BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

A tool to nowcast tourist overnight stays with payment data and complementary indicators

by Marta Crispino and Vincenzo Mariani

February 2023

Number 746

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

A tool to nowcast tourist overnight stays with payment data and complementary indicators

by Marta Crispino and Vincenzo Mariani

Number 746 – February 2023

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of the Bank of Italy

A TOOL TO NOWCAST TOURIST OVERNIGHT STAYS WITH PAYMENT DATA AND COMLPLEMENTARY INDICATORS

by Marta Crispino† and Vincenzo Mariani‡

Abstract

This paper proposes a strategy for nowcasting tourist overnight stays in Italy by exploiting payment card data and Google Search indices. The strategy is applied to national and regional overnight stays at a time of a significant and unanticipated shock to tourism flows and payment habits (the COVID-19 pandemic). Our results show that indicators based on payment data are very informative for predicting tourist volumes, both at the national and at the regional level. Instead, the predictive power of Google Search data is more limited.

JEL Classification: L83, C53, C55, F47.
Keywords: tourism, time series, payment card data, Google trends, nowcasting.
DOI: 10.32057/0.QEF.2022.0746

Contents

† Bank of Italy, Via Nazionale, 91 - 00184 Rome, Italy, marta.crispino@bancaditalia.it
‡ Bank of Italy, Corso Cavour, 5 - 70121 Bari, Italy, vincenzo.mariani@bancaditalia.it

1 Introduction[1]

According to available estimates, in Italy tourism accounts directly for 5 per cent of the GDP and for 6 per cent of employment, well above the average in the OECD countries (Petrella et al., 2019). Taking into account the indirect contribution of tourism, the figures are remarkably higher (respectively around 13 and 15 per cent).[2] In 2018, before the spread of the COVID-19 pandemic, according to the World Tourism Organization, Italy was the fifth major destination of international tourist flows, with around 62 million visits (UNWTO, 2021); moreover, in the same year, roughly 65 million residents visited the country (Istat, 2021). The number of arrivals from abroad has increased by roughly 4 per cent each year on average during the previous decade, and tourism has been one of the best performing sectors in the Italian economy. Given the relevance and growth of this sector, the accurate and timely estimate of tourist flows is becoming increasingly important in tracking economic activity. Moreover, their quantification may also be useful for compiling other statistics, as for instance the travel item in the balance of payments (Carboni et al., 2022).

Motivated by these facts, this paper introduces a new strategy to nowcast monthly tourism. Specifically, our target variable is the total number of overnight stays, which is a statistics published on a monthly basis and with delay by the Italian national office of statistics (Istat) and by Eurostat.

We present two exercises. First, we consider the estimation of overnight stays at the national level. In this case, the goal is to nowcast the official statistics published by Istat with a delay of few months. This exercise is particularly useful at the end of the summer period, as it gives real-time information during the peak tourist season, when the contribution of tourism to GDP and employment is particularly large. In particular, official statistics about overnight stays of the summer months are available in the following January. However, our choice to focus the analysis on the summer months does not prevent to use the models proposed at different periods of the year.

In the second exercise, we focus on regional tourist volumes. The regional dimension is important because most policies on tourism are deployed by regional administrations.[3] Such

[1]The views expressed in this paper are those of the authors and do not involve the responsibility of the Bank of Italy and/or the Eurosystem. We thank Matteo Alpino, Valentina Aprigliano, Laura Bartiloro, Andrea Carboni, Costanza Catalano, Andrea Doria, Simone Emiliozzi, Silvia Fabiani, Sara Lamboglia, Michele Loberto, Juri Marcucci, Alessandro Moro and Alfonso Rosolia for fruitful discussions and suggestions. We would also like to thank Marco Langiulli and Luca Bastianelli for providing useful information about the payment card data.

[2]Other estimates produced by Istat (2020) indicate that the direct and indirect contribution to the value added are around 6 and 15 per cent, respectively.

[3]Examples include publicly financed advertisement campaigns or local transport organizations. More in

policies, jointly with other structural and temporary factors, generate geographical heterogeneity in tourism flows. Official regional statistics are published by Istat with much more delay than national data: generally, monthly data on the previous year are available in June.[4] In some cases, provisional regional data can be obtained earlier from the regional tourist offices, which are compelled by law to collect and deliver them to Istat. However, they often do not make them available before the official publication by Istat itself. Moreover, the quality of provisional data may be quite low because the statistical procedures of harmonization and quality are performed by Istat. We show how and to what extent overnight stays in a region can be nowcast using national data on overnight stays and additional predictors. In particular, at the beginning of October official regional statistics are available up to December of the previous year, and estimation can be conducted, as we show, for the nine missing months (from January to September included).

The two exercises differ mainly for two reasons. First, the estimation window for the regional exercise is longer, as regional data are published with more delay (to this extent the exercise is more demanding). Second, when estimating regional overnight stays, national stays are available for most of the estimation period. Therefore, they can be used as a predictor in the estimation procedure (to this extent the exercise is less demanding). In both cases, we have no ambition to project the series to the future.

The regional exercise is carried out on data from Puglia and Veneto, two very different regions in terms of tourist flows, but it is immediately generalizable to all Italian regions. Veneto is historically the Italian region with the largest tourist volumes and is located in the North of the country. Puglia is in the South, and compared to Veneto features lower tourist flows, mostly domestic, although it experienced a sharp increase in tourism relevance in the decade before the COVID-19 pandemic. The number of overnight stays in this region shows higher seasonality than in Veneto, mainly because seaside tourism is prevalent. On the contrary, in Veneto, flows are relatively more uniformly distributed across months, since cultural tourism is more widespread within the solar year, and winter tourism is important for this region.

In order to account for seasonality, we consider as a benchmark a simple econometric model with seasonal adjustments (SARIMA). We evaluate its performance in comparison to extended versions of this model, augmented with exogenous variables (SARIMAX or dynamic

general, following the constitutional reform of 2001, the regional administrations have seen an extension of their powers on many tourism issues.

[4]This means that until June of a given year, the regional data are only available until December of two years before.

regression model). In particular, we include information from payment card data and web searches, considering, for the latter, the existence of a possible time lag with respect to the target variable. The SARIMA model assumes that future values of the target variable linearly depend on its past values, as well as on the values of past stochastic shocks. The SARIMAX model, instead, improves the nowcasting performance by exploiting additional information from other external observable variables which are timely available. Such variables may be able to capture the occurrence of unexpected events which affect tourist flows, but that are not captured by the lagged variables included in the benchmark model.

By its nature, predicting tourist flows is a challenging exercise, as idiosyncratic shocks in the host and tourists' origin countries may affect travel decisions significantly (de Kort, 2017). These "disturbances" can be of very heterogeneous nature: for instance, they may be due to the environmental, economic or political situation or they may be related to social or natural factors that are difficult to predict, making a given area touristically more (or less) attractive for a short or a long period. An extreme case is the COVID-19 pandemic, which, jointly with all restrictions to mobility, has been an extremely negative shock to tourist flows (Demma, 2021). Other examples of shocks, such as big cultural and sports events (with a positive effect) or natural disasters (with a negative effect), may display similar intensity, but may be eventually more geographically localized to specific regions.

Our analysis reveals that the models augmented with the proposed complementary indicators very often outperform the benchmark model. Better performances of the augmented models are due, in particular, to the introduction of payment card data as explanatory variables: this result holds even in 2020, despite the huge change in the consumption and payment habits induced by the pandemic (Ardizzi et al., 2021, Della Corte et al., 2021) and by the measures introduced to limit the spread of the virus. The predictive power of indices based on Google Trend (GT) series is more limited, as already shown by Antolini and Grassini (2019), but their introduction in our regional models is useful because they have in some cases a positive impact on the predictive performance.

The rest of the paper is structured as follows. After Section 2, which proposes a short literature review, Section 3 describes the data; Section 4 presents the estimation setting and the timing of our target variable. Section 5 discusses briefly the models we use for the prediction task and the empirical strategies adopted in the two exercises. Section 6 illustrates the results of the estimations and Section 7 concludes.

