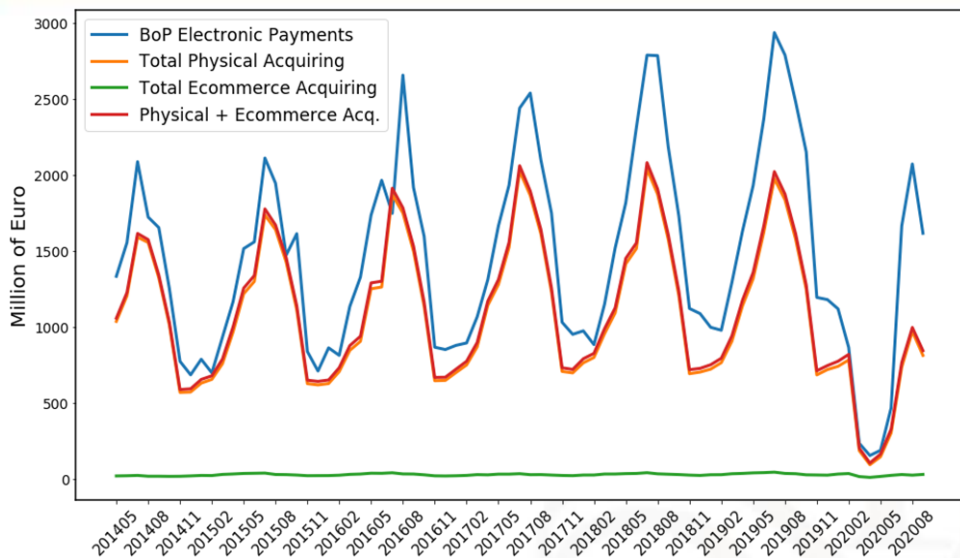
→ Aggregated by all the variables above

Can we use electronic payment data to estimate tourist expenditure?

Can we build a model to provide timelier provisional estimate?

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

Italian card & foreign POS/website ➡ Italian expenditure abroad



Foreigner travelers' expenditure in Italy

Acquiring= electronic transactions made by foreign payment cards in Italy

BoP electronic payments= official bank of Italy statistics (expenses declared to be paid by payment cards)

Correlation by categories

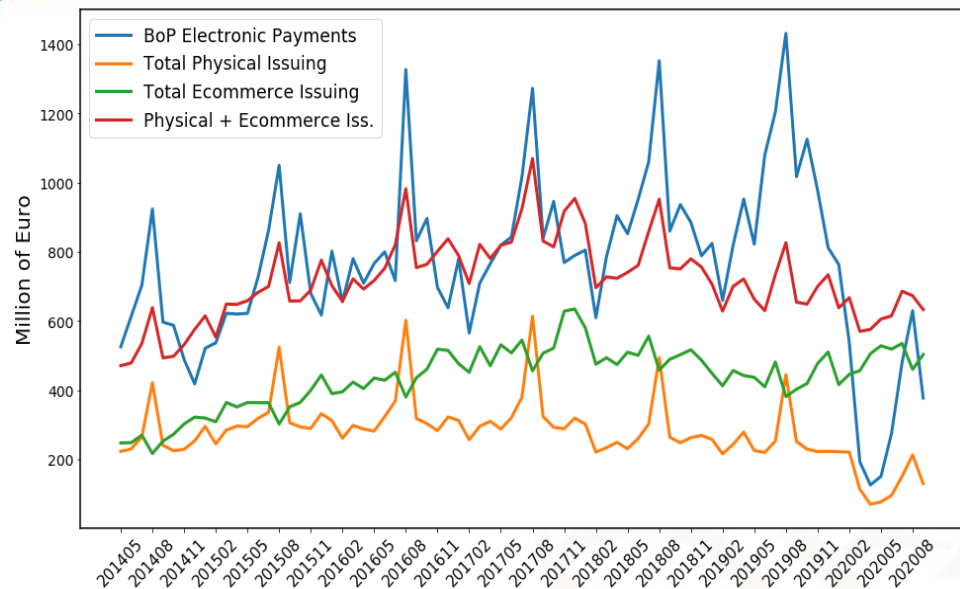
Very high correlation

High correlation (for certain categories)

