Article

TOURISM ECONOMICS
The Business and Finance of Tourism and Recreation

Tourism Economics
2021, Vol. 27(1) 129–148

DOI: 10.1177/1354816619879584
journals.sagepub.com/home/teu
SAGE

Forecasting tourist arrivals: Google Trends meets mixed-frequency data

Tomas Havranek
Charles University, Prague

Ayaz Zeynalov
University of Economics, Prague

Abstract
In this article, we examine the usefulness of Google Trends data in predicting monthly tourist arrivals and overnight stays in Prague during the period between January 2010 and December 2016. We offer two contributions. First, we analyze whether Google Trends provides significant forecasting improvements over models without search data. Second, we assess whether a high-frequency variable (weekly Google Trends) is more useful for accurate forecasting than a low-frequency variable (monthly tourist arrivals) using mixed-data sampling (MIDAS). Our results suggest the potential of Google Trends to offer more accurate predictions in the context of tourism: we find that Google Trends information, both 2 months and 1 week ahead of arrivals, is useful for predicting the actual number of tourist arrivals. The MIDAS forecasting model employing weekly Google Trends data outperforms models using monthly Google Trends data and models without Google Trends data.

Keywords
forecasting, Google Trends, mixed-frequency data, tourism

Introduction

Municipal authorities and hotel chain management need good estimates of the number of people who will visit their city in a given month. To generate such estimates, the most frequently used models are simple time series autoregression models. (These are in contrast to the more structural gravity models, which examine the determinants of bilateral tourism flows to and from various

Corresponding author:
Ayaz Zeynalov, Jan Masaryk Centre of International Studies, Faculty of International Relations, University of Economics, Prague, W. Churchill 4, 130 67 Prague 3, Czech Republic.
Emails: ayaz.zeynalov@vse.cz

regions but are of little use in forecasting the aggregate tourism flow to a single city.) Several recent studies have employed Google Trends data to enrich the basic autoregression models for aggregate tourism flow forecasting, and we build on this new stream of literature. Our main contribution is our combination of Google Trends data and mixed-frequency modeling using mixed-data sampling (MIDAS) models. Because tourism arrivals are available on a monthly basis, while search data are also available but at higher frequencies, mixed-frequency modeling allows us to use more up-to-date information. Employing data for Prague, we show that such an enriched mixed-frequency model dominates the commonly used alternatives.

The years 2018 and 2019 have witnessed a flurry of new techniques and data sources for tourism modeling in general and heightened activity based on Google Trends data in particular. Recent research by Bokelmann and Lessmann (2019) shows that Google Trends is a useful tool for forecasting short-term tourism demand in Germany. Law et al. (2019) forecast monthly Macau tourist arrivals to demonstrate that the deep learning approach significantly outperforms vector regression and artificial neural network (ANN) models. Clark et al. (2019) emphasize the advantage of Google Trends in forecasting visitation for most national parks in the United States. Li and Law (2020) examine whether decomposed search engine data can be used to improve accuracy in forecasting tourism demand to predict monthly tourist arrivals from nine countries to Hong Kong. Camacho and Pacce (2018) use a dynamic factor approach to produce real-time predictions for the overnight stays of travelers in Spain. Wen et al. (2019) combine the linear and nonlinear features of component models and claim that the new "hybrid" model is better than other models when forecasting tourist arrivals in Hong Kong from mainland China.

In this article, we evaluate three different weighted MIDAS models using weekly data, ARIMA(1,1,1) with monthly Google Trends information, and a model without search data. The main objective is to assess whether Google Trends information provides significant benefits to the evaluation and forecasting of tourist arrivals and overnight stays in Prague and whether models with higher frequency