2 Literature review

The literature aiming at forecasting tourism has mainly focused on three aggregates: tourist expenditure, international tourist arrivals and length of stay (de Kort, 2017, Carboni et al., 2022). Empirical analyses tend to agree that time series models perform better than explanatory models with pull and push factors[5] as predictive variables (Antolini and Grassini, 2019). Most works use as a benchmark a simple autoregressive model with seasonal effects. Performances of competing model specifications with additional information are then tested against the benchmark.

Payment card data have already been shown to be informative to nowcast macroeconomic variables, for instance household consumption (see e.g. Aladangady et al., 2021, Verbaan et al., 2017, Aastveit et al., 2020) and gross domestic product (Aprigliano et al., 2019, Galbraith and Tkacz, 2018). However, to the best of our knowledge the use of payment card data to forecast tourist flows is a novelty of this paper.

On the contrary, the use of indices based on web searches as prediction variables for tourism flows is not new, and their forecasting capability, as in the case of payments by cards, will likely increase as time goes by, considering the growing share of goods and services that are bought through the internet. The use of web searches (in particular, GT series) for predicting tourist activity has been first proposed in the seminal paper by Varian and Choi (2009).[6] Using GT series Artola and Martínez-Galán (2012) obtain a one-month leading indicator of the flow of British tourists to Spain, but the improvement in forecasting provided by their short-term models is limited. Using a dynamic factor approach and a real-time database reproducing the exact information available at each particular point in time, Camacho and Pacce (2018) find that models including Google's query volume indices outperform models that exclude them. Other studies applying time series techniques for the prediction of tourist flows are Park et al. (2017), which focus on the Japanese tourism in South Korea, finding that Google-augmented models perform significantly better than standard time series models in terms of short-term forecasting accuracy. Havranek and Zeynalov (2019) and Bangwayo-Skeete and Skeete (2015) study the performances of mixed data sampling (MIDAS) models in the context of tourist flows. Yang et al. (2015) compare the

[5] According to the psychological literature, people travel because they are "pushed" into making travel decisions by internal, psychological forces, and "pulled" by the external forces of the destination attributes (Yoon and Uysal, 2005).

[6] GT series have been used in many other fields. Examples are: prediction of unemployment (D'Amuri and Marcucci, 2017, Askitas et al., 2009, for USA and Germany respectively), trends in the housing market (Wu and Brynjolfsson, 2015, Webb, 2009); consumer confidence (Della Penna and Huang, 2009).

performances of GT with the ones from Baidu, a search engine which has larger market share than Google Search in China.

As far as we know, the only application on Italian data is by Antolini and Grassini (2019), who use GT series to predict international tourist arrivals in the country finding no evidence of a significant contribution in terms of forecasting accuracy. According to the authors, a possible explanation for this finding is the nature of the data: the time series of foreign arrivals in the country shows a very regular pattern that can be effectively represented and forecast by a univariate time series model.

Some works apply machine learning techniques to the problem of forecasting tourist volumes (see e.g. de Kort, 2017, Feng et al., 2019, Breiman, 2001, Sun et al., 2019, Ji-yuan et al., 2017, Law et al., 2019). These techniques are potentially well suited to account for the large amount of web-search data available. In the early stages of this paper we experimented with some of these techniques, without finding any significant improvement in performance with respect to the simpler, yet effective, more traditional econometric models we use.

3 Data description

3.1 Overnight stays

Our target variable is overnight stays (or nights spent), a statistic defined in the Regulation (EU) No 692/2011 of the European Parliament, as the number of nights that a guest/tourist (resident or non-resident) spends (sleeps or stays) in a registered accommodation (Eurostat, 2021), such as hotels, bed and breakfast or camping sites.[7] All accommodation facilities (including those using online platform services) must communicate electronically to the police authorities the details of their guests as well as their number of overnight stays. This communication[8] has no fiscal purposes, but non-compliance can potentially lead to consequences for the manager. This flow of information is paralleled by an additional one with statistical purposes only[9], which goes to the regional administration statistics offices (or to other similar offices chosen by the regional administration) that collect the data and deliver them to Istat, which is in charge of their control, harmonization and dissemination, consistently with the European regulations.

[7] Overnight stays are the product of the number of the arrivals and the number of nights spent per each tourist arrived. In the estimations, we do not consider arrivals, but we focus exclusively on overnight stays. Over the short run, the dynamic of the two statistics is very similar; over the last ten years, arrivals have grown more because the average number of nights spent per tourist has declined sharply.

[8] Regulated by the "Testo unico di Pubblica Sicurezza" (art.109).

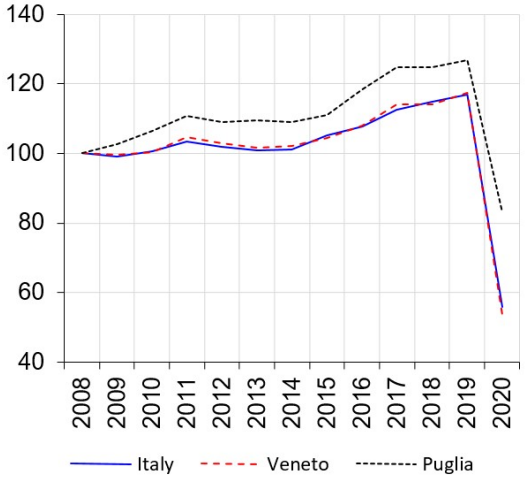
[9] Due according to the Regulation EU 692/2011.

In 2019, more than 430 million overnight stays were registered in the country according to the official data, corresponding to more than 130 million of arrivals. More than 16 per cent of the nights spent pertain to an accommodation located in Veneto and almost 4 per cent to Puglia, the two regions we consider in our estimation exercises.

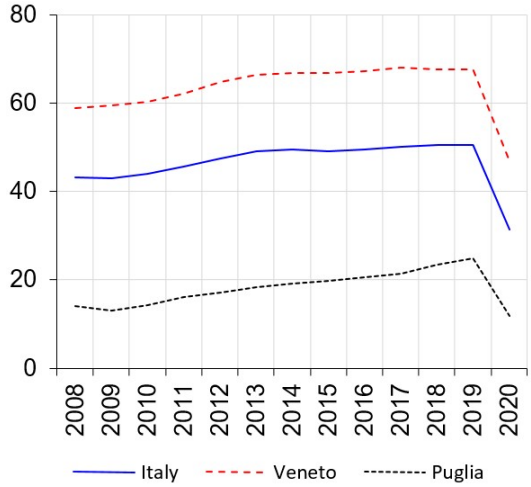
Figure 1a represents the number of overnight stays in the three areas (Italy, Veneto and Puglia). Overall, they display a positive time trend with an increase of approximately 17 per cent in both Italy and Veneto, and of 27 per cent in Puglia, between 2008 and 2019. In 2020 the number of nights spent halved with respect to the previous year in both Italy and Veneto, mainly as a consequence of the mobility restrictions due to the pandemic. In Puglia overnight stays declined less severely (by one third), likely because of the lower share, in this region, of international tourists (Demma, 2021). Figure 1b presents this share: in Italy, international tourists accounted for around half of the total in 2019 and the share declined to one third in 2020. The pattern is very similar for Veneto and Puglia, but the share of inbound tourists in Veneto is three times larger than in Puglia.

Figure 1: Overnight stays in Italy, Veneto and Puglia (period: 2008-2021)

(a) Annual overnight stays



(b) Share of stays by inbound tourists



Note: base is 2008 in the same area.

When disaggregated at the monthly level, the series of overnight stays display strong seasonality: in 2019, almost 14 per cent of the registrations to an accommodation in Italy were in August and 40 per cent were concentrated between June and August; the corresponding figures are even higher for overnight stays (because during summer the length of

stay increases) and for most southern regions (where local tourists represent a larger share). Seasonality is particularly strong in Puglia, where nights spent during August account for around 30 per cent of the annual volume.

3.2 Payment card data

The Bank of Italy acquires, free of charge and on a best-effort basis, a sample of debit and credit card payments by the Italian clearing and processing system, which we call *Pago*, a fictional name.