Transactions on POS (physical) →

E-commerce transactions
on different lags →

	CLOTHING	GROCERIES	HOTELS AND RESTAURANTS	HOME	CASH ADVANCE	WORK	RETAIL	SERVICES	MOBILE AND WEB	TRAVEL AND TRANSPORTS	TOTAL
Transactions on POS (physical)	0.83	0.94	0.98	0.9	0.96	0.88	0.92	0.18	-0.05	0.94	0.97
E-commerce transactions on different lags											
-4	0.28	-0.02	0.33	0.4		0.17	0.24	0.48	0.22	0.04	0.02
-3	0.23	-0.2	0.56	0.51		0.45	0.2	0.57	0.25	0.35	0.46
-2	0.16	-0.29	0.68	0.45		0.58	0.08	0.45	0.26	0.53	0.75
-1	0.25	-0.26	0.61	0.43		0.67	0.07	0.39	0.22	0.58	0.89
0	0.33	-0.23	0.4	0.31		0.62	0.06	0.2	0.19	0.44	0.75



high correlation only on
total and few categories

No significant correlation
but hotels and restaurants



Transactions on POS (physical) →

E-commerce transactions
on different lags →



-4
-3
-2
-1
0

Correlation by categories										
CLOTHING	GROCERIES	HOTELS AND RESTAURANTS	HOME	CASH ADVANCE	WORK	RETAIL	SERVICES	MOBILE AND WEB	TRAVEL AND TRANSPORTS	TOTAL
0.16	0.49	0.78	0.14	0.28	0.7	0.38	0.65	0.19	0.66	0.65
-0.08	-0.02	0.41	-0.09		0.07	0.18	0.22	-0.24	0.33	0.05
-0.16	-0.13	0.54	-0.16		0	0.12	0.17	-0.3	0.36	0.08
-0.28	-0.27	0.61	-0.24		-0.19	0.02	-0.04	-0.47	0.25	0.04
-0.24	-0.28	0.72	-0.24		-0.18	0.21	-0.04	-0.51	0.14	0.14
-0.27	-0.35	0.7	-0.28		-0.3	0.12	-0.05	-0.57	-0.1	-0.02

General problems in using payment data for tourism statistics:

- The nationality of the card is a proxy of the traveler's residence (issues with Revolut, N26..)
- Confidentiality issues allow only aggregated data:
 - you do not see the single payment card transaction 
 - Difficult to tell apart transactions made by foreigners living permanently in Italy using foreigners payment cards/ Italians living permanently abroad using Italian cards
- Share of the paytech company unknown:
 - we cannot expand the data to the reference population
- Lack of info on cash and wire payments, and on the type of the trip (business/holiday) 
- Identification of the transactions that are related to tourism from the ones that are not, based on the Merchant categories (especially in the e-commerce). Identification of good for resale

General problems in using payment data for tourism statistics:

- The nationality of the card is a proxy of the traveler's residence (issues with Revolut, N26..)
- Confidentiality issues allow only aggregated data:
 - you do not see the single payment card transaction 
 - Difficult to tell apart transactions made by foreigners living permanently in Italy using foreigners payment cards/ Italians living permanently abroad using Italian cards
- Share of the paytech company unknown: [outsourcing the estimates to the Paytech company as done with MPD?]
 - we cannot expand the data to the reference population
- Lack of info on cash and wire payments, and on the type of the trip (business/holiday) 
- Identification of the transactions that are related to tourism from the ones that are not, based on the Merchant categories (especially in the e-commerce). Identification of good for resale

Transactions through Digital International Platforms



Transactions through Digital International Platforms



1. Transactions not covered/misallocation by counterpart country

In tourism statistics: a French traveler pays an accommodation in Italy (Italian credits)