The data cover the period starting from May 2014, and are updated twice per month with around 10 days of delay. They consist of digital transactions registered by both the *acquiring* side, that is, the merchant side, and the *issuing* side, that is, the cardholder side. In the first case (*acquiring*), data include information on all payments made on a *Pago* point-of-sale (POS), irrespective of the card issuer (which may or may not be *Pago*). In the second case (*issuing*), data include all the payments made with a card issued by *Pago*, irrespective of the issuer of the POS. The dataset is thus not necessarily representative of the non-cash payments in Italy, as it includes information on the *Pago* customers only, whose share of the total payment services clients on the Italian territory is unknown and heterogeneous in time and space. The non-representativeness of the sample is not necessarily an issue for us because we are only interested in an indicator that correlates with the dynamic of the tourist flows for forecasting purposes, while we are not interested in estimating consumption or payments *per se*. Nevertheless, changes in the representativeness of the sample, due for instance to shocks to the preference for electronic payments may reduce the forecasting power of payment data.

The transactions included in the dataset are aggregated by the provider of the data according to: a) 11 product categories, b) type of payment card (credit, debit, prepaid), c) technology used for the payment (e-commerce, physical), and d) geographical area (country and region) of the branch of the bank connected to the *Pago* POS or card.[10]

For the scope of this paper, we only use as exogenous variables *acquiring* transactions, that is the value of payments made on a POS provided by *Pago*. The motivation is that we are interested in building an indicator correlated with the tourist demand in a given area (specifically, Italy, Puglia and Veneto), independently of the region of residence of the cardholder. We then select only the physical transactions of the categories *hotels and restaurants*, and *travel and transport*, obtaining a monthly series of payments expressed in

[10]The product categories are: 1) clothing, 2) hotels and restaurants, 3) food, 4) home, 5) cash advance, 6) work, 7) retail, 8) services, 9) telephony, 10) travel and transport, 11) not defined.

euros. From this series, we extract an additional one, selecting only payments made with card issued by Italian banks, which is plausibly more informative on the payments made by local tourists.

3.3 Google Trends

We use, among the predictors of overnight stays, an index based on the number of queries for words or expressions related to tourism in each area of interest, obtained from GT.[11] GT is an algorithm set up by Google, which provides, for a given word or expression, the number of web searches that have been made on Google Search, in a given period of time, and in a given geographical region.[12] A feature of GT, which is useful for us, is the possibility to select the original scope of the web query among various search categories, including the category *Travels* which is the one we use in this paper.[13]

The routine we develop to extract the GT series is the following. First, we choose the search queries by selecting, separately for each area, the top-10 destinations according to Tripadvisor, an internet source specialized in tourism.[14]. The queries are reported in Table 1.

Table 1: Keywords selected with Tripadvisor for each area.

Italy	**Veneto**	**Puglia**
Roma	Venezia	Vieste
Venezia	Jesolo	Gallipoli
Firenze	Verona	Otranto
Milano	Bibione	Porto Cesareo
Napoli	Caorle	Polignano a Mare
Sorrento	Abano Terme	Lecce
Jesolo	Mestre	Ostuni
Taormina	Bardalino	Bari
Riccione	Peschiera del Garda	Peschici
Rimini	Malcesine	Monopoli

[11] In 2021, Google Trends had around 92 per cent of the worldwide search engine marked share (Statcounter, 2017).

[12] More precisely, the counts, available on a monthly basis, are reported only if exceeding an unknown threshold based on the geographical location, and are measured with a 0-100 index (normalized on the chosen time window). Importantly, the series generated by GT does not provide absolute numbers of searches, but a relative frequency of them. They represent the popularity of the searches for a keyword with respect to the total searches in the geographical area and time period selected, measured in relative terms.

[13] The classification into categories is done by an algorithm set-up by Google about which Google does not release details.

[14] For instance, for Veneto, we consider the top-10 destinations specified in the following webpage: `https://www.tripadvisor.it/Tourism-g187866-Veneto-Vacations.html`.

We then add other generic keywords, such as the name of the area of interest (in Italian, English, French and German), alone, plus the same name concatenated with keywords related to tourism (for instance, "mare" (sea), "spiagge" (beaches), "montagne" (mountains), "lago" (lake), "agriturismo" (agritourism), "parchi" (parks), booking, hotel, airb&b, resort).

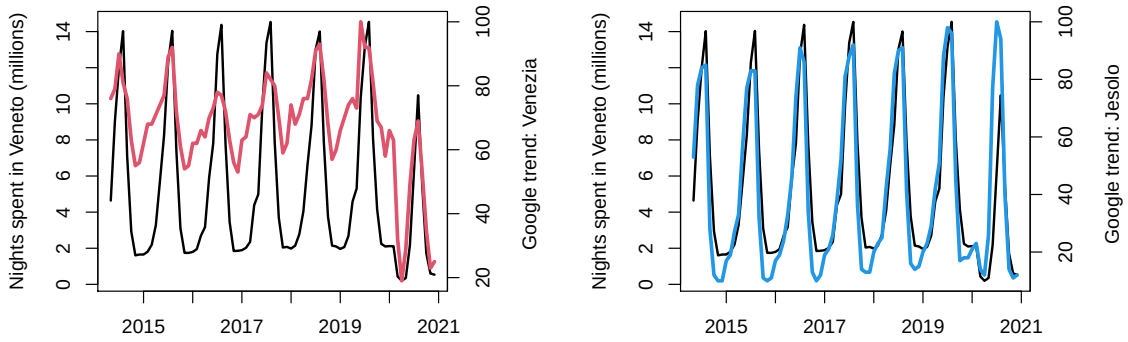
Second, after selecting the period of time, we download from the *Travel* category the monthly GT series corresponding to searches from any location (Italian or abroad). Given that the travel is often after the web search, as in Feng et al. (2019), we choose the *optimal* lag period (between 0 and 12 months) of each series with respect to the target variable, maximizing the Pearson correlation coefficient. Finally, we keep a GT series if its final correlation coefficient with the target variable is above 0.5.

As an example, Figure 2 shows the dynamic of the target variable (overnight stays) for the region Veneto (in black) along with the GT series corresponding to the *Venezia* keyword (in red), left panel, and the *Jesolo* keyword (in blue), right panel (notice that *Venezia* and *Jesolo* are the top 2-destinations for Veneto and among the 10-destinations for Italy, according to Tripadvisor). The high correlation (0.75 for *Venezia*, and 0.89 for *Jesolo*) between the GT series and the target largely reflects their strong seasonal component, although a reasonable correlation persists when these are calculated on the basis of the seasonally adjusted series (0.78 for *Venezia*, and 0.38 for *Jesolo*).

Figure 2: Overnight stays in Veneto (period: 2015-2021)

(a) Overnight stays in Veneto (black line) and GT series with keyword *Venezia*

(b) Overnight stays in Veneto (black line) and GT series with keyword *Jesolo*



Note: Both GT series are not lagged; all series are not seasonally-adjusted.

After having selected the GT series according to the outlined procedure, in order to summarize them into a few non-correlated time series which explain the majority of their

variation, we run a principal component analysis (PCA). In a second moment, we select the number of principal components (PCs) and use them as external regressors in the proposed econometric models.[15]

4 The estimation setting

As discussed in the introduction, we consider two estimation exercises. In the first, that we call "national setting", we are interested in nowcasting the number of overnight stays in the entire Italian territory, $y_{t,IT}$, with the aid of up-to-date information on payment card data and web searches. In the second, the "regional setting", we are interested in nowcasting the number of overnight stays in a given Italian region r, $y_{t,r}$, with the aid of up-to-date information on payment card data and web searches as in the "national setting" case, plus official data on stays at the national level (for the months when they are available). In this exercise we focus on Puglia and Veneto.

Figure 3 visually represents the two estimation settings, which are also described precisely in the following. Given a generic time series Y_t, we use a shortcut to indicate fractions of it: hence $Y_{t_1:t_2} = (Y_t)_{t=t_1}^{t_2}$, is the series Y_t from t_1 to t_2.