In electronic payment data:

- Transaction from France to The Netherlands → not covered
- Transaction from The Netherlands to Italy (IF made by payment cards) → misallocation

Underestimation of Italian credits

Transactions through Digital International Platforms



2. Misallocation by counterpart country

In tourism statistics: an Italian traveler pays an accommodation in France (Italian debits)

In electronic payment data:

- Transaction from Italy to The Netherlands → misallocation

Transactions through Digital International Platforms



3. False positive: domestic trip seen as international

In tourism statistics: domestic tourism

In electronic payment data:

- Transaction from Italy to The Netherlands → misallocation
- Transaction from The Netherlands to Italy (IF made by payment cards) → misallocation

Overestimation of Italian debits
Overestimation of Italian credits

Transactions through Digital International Platforms



4. Time of recording issue

In tourism statistics: trip made in August

In electronic payment data:

- Transaction from Italy to The Netherlands in April → seen as trip made in April

Time of recording issue common to ALL on-line transaction

Transactions through Travel Agencies/Tour operators



Only one merchant category code to describe transactions towards travel agencies/tour operators

- Impossible to identify the purchased services (accommodation/transports/package tour...)
- Same time of recording issues/misallocation issues/transitions not covered as Digital platforms, depending where the travel agency is located

Goal: forecasting credits and debits of the total travel expenditure

Data: monthly data from January 2015 to August 2020

Tested some forecast models:

- Ridge
- Lasso
- Lasso with positive coefficients
- Regression trees
- Boosted regression trees

Training set: years 2015-2017

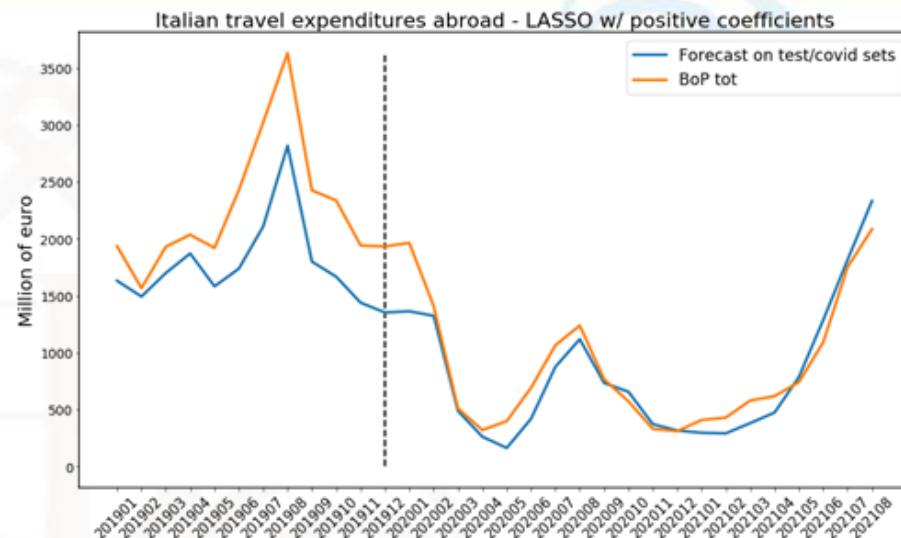
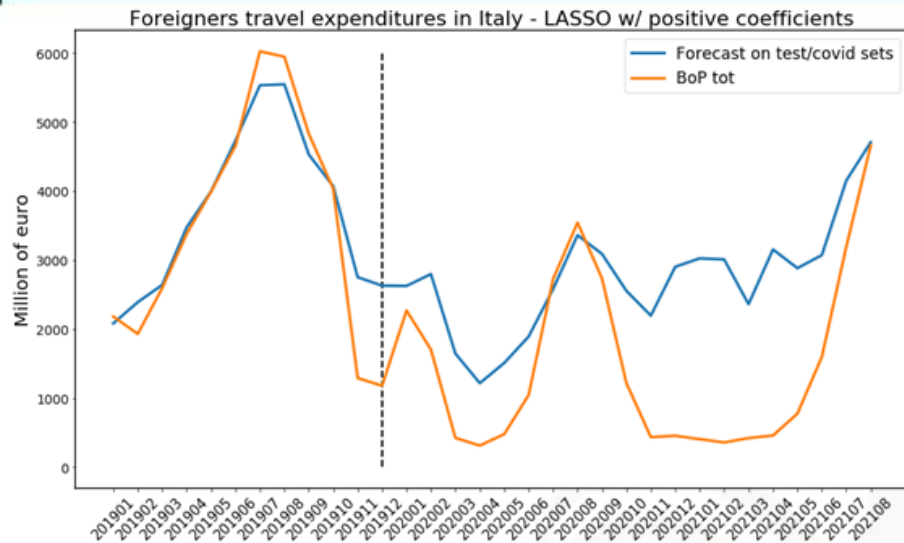
Validation set: year 2018 (for choosing the discount parameter of the Lasso/Ridge models)