National setting: $y_{t,IT}$ is the total number of nights spent in Italy at time t. Reliable data are available with a lag of four months.[16] The "now" is $T = (month/year) = 10/Y$, that is October of a given year Y. We are interested in the summer estimate of $y_{t,IT}$ (i.e. for July, August and September). For the estimation, we use:

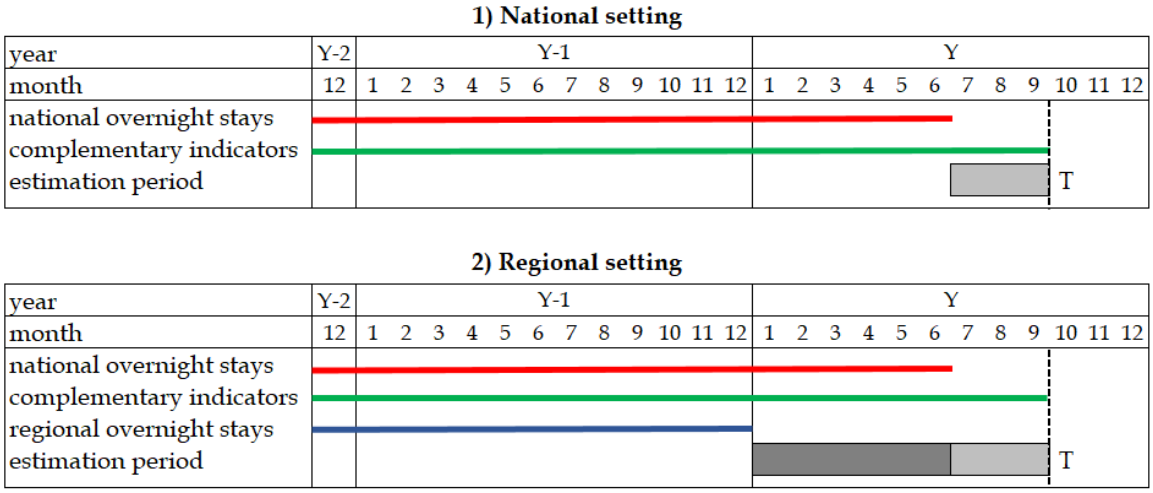
- The total number of nights spent in Italy up to June of the same year, $y_{0:(T-4),IT}$, $T - 4 = 6/Y$;
- Complementary indicators (payment card plus web queries indices) up to September, $\boldsymbol{X}_{0:(T-1)}$, $T - 1 = 9/Y$.

Regional setting: $y_{t,r}$ is the total number of nights spent in a given Italian region r at time t. The data are available with a lag of many months (usually the monthly data for the previous year $Y - 1$ are available around June of a given year Y). The "now" is $T = 10/Y$

[15] In each case we use the scree plot to select the number of PCs, which, as a consequence, may differ according to the area of interest.

[16] Provisional data are delivered from Istat to Eurostat within 56 days from the end of the previous month. Eurostat publishes the data some days after, but they are then revised within four months: at that date, the statistics are published by Istat, too.

Figure 3: Timing: observation and estimation periods in the national and in the regional settings.



Notes: The figure represents visually the timing for both exercises. In the national setting, the night stays are estimated for 3 months in one step, using information on past stays and complementary indicators. In the regional setting, the night stays are estimated for 9 months in two steps. In step (A), regional night stays are predicted using national stays and complementary indicators (6 months); in step (B), only the latter are used (3 months).

as before. We are interested in the summer estimate of $y_{t,r}$ (i.e. for June, July, August and September). We use:

- The regional overnight stays up to December of the previous year, $y_{0:(T-10),r}$, $T-10 = 12/Y - 1$;
- The national overnight stays up to June of the same year, $y_{0:(T-4),IT}$, $T-4 = 6/Y$;
- Complementary indicators (payment card plus web queries) up to September of the same year, $\boldsymbol{X}_{0:(T-1)}$, $T-1 = 9/Y$.

Notice that while the complementary indicators are available for all the months when regional stays are to be estimated, national stays are available only for a sub-period (i.e. until June).

In both settings the training is performed starting from May 2014, because it is the first month when all the complementary indicators are available.[17] We then report the out-of-sample performance of the models starting from the summer 2017, so that we have three years of monthly data to train our models and four years to test their performance.

[17]Unfortunately, as already pointed out, data on payments are not available before 2014.

5 Method

We adopt a standard SARIMA (seasonal autoregressive integrated moving average) model with explanatory variables (also known as SARIMAX, regression with SARIMA errors, or dynamic regression model)

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + \eta_t \tag{1}$$

where, y_t is the number of overnight stays at time t, $\mathbf{x}_t = [x_{1,t}, ..., x_{p,t}]$ is the vector containing p explanatory variables at time t, and η_t is a univariate SARIMA model denoted by ARIMA$(p,d,q)(P,D,Q)_m$, where m is the number of observations per year (12 in our case). The ARIMA$(p,d,q)(P,D,Q)_m$ model takes the form

$$\Phi_P(L^m)\phi_p(L)\Delta_m^D\Delta^d\eta_t = \Theta_Q(L^m)\theta_q(L)\epsilon_t, \tag{2}$$

where L is the lag operator; $\phi_p(L)$ and $\theta_q(L)$ are polynomial functions of L (respectively of order p and q) representing the autoregressive and moving average components; $\Phi_P(L^m)$ and $\Theta_Q(L^m)$ are polynomial functions of L^m (respectively of order P and Q) representing the seasonal autoregressive and moving average components. Finally, Δ^d and Δ_m^D are respectively ordinary and seasonal difference components[18].

While the SARIMA model predicts future values based on the past values of the target variable only,[19] the SARIMAX uses external data in the forecasts, as it includes exogenous (predictor) variables. The SARIMAX model is therefore similar to a multivariate regression model, but it also allows to take into account the autocorrelation that may be present in the residuals of the regression.

5.1 Method for the national setting

The goal is to obtain the estimates of tourist overnight stays in Italy ($y_{t,IT}$) for the summer months of the current year. We assume being in October of year Y (10/Y), when the outcome variable is available until June (6/Y). The SARIMA benchmark model is fitted to $y_{t,IT}, t \leq 6/Y$, and forecast three-steps ahead, obtaining the point estimates corresponding to July, August and September. The same fitting and forecasting are applied to the different versions of the augmented SARIMAX, which include combinations of the *Pago* indicators and the first principal components of the GT as external regressors. The number of principal

[18] Note that, by definition, $Ly_t = y_{t-1}$, and $L^m y_t = y_{t-m}$.

[19] It assumes that future values of the target linearly depend on its past values and on the values of past (stochastic) shocks.

components included is selected visually by detecting an elbow in the scree plot.

5.2 Method for the regional setting

The goal is to obtain an estimate of the tourist overnight stays in a given region r ($y_{t,r}$) for the summer months of the current year. As in the national setting, we assume being in October of year Y (10/Y), and want to obtain estimates for the summer months, from June to September included, of year Y. Recall that in October the target variable, $y_{t,r}$ is only available until December of the previous year (12/Y-1).

We divide the nowcast into two steps: In step (A) we estimate $y_{t,r}$ for the first six months of the current year, exploiting the national series of overnight stays (which is available until 6/Y included), and the complementary indicators. We obtain the predicted values $\hat{y}_{t,r}$ for $1/Y \leq t \leq 6/Y$. In step (B) we estimate $\tilde{y}_{t,r}$ for the three remaining months (7/Y to 9/Y), where

$$\tilde{y}_{t,r} = \begin{cases} y_{t,r}, & \text{if } t \leq 12/(Y-1) \\ \hat{y}_{t,r}, & \text{if } 1/Y \leq t \leq 6/Y \end{cases} \tag{3}$$

using only the complementary indicators, because for the summer months the national overnight stays are not available.