Test set: year 2019

Supplementary test set: covid year 2020

	Foreigners in Italy			Italians abroad		
	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

Electronic payment data: what we can do with them



physical	Foreigners	Italians	e-commerce	Foreigners	Italians
HOTELS AND RESTAURANTS	X		HOTELS AND RESTAURANTS (lag 0)		X
GROCERIES	X		HOTELS AND RESTAURANTS (lag -2)		X
CASH ADVANCE	X	X	HOTELS AND RESTAURANTS (lag -3)		X
CLOTHING		X	HOTELS AND RESTAURANTS (lag -4)	X	
TRAVELS AND TRANSPORTS	X	X	HOME (lag -1)	X	
			WORK (lag 0)	X	
			WORK (lag -3)	X	
			SERVICES (lag -1)	X	
			TRAVELS AND TRANSPORTS (lag 0)	X	X
			TRAVELS AND TRANSPORTS (lag -1)	X	
			TRAVELS AND TRANSPORTS (lag -2)		X
			TRAVELS AND TRANSPORTS (lag -3)		X
			TRAVELS AND TRANSPORTS (lag -4)		X

Features selection

Take away:



- Important issues inherent of the database:
 - misallocation by counterpart country and time of recording
 - transactions not covered
 - DIPs and travel agencies
 - misclassification of domestic trips
- Some problem might be solved by more granular data (more payment categories, identification of the single payment card, description of the transaction...)
- Proved useful to be a valid input of preliminary tourist expenditure estimates

Take away:



- Important issues inherent of the database:
 - misallocation by counterpart country and time of recording
 - transactions not covered
 - DIPs and travel agencies
 - misclassification of domestic trips
- Some problem might be solved by more granular data (more payment categories, identification of the single payment card, description of the transaction...)
- Proved useful to be a valid input of preliminary tourist expenditure estimates

Questions?

Google Trends 

Website that analyses the popularity of search queries on Google



Input:

- Keywords that must have been included in the query
- Reference period
- Geographic area where conducted the web search
- Category of the search

Example:

- 'Italy'
- 2006-2019
- Germany, France, Spain , UK, USA
- Travel

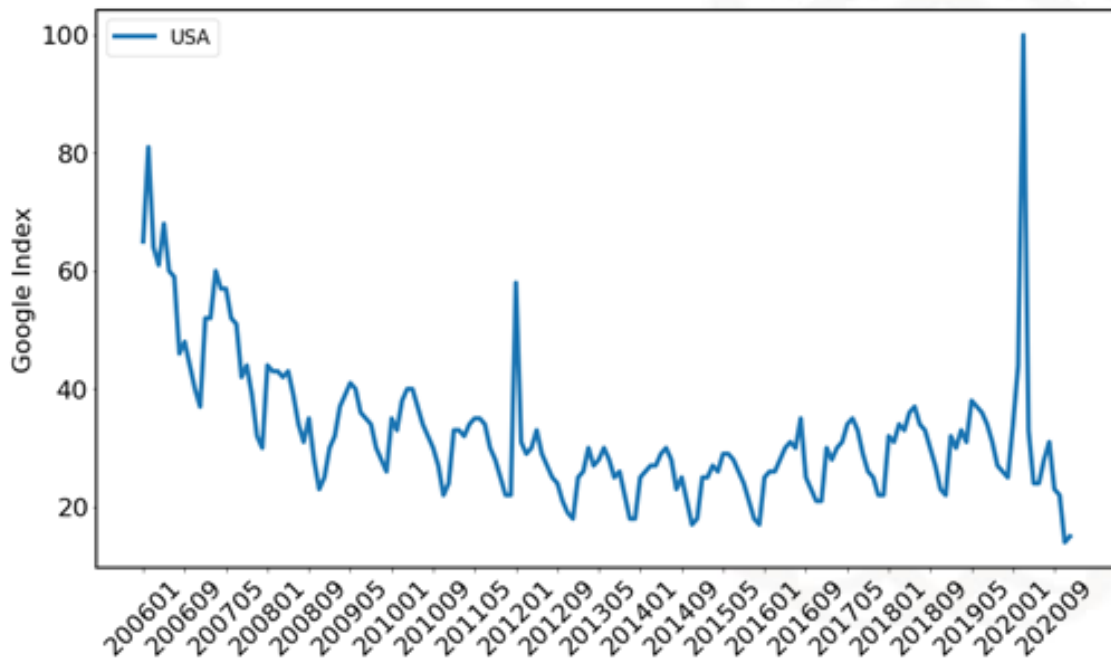
Output: for each day/week/month in the time frame it selects the queries containing the chosen keywords in the selected geographic areas and category, producing the GT index:

GT=100	when reached the maximum volume of queries over the selected time frame
GT=x	when reached x% of the volume of queries of GT=100
GT=0	no queries

...it's a relative index!

<https://trends.google.it/trends/?geo=IT>

Website that analyses the popularity of search queries on Google



Example:

- 'Italy'
- 2006-2019
- USA
- Travel

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

The model: seasonal AR(1)

$N_{c,t}$ = number of travelers from country c in month t

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

Known from the survey

Lag chosen as the one minimizing the out-of- sample forecasting performance in terms of Mean Square Error

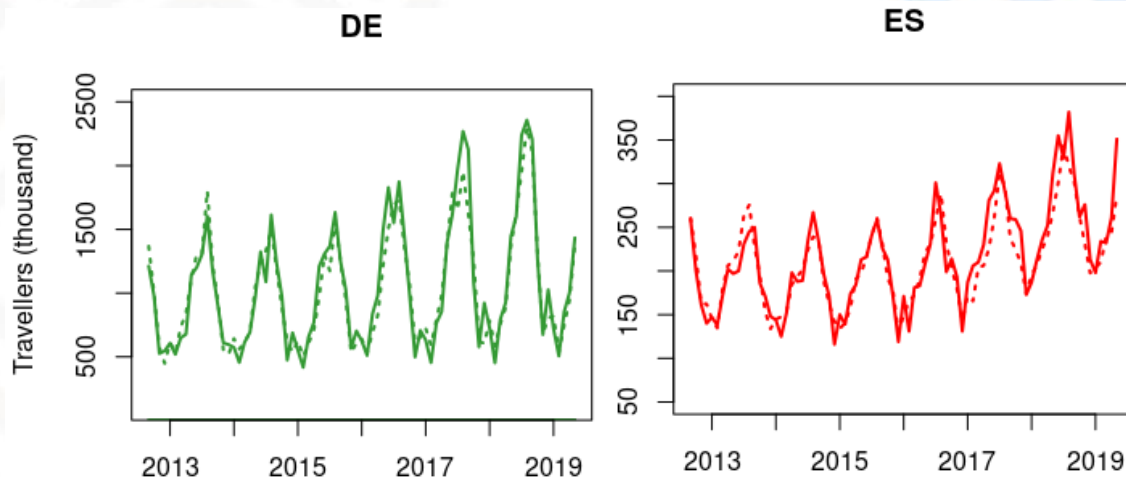
GT parameters:

- Keyword: 'Italy'
- Time: 2006-2019
- Location: Germany, France, Spain, UK, USA
- Category: 'Travel'

- one-step ahead forecast with expanding windows approach

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.

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



Lag=0

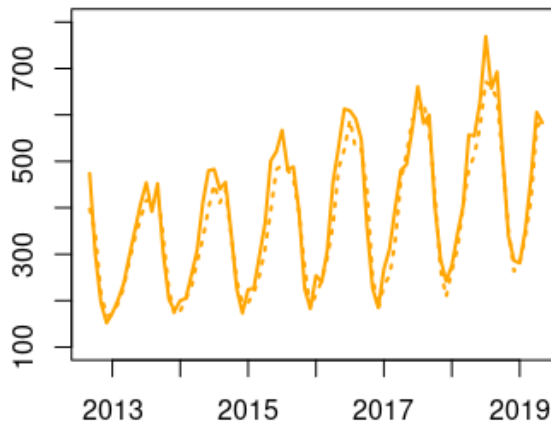
Solid line: observed data; Dashed line: predicted data

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.

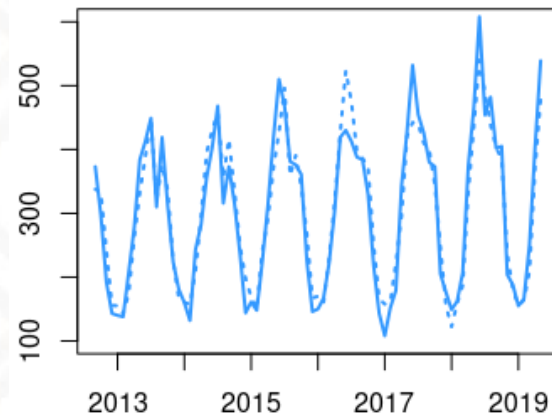
	(1) DE	(2) ES	(3) UK	(4) US
N_{t-1}	0.20*** (0.04)	0.34*** (0.05)	0.33*** (0.05)	0.13*** (0.04)
N_{t-12}	0.73*** (0.05)	0.46*** (0.05)	0.76*** (0.05)	0.89*** (0.04)
GT_{t-1}	6.01*** (1.17)	2.11*** (0.29)	-1.50*** (0.44)	-0.91*** (0.24)
Const	-182.49*** (43.34)	-22.78* (12.07)	29.87* (15.52)	38.84***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

Travellers (thousand)

UK



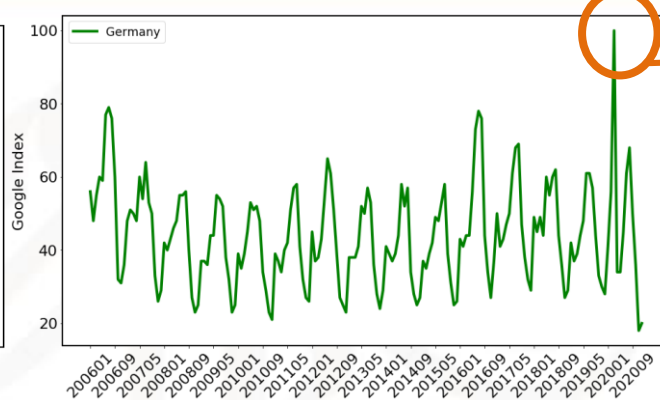
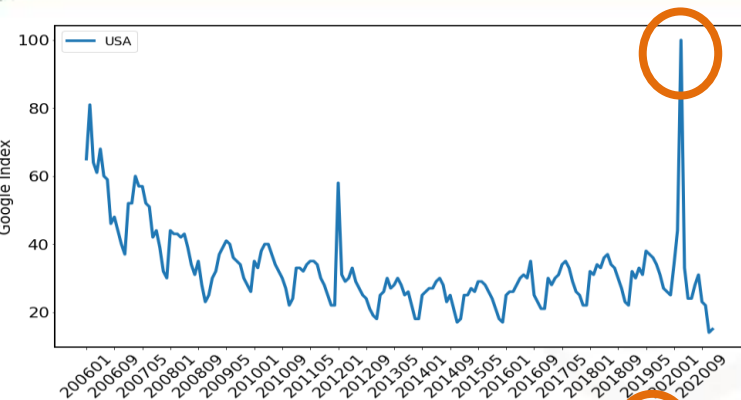
US