In practice, step (A) goes as follows. Let the percentage share of regional to national overnight stays $y_t = \frac{y_{t,r}}{y_{t,IT}}$ be the target variable.[20] First, we fit a SARIMAX model to $y_t, t \leq T_1$, where $T_1 = 12/(Y-1)$, with the percentage share of regional to national *Pago* payments and the first principal components of the GT as external regressors.[21] Second, we forecast y_t for 6-steps ahead, obtaining six point estimates $\hat{y}_t, t = T_1 + 1, ..., T_1 + 6$, and third, we obtain the six values of the regional tourist overnight stays series by multiplying the forecast values by the actual value of the Italian series: $\hat{y}_{t,r} = \hat{y}_t \cdot y_{t,IT}, t = T_1 + 1, ..., T_1 + 6$.

Equipped with the estimates of the first six months, in step (B), we fit a SARIMAX model on the regional overnight stays, $\tilde{y}_{t,r}$, and forecast the remaining three months, using the regional *Pago* indicators and the first principal components of the GT as external regressors. Note that in this step the target, $\tilde{y}_{t,r}$, is a combination of the actual values $y_{t,r}$ and of the

[20] The reason why step (A) considers as target the time series of the shares, in place of the more natural one in levels, is that we do not want to use $y_{t,IT}$ as exogenous variable in the procedure, because it is too correlated with the $y_{t,r}$. As a consequence, the impact of the other covariates would be artificially smaller. With this model instead, we are able to capture both the correlation between $y_{t,r}$ and $y_{t,IT}$, and the residual correlation between $y_{t,r}$ and the other regressors.

[21] The number of principal components to include in the model is selected for each different year and region, visually, by detecting an elbow in the scree plot. Contrary to the other variables, the principal components of the GT series are included in levels, because GT series are provided by Google as indices and can not be recalculated in regional over national shares.

estimates from step (A) $\hat{y}_{t,r}$ (see eq. (3)). Hence, when constructing the prediction intervals one needs to take into account the fact that, given the auto-regressive structure of the model, step (B) uses the information from step (A) (i.e. the overnight stays estimated for the first six months of the year). To deal with this issue, we perform step (B) in three different scenarios: the *average season scenario*, i.e. with $\hat{y}_{t,r}$ of eq. (3) equal to the mean estimate of step (A); the *low season scenario*, i.e. with $\hat{y}_{t,r}$ of eq. (3) equal to the lower limit of the 80% prediction intervals of the overnight stays estimated in step (A); and the *high season scenario*, i.e. with $\hat{y}_{t,r}$ of eq. (3) equal to the upper limit of the 80% prediction intervals of the overnight stays estimated in step (A). We then construct the lower and upper bounds of the prediction intervals by combining the results of the three scenarios above. This procedure allows to propagate the uncertainty of the estimates of step (A) into the forecasts of step (B).

6 Results

For each of the two settings considered, we try many models (that is, different choices for the (p, d, q, P, D, Q) parameters), by using the stepwise algorithm outlined in Hyndman and Khandakar (2008).[22] We select the best model through the Bayesian Information Criterion (BIC) computed in-sample. As far as the predictors are concerned, our models include, alternatively *Pago* or GT indices; we also show models where both types of regressors are included. We evaluate the out-of-sample performances of our models relative to the benchmark model (SARIMA) through the mean absolute percentage error (MAPE) of the h-steps ahead forecasts. We decided to employ the MAPE instead of more traditional measures of error such as the RMSE because, being a relative error, it is more easily comparable in different scenarios. In fact, the MAPE represents the error in percentage terms, and therefore it is not affected by the size of the target variable.

We focus on the summer months and report both the forecasts in levels, and in percentage change relative to the previous year.[23] All the statistics are calculated on expanding windows, that is, on all available historical data.

[22]The algorithm combines unit root tests, minimization of the Akaike Information Criterion (AIC) and MLE to obtain an ARIMA model.

[23]Recall that in the regional exercise we consider the months from June to September included. In the national case we exclude June, because official data for that month are already available in October of the same year.

6.1 National setting

In Table 2 we report the out-of-sample performance of each model for the national setting: in particular, the table shows the MAPE of the forecasts for the entire summer season (July, August and September) for 2017, 2018, 2019 and 2020 (the "COVID year"). We also show the average MAPE measured over the entire period (2017-2020) and the average MAPE measured during "regular times" (that is, during the pre-COVID years, 2017-2019). The reason is that we expect the augmented models to perform very differently with respect to the benchmark during regular times and during the COVID year.

Table 2: MAPE results of different specifications of the model

Model	2017	2018	2019	2020	ave. 17-19	ave. 17-20
0. benchmark (SARIMA)	1.55	1.49	1.33	61.63	1.46	16.50
1. `Pago Ita`	0.63	4.72	0.03	2.68	1.79	2.02
2. `Pago all`	0.32	4.22	0.51	23.05	1.68	7.02
3. `Pago Ita + Pago all`	0.49	0.02	0.77	11.23	0.43	3.13
4. `GT only (2 PCs)`	0.50	1.64	1.29	40.51	1.14	10.98
5. `Pago Ita + GT (2 PCs)`	0.29	4.78	0.67	31.83	1.91	9.39
6. `Pago all + GT (2 PCs)`	2.39	4.38	1.24	15.50	2.67	5.88
7. `Pago Ita + Pago all + GT (2 PCs)`	0.37	1.10	1.55	7.28	1.00	2.57

Notes: The first row represents the MAPE in each year (summer season: July to September) for the SARIMA model (no additional covariates). In the following rows, we add the indicated variables and show the corresponding MAPE. The best performing model in each period is indicated in gray.

The first takeaway from the table is that all augmented models perform better than the benchmark if one considers the entire 2017-2020 period (last column of the table). Overall, the best performing model is model 1, which is the SARIMAX with the *Pago* index built by aggregating Italian cards only (`Pago Ita`).[24] The second best performing model over the same time span is model 7, that is the SARIMAX with `Pago ita` plus the *Pago* index built by aggregating all cards (`Pago all`),[25] and the first two principal components of the GT (`GT (2 PCs)`). In both cases the MAPE is around 2 (2.02 and 2.57, precisely), while the same indicator for the benchmark model is 16.5. It does not surprise that most of the improvement of the augmented models spurs from the year 2020, when the pandemic negatively affected the tourist sector. In this case, the addition of external predictors to the benchmark model considerably improves the nowcasting performance, because it helps to quickly adapt to the

[24]This monthly index corresponds to the total amount of acquiring transactions (made on a *Pago* Italian POS) in the tourist industry categories (hotel and restaurant plus travel and transport) made by cards issued by Italian banks.

[25]This corresponds to the total amount of acquiring transactions (made on a *Pago* Italian POS) in the tourist industry categories (hotel and restaurant plus travel and transport) made by cards issued by any national or international bank.

abrupt and unanticipated decline in tourist volumes. Nevertheless, model 7 performs slightly better than the benchmark even during regular times (second to the last column of the table), while the performance of model 1, in the same period, is comparable to the benchmark (the MAPE of model 1 is 1.79, the one of the benchmark is 1.46). The best model in regular times is model 3, which is the SARIMAX with `Pago Ita` and `Pago all` as predictors.

In Table 3 we report the forecasting results obtained with model 1, both in levels (upper panel) and in yearly percentage change.