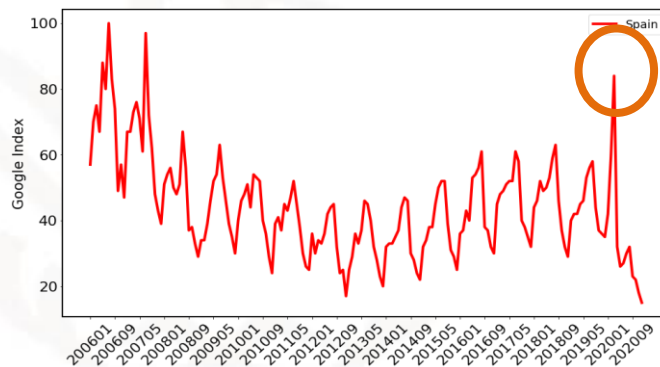
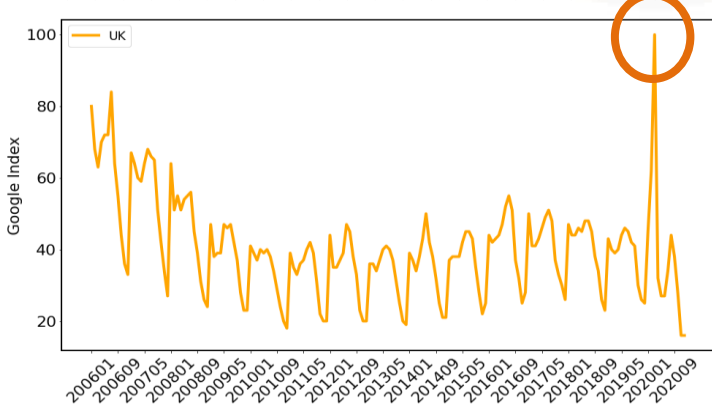
UK lag=4 US lag=6

Solid line: observed data; Dashed line: predicted data

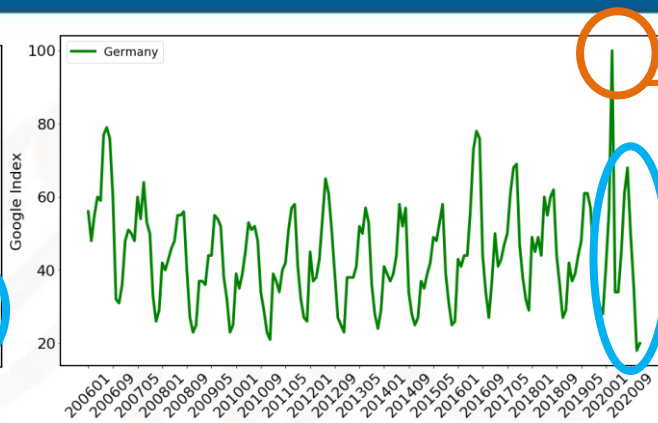
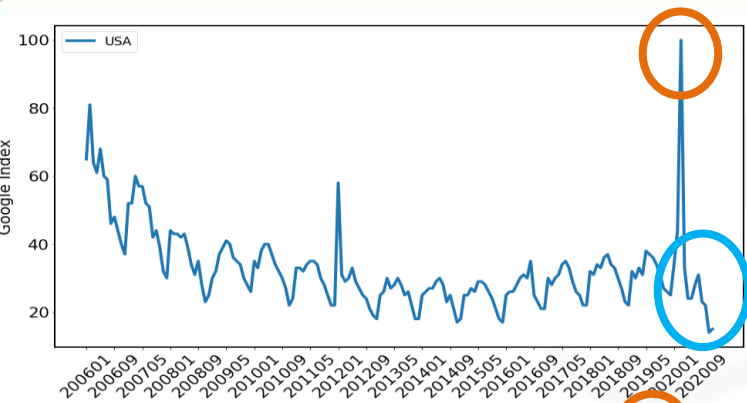
Limits of Google Trends



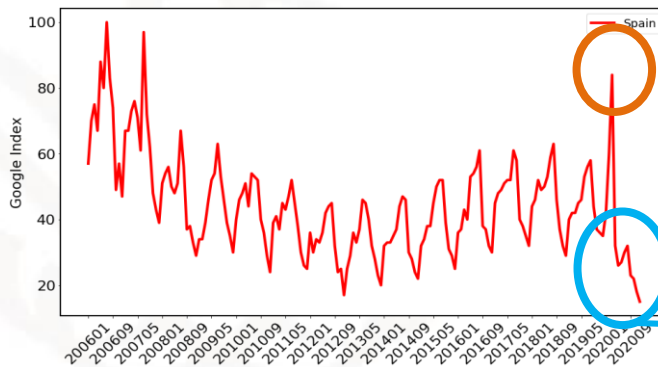
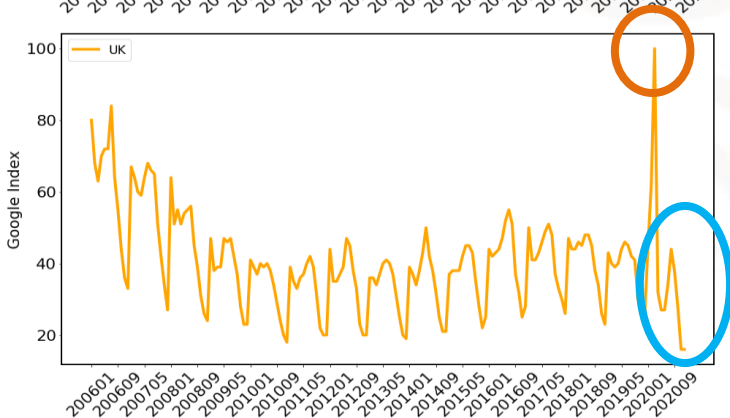
Peak of search queries
in March 2020 while
Italy was blocking the
tourist inflow



Limits of Google Trends



Peak of search queries in March 2020 while Italy was blocking the tourist inflow



Drop of smaller magnitude that the 'real' tourist presence

Take away:



- In presence of extraordinary events (like the covid-19 pandemic) the Google classification seems to be less effective with higher risk of outliers
- Useful as explanatory variable for estimating the number of international travelers
- Promising source of information but the possible noise of the index could be misleading. Use of other/more words as search queries (maybe referring to specific Italian tourist destinations) could generate more accurate results

Take away:



- In presence of extraordinary events (like the covid-19 pandemic) the Google classification seems to be less effective with higher risk of outliers
- Useful as explanatory variable for estimating the number of international travelers
- Promising source of information but the possible noise of the index could be misleading. Use of other/more words as search queries (maybe referring to specific Italian tourist destinations) could generate more accurate results

Questions?

Thank you



Duane Hanson *Tourists II* (1988)
Photo: Saatchi Gallery

- 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.
- Askitaş, 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.

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.
- UNWTO (2008), [International Recommendations for Tourism Statistics 2008 \(IRTS 2008\)](https://unstats.un.org/unsd/tourism/methodology.asp), 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.