Table 3: Forecasting results (model 1): aggregated national overnight stays (levels and percentage change)

	2017	2018	2019	2020
Levels				
Official data	203.76	203.62	205.65	131.44
Forecast	202.47	213.23	205.71	134.97
	[195.84,209.1]	[207.04,219.41]	[199.45,211.98]	[115.74,154.2]
Percentage change				
Official data	2.22	-0.07	1.00	-36.09
Forecast	1.57	4.65	1.03	-34.37

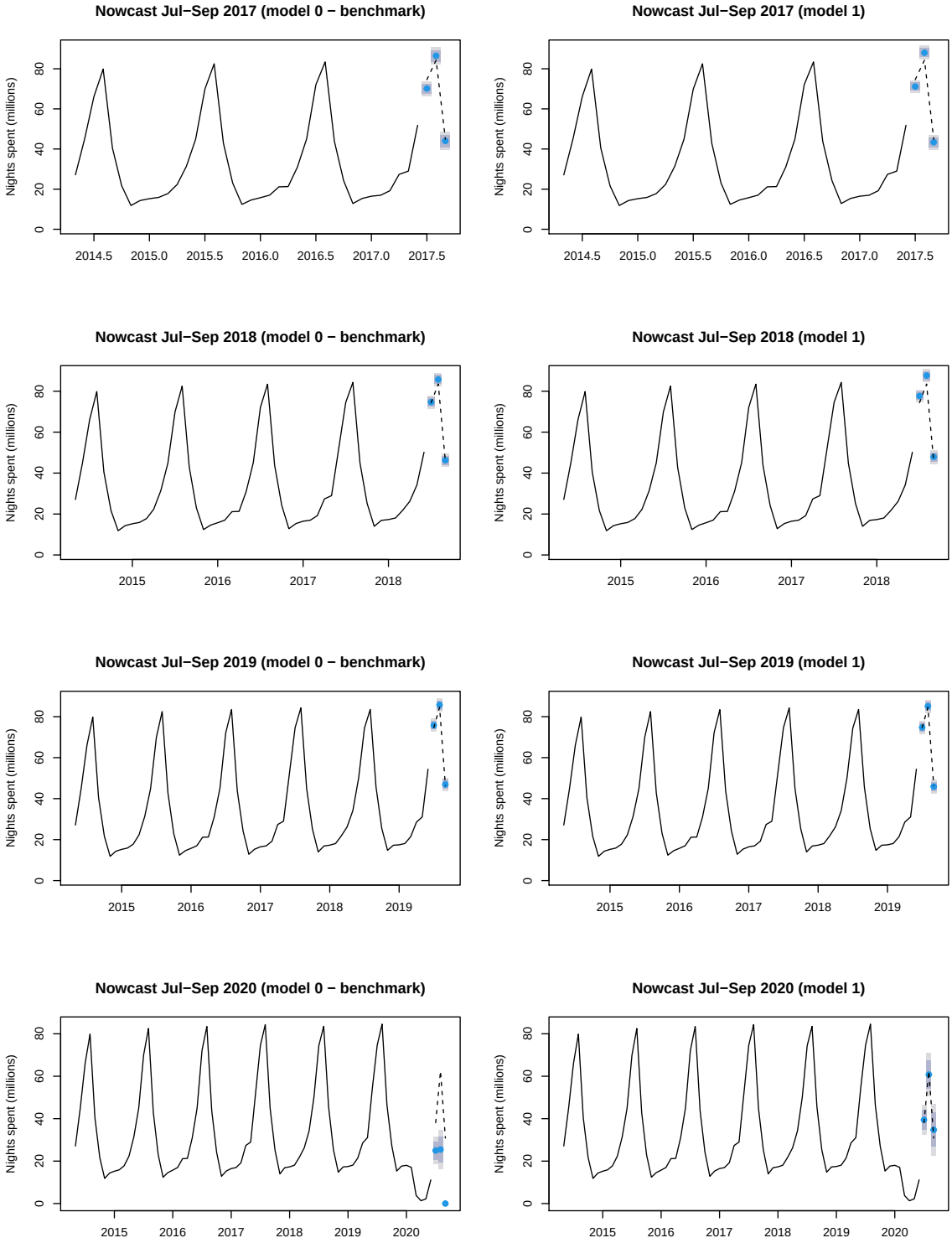
Notes: Results for the summer season (July to September) of each year. In the panel above (levels), the rows represent the official number of stays published by Istat (in millions) and the point estimates obtained with model 1 (in millions). 80% confidence intervals in brackets. In the panel below (percentage change), the rows represent the official data, and the forecasts of the yearly percentage change of overnight stays.

We notice that model 1 performs very well, apart from the year 2018, where there is a substantial overestimation of the official value published by Istat (forecast: 213 million; official data: 204 million). Note that summer 2018 had approximately the same volume of tourists of 2017 (which results in an official value of -0.07 year percentage change), in contrast with a positive trend of the previous summers (see also Figure 4). As a consequence, in this year even the benchmark model overestimates the correct figure, while model 3, which includes both `Pago Ita` and `Pago all`[26], has a very low MAPE value (0.02). This suggests that in year 2018, the index built by aggregating only payments made by cards issued by foreign banks has more predictive power with respect to the other two *Pago* series.[27] The richer model 7, summarized in Table 4, also performs quite well, especially before the year 2020. In 2020, instead, the forecast value is smaller than the official value, implying an estimated percentage change relative to 2019 of almost -41% (5% less than the Istat value -36%).

[26]This model is equivalent to including both `Pago Ita` and `Pago foreigners = Pago all - Pago Ita`.

[27]In practice, this happens because the total number of payments by cards issued by foreign banks is more stable in the years considered, while the payments by Italian-issued cards show a positive trend (data not shown for confidentiality reasons).

Figure 4: Predicted number of night stays by model 0 and by model 1



Notes: The predictions are represented by the blue dots, with 80% (dark-gray) and 95% (light-gray) confidence intervals. The dotted lines corresponds to the official values released by Istat.

Table 4: Forecasting results (model 7): aggregated national overnight stays (levels and percentage change)

	2017	2018	2019	2020
Levels				
Official data	203.76	203.62	205.65	131.44
Forecast	204.51	205.85	202.48	121.87
	[200.2,208.83]	[202.16,209.54]	[198.31,206.64]	[117.26,126.48]
Percentage change				
Official data	2.22	-0.07	1.00	-36.09
Forecast	2.60	1.03	-0.56	-40.74

Notes: Results for the summer season (July to September) of each year. In the panel above (levels), the rows represent the official number of stays published by Istat (in millions) and the point estimates obtained with model 1 (in millions). 80% confidence intervals in brackets. In the panel below (percentage changes), the rows represent the official data, and the forecasts of the yearly percentage change of overnight stays.

6.2 Puglia

As already discussed, the regional estimation consists of two steps: in step (A), the number of regional night stays is estimated for 6 periods ahead, using the national figures and the complementary exogenous variables; in step (B) the target variable is forecasted for the next 3 periods using only the complementary exogenous variables as predictors.

We experimented with different choices of exogenous variables and found that the SARIMAX with the *Pago* index built by aggregating transactions by Italian cards in POS located in Puglia as the unique exogenous variable (model 1, `Pago Ita (P)`) is the best model (see Table 5) in terms of average MAPE over the entire period (last column of the table). This result is consistent with what has been found in the national setting. Specifically, the average MAPE of summer estimates for this model is almost 3 while for the benchmark it is more than 15, a considerable improvement in the estimates, due to the results for the year 2020. In fact, in 2020, the MAPE of the benchmark is almost 55, while the MAPE of model 1 is only approximately 1. Differently from the national case, no model outperforms the benchmark on average during regular times (second to the last column of the table). Apparently, the seasonal and trend components included in the SARIMA are able to capture relatively well the evolution of regional night stays for Puglia.

In Table 6 we present the predictions for the overnight stays in Puglia with model 1. The first row reports the official values published by Istat, while rows 2-4 show the estimation results for the three scenarios considered (average, high and low season, respectively), as explained in Section 5.2. The presence of the three scenarios gives us a heuristic rule to be more conservative and to propagate the uncertainty of step (A) in the estimates of step (B):

Table 5: Puglia. MAPE results of different specifications of the model

Model	2017	2018	2019	2020	ave. 17-19	ave. 17-20
0. benchmark (SARIMA)	1.19	3.14	2.51	54.46	2.28	15.32
1. `Pago Ita (P)`	3.15	5.08	2.70	1.03	3.64	2.99
2. `Pago all (P)`	0.37	6.31	3.24	16.89	3.31	6.70
3. `Pago Ita (P) + Pago all (P)`	0.71	8.19	3.31	19.53	4.07	7.93
4. `GT only (2 PCs)`	2.72	3.06	2.81	10.21	2.86	4.70
5. `Pago Ita (P) + GT (2 PCs)`	2.04	4.50	3.70	2.48	3.41	3.18
6. `Pago all (P) + GT (2 PCs)`	2.53	5.92	2.94	11.98	3.80	5.84
7. `Pago Ita (P) + Pago all (P) + GT (2 PCs)`	0.85	9.00	2.88	18.59	4.24	7.83

Notes: The first row represents the MAPE in each year (summer season: June to September) for the SARIMA model (no additional covariates). In the following rows, we add the indicated variables and show the corresponding MAPE. The best performing model in each period is indicated in gray.

Table 6: Puglia. Forecasting results (model 1): aggregated regional overnight stays (levels and percentage change)

	2017	2018	2019	2020
Levels				
Official data	11.95	11.71	11.78	8.63
Forecast [Average]	12.33	12.3	12.1	8.52
	[12.05,12.6]	[12.06,12.54]	[11.8,12.4]	[7.86,9.18]
Forecast [High]	12.55	12.45	12.25	8.65
	[12.27,12.84]	[12.21,12.69]	[11.95,12.56]	[8.01,9.3]
Forecast [Low]	12.14	12.13	11.95	8.33
	[11.86,12.41]	[11.88,12.39]	[11.65,12.25]	[7.66,9]
Percentage change				
Official data	5.23	-2.06	0.64	-26.75
Forecast [Average]	8.54	2.92	3.36	-27.68
Forecast [High]	10.51	4.18	4.66	-26.55
Forecast [Low]	6.87	1.51	2.08	-29.32

Notes: Results for the summer season (June to September) of each year. In the panel above (levels), the rows represent the official number of stays published by Istat (in millions), the point estimates of the average season scenario, the point estimates of the high season scenario and the point estimates of the low season scenario (see Section 5.2). 80% confidence intervals in brackets. In the panel below (percentage changes), the rows represent the official and the forecasts of the yearly percentage change of overnight stays obtained in the three scenarios.

in fact, we could take as final estimate the average over the 3 scenarios, with final CI built by taking the two more extreme points of the CIs of the scenarios above. In the lower panel of the same table we report the official and predicted yearly percentage changes of the total overnight stays of summer relative to the same period of the year before. We see that the estimates of the average scenario are quite far from the official values (in 2018 our estimate is almost 5 points larger than the official value). However, in 2020 our model is able to predict well the abrupt decline in the tourist sector (average scenario forecast value: -27.68%; official

value released by Istat: -26.75%).

6.3 Veneto

Applying the same procedure to Veneto, we found that the best model, in terms of average MAPE, is the SARIMAX with the *Pago* index built by aggregating all (Italian and foreign) payment by cards in Veneto (`Pago all (V)`) plus the first three principal components of the GT series for Veneto as exogenous variables (model 6, see Table 7). In particular, the average MAPE of this model for the entire period (last column of the table) is slightly smaller than 5 (4.68), while the one of the benchmark model is more than 16 (16.05), a considerable improvement in the estimates, even if more limited with respect to the national case and to the case of Puglia.

Table 7: Veneto. MAPE results of different specifications of the model

Model	2017	2018	2019	2020	ave. 17-19	ave. 17-20
0. benchmark (SARIMA)	4.47	5.90	0.45	53.39	3.61	16.05
1. `Pago Ita (V)`	2.50	7.88	1.25	25.05	3.88	9.17
2. `Pago all (V)`	1.50	3.08	0.49	18.34	1.69	5.85
3. `Pago Ita (V) + Pago all (V)`	3.99	1.85	1.07	12.70	2.30	4.90
4. `GT only (3 PCs)`	2.85	5.25	0.61	11.21	2.90	4.98
5. `Pago Ita (V) + GT (3 PCs)`	1.80	7.72	0.22	13.92	3.25	5.92
6. `Pago all (V) + GT (3 PCs)`	4.57	4.12	0.21	9.83	2.97	4.68
7. `Pago Ita (V) + Pago all (V) + GT (3 PCs)`	4.00	2.85	0.52	13.10	2.46	5.12

Notes: The first row represents the MAPE in each year (summer season: July to September) for the SARIMA model (no additional covariates). In the following rows, we add the indicated variables and show the corresponding MAPE. The best performing model in each period is indicated in gray.

As before, the low value of the average MAPE for the best model is driven mostly by the year 2020. Considering only regular times (second to the last column of the table), the best-performing model is model 2, which includes `Pago all (V)` only. This difference with respect to the national case and the case of Puglia (where the models including `Pago Ita (V)` generally perform better) may be due to the higher share of international tourist flows to Veneto in regular times. Figure 1b shows in fact that in the window 2017-2019, the share of tourism from abroad is almost 70% in Veneto, around 50% in Italy, and only around 20% in Puglia.

In Table 8, we present the forecasting results obtained with model 6. Note that, as in the case of Puglia, the forecasts of the overnight stays in the average season scenario (second row of the table) are generally upward biased (apart from 2019) and include the true value

Table 8: Veneto. Forecasting results (model 6): aggregated regional overnight stays (levels and percentage change)

	2017	2018	2019	2020
Levels				
Official data	45.73	43.69	46.12	24.42
Forecast [Average]	47.83	45.49	46.03	26.52
	[46.77,48.89]	[44.06,46.91]	[44.77,47.29]	[24.97,28.08]
Forecast [High]	48.22	45.72	46.12	27.09
	[47.18,49.27]	[44.33,47.1]	[44.82,47.42]	[25.54,28.63]
Forecast [Low]	47.43	45.24	45.95	25.96
	[46.29,48.56]	[43.7,46.77]	[44.7,47.21]	[24.39,27.54]
Percentage change				
Official data	7.09	-4.48	5.58	-47.05
Forecast [Average]	11.99	-0.54	5.36	-42.50
Forecast [High]	12.92	-0.04	5.57	-41.27
Forecast [Low]	11.05	-1.09	5.19	-43.71

Notes: Results for the summer season (June to September) of each year. In the panel above (levels), the rows represent the official number of stays published by Istat (in millions), the point estimates of the average season scenario, the point estimates of the high season scenario and the point estimates of the low season scenario (see Section 5.2). 80% confidence intervals in brackets. In the panel below (percentage changes), the rows represent the official and the forecasts of the yearly percentage change of overnight stays obtained in the three scenarios.

in the CI only in 2019. However, the CIs built taking as bounds the more extreme points of the CIs of the three scenarios, contains the official values in all years apart from 2017.

7 Conclusion

We presented a strategy to nowcast tourist volumes in Italy and its regions, with a simple econometric model augmented with complementary indicators based on payment card and internet search data.

Concerning payment card data, our results show that some indicators built from this dataset are highly informative for predicting tourist volumes, especially in the national setting. The indicator that contributes the most to nowcasting changes depending on the considered area: the best performing model for Veneto includes payments of both Italian and foreign cards, while the best one for Puglia and for the national case includes the payments of cards issued by Italian banks only; this difference may be due to the higher share of inbound foreign tourists which characterizes Veneto with respect to the other areas.

The indices built from the Google Trends series have a small impact on the out-of-sample performances when we consider the national setting, but they improve the performance in the regional setting, at least in the case of Veneto. There are multiple possible explanations for

this result. The first one is that the "short" list of keywords for Italy, like the one we chose, is not sufficiently detailed for capturing tourist flows in a large area like the entire Italian territory. In fact, the top$-n$ destinations in Veneto and Puglia account for a much larger share of overnight stays in each of the two regions than the top$-n$ destinations in Italy. If this is the case, enriching the list of keywords may potentially increase the prediction accuracy of the models at the national level. A second possible explanation is that web searches for small municipalities, which are more often included as regional keywords, are more likely to be motivated by travel planning: in other terms, apart from tourism purposes, there are several reasons one may search on the internet the keywords "Rome" or "Milano", but perhaps very few ones to search "Vieste" or "Bibione". A third, possibly more plausible reason, is that the predictive power of such indices increases with the share of foreign tourists, because these are more likely to collect information on travel destinations through the web.

We illustrated the results of the model for the national setting and two specific Italian regions, but the outlined methodologies can be immediately applied to any other region (or, eventually, to Italian macro-regions). Similarly, we focused on summer, but the strategy is generalizable to any other time-interval of interest. Note, however, that our results show that the best performing indicators, among the ones considered, are not stable when applied to different regions and to different time intervals. This fact is not surprising and it is a well known result in the empirical forecasting literature (see, for instance, Giacomini and Rossi, 2010, Rossi and Sekhposyan, 2010), where it is documented that the predictive ability of forecasting models generally varies through time, implying that different predictors may perform well in some periods of time and specific areas, but not in others.

A limitation of the proposed strategy is that it can only produce a nowcast of the variable interest. To overcome this limitation, one may forecast the indicators, and then estimate the model. However, this procedure would generate many issues regarding the correct estimation and propagation of the errors of the first stage estimates which put it beyond the scope of this paper.

This work is a first attempt to provide up-to-date indicators for the tourist industry in Italy and its regions, and leaves room for improvements and further investigations. One way to improve the method is to extend it in order to include some additivity constraints, in the same spirit as the indicator of the regional economic cycle outlined in Di Giacinto et al. (2019). The idea is to exploit the fact that all the regional series have to sum up to the national one, in each time period, in order to improve the current biases in the estimates, and do this by adding a contemporaneous aggregation constraint based, for instance, on a

proportional adjustment (see e.g. Fonzo and Marini, 2011). A second direction for future developments is to include other variables which may prove useful for nowcasting. Examples are mobile phone data (in particular, the number of foreign mobile SIM cards used in a given area), data on job hiring and separations for workers in the tourist sector, or information on the use of social safety nets (as, for instance, "Cassa integrazione guadagni") by workers employed in tourism firms. Both directions will be explored in future developments of this paper.

In general, we are aware that the proposed methods may be over-simplistic, and that more sophisticated and up-to-date econometric models may be preferred. However, while many papers have already discussed the pros and cons of using web search data, to the best of our knowledge the use of payment card data for predicting tourist flows is a novelty. Moreover, even if our specifications are quite parsimonious in terms of data requirements, their predictive power is generally higher than the one from a SARIMA benchmark model. As a consequence, we believe the discussion of these models represents a first valuable step toward a deeper understanding of the informativeness of the indices proposed and their relevance for nowcasting tourist volumes.

References

Aastveit, K. A., Fastbø, T. M., Granziera, E., Paulsen, K. S., and Torstensen, K. N. (2020). Nowcasting Norwegian household consumption with debit card transaction data. *Norges Bank*.

Aladangady, A., Aron-Dine, S., Dunn, W., Feiveson, L., Lengermann, P., and Sahm, C. (2021). *From Transactions Data to Economic Statistics: Constructing Real-Time, High-Frequency, Geographic Measures of Consumer Spending*. University of Chicago Press.

Antolini, F. and Grassini, L. (2019). Foreign arrivals nowcasting in Italy with Google Trends data. *Quality & Quantity*, 53.

Aprigliano, V., Ardizzi, G., and Monteforte, L. (2019). Using Payment System Data to Forecast Economic Activity. *International Journal of Central Banking*, 15(4):55–80.

Ardizzi, G., Nobili, A., and Rocco, G. (2021). A game changer in payment habits: Evidence from daily data during a pandemic. *Social Science Research Network*.

Artola, C. and Martínez-Galán, E. (2012). Tracking the future on the web: construction of leading indicators using internet searches. *Banco de Espana Occasional Paper*, (1203).

Askitas, N., Zimmermann, K., (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly (formerly: Konjunkturpolitik)*, 55:107–120.

Bangwayo-Skeete, 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.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Camacho, M. and Pacce, M. J. (2018). Forecasting travellers in Spain with Google's search volume indices. *Tourism Economics*, 24(4):434–448.

Carboni, A., Doria, C., and Catalano, C. (2022). How big data can improve the quality of tourism statistics? The Bank of Italy experience in compiling the BoP travel item.

de Kort, R. E. (2017). Forecasting tourism demand through search queries and machine learning. *IFC Bulletins*, 44.

Della Corte, V., Doria, C., and Oddo, G. (2021). The impact of Covid-19 on international tourism flows to Italy: evidence from mobile phone data.

Della Penna, N. and Huang, H. (2009). Constructing Consumer Sentiment Index for U.S. Using Google Searches. Working Papers 2009-26, University of Alberta, Department of Economics.

Demma, C. (2021). Il settore turistico e la pandemia di Covid-19. *Note Covid-19*.

Di Giacinto, V., Monteforte, L., Filippone, A., Montaruli, F., and Ropele, T. (2019). ITER: a quartely indicator of regional economic activity in Italy. *Questioni di Economia e Finanza*, (489).

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.

Eurostat (2021). Eurostat statistics explained. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Nights_spent`. Accessed: 2021-07-26.

Feng, Y., Li, G., Sun, X., and Li, J. (2019). Forecasting the number of inbound tourists with google trends. *Procedia Computer Science*, 162:628 – 633.

Fonzo, T. D. and Marini, M. (2011). Simultaneous and two-step reconciliation of systems of time series: methodological and practical issues. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 60(2):143–164.

Galbraith, J. W. and Tkacz, G. (2018). Nowcasting with payments system data. *International Journal of Forecasting*, 34(2):366–376.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

Havranek, T. and Zeynalov, A. (2019). Forecasting tourist arrivals: Google trends meets mixed-frequency data. *Tourism Economics*.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software, Articles*, 27(3):1–22.

Istat (2020). Conto satellite del Turismo per l'Italia, year 2017. Statistiche report. `https://www.istat.it/it/files//2020/06/Conto-satellite-turismo.pdf`. Accessed: 2022-01-05.

Istat (2021). Occupancy in collective tourist accomodation. `https://www.unwto.org/country-profile-outbound-tourism`. Accessed: 2021-07-20.

Ji-yuan, W., Geng, P., and Shou-yang, W. (2017). Model selection on tourism forecasting: A comparison between Bayesian model averaging and Lasso. *African Journal of Business Management*, 11:158–167.

Law, R., Li, G., Fong, D. K. C., and Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, 75:410 – 423.

Park, S., Lee, J., and Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing*, 34(3):357–368.

Petrella, A., Torrini, R., Barone, G., Beretta, E., Breda, E., Cappariello, R., Ciaccio, G., Conti, L., David, F., Degasperi, P., Di Gioia, A., Felettigh, A., Filippone, A., Firpo, G., Gallo, M., Guaitini, P., Papini, G., Passiglia, P., Quintiliani, F., Roma, G., Romano, V.,

and Scalise, D. (2019). Turismo in Italia: numeri e potenziale di sviluppo. *Questioni di Economia e Finanza*, 606:1 – 113.

Rossi, B. and Sekhposyan, T. (2010). Have economic models' forecasting performance for US output growth and inflation changed over time, and when? *International Journal of Forecasting*, 26(4):808–835.

Statcounter (2017). Search Engine Market Share Worldwide. `https://gs.statcounter.com/search-engine-market-share#quarterly-200901-201702`. Accessed: 2021-07-20.

Sun, S., Wei, Y., Tsui, K.-L., and Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, 70:1 – 10.

UNWTO (2021). Data on outbound tourism by country. `https://www.unwto.org/country-profile-outbound-tourism`. Accessed: 2021-07-20.

Varian, H. and Choi, H. (2009). Predicting the Present with Google Trends. *Economic Record*, 88.

Verbaan, R., Bolt, W., and van der Cruijsen, C. (2017). Using debit card payments data for nowcasting Dutch household consumption. DNB Working Papers 571, Netherlands Central Bank, Research Department.

Webb, G. (2009). Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures. *Issues in Information Systems*, 10.

Wu, L. and Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In *Economic Analysis of the Digital Economy*, pages 89–118. National Bureau of Economic Research, Inc.

Yang, X., Pan, B., Evans, J. A., and Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, 46:386 – 397.

Yoon, Y. and Uysal, M. (2005). An examination of the effects of motivation and satisfaction on destination loyalty: a structural model. *Tourism Management*, 26(1):45–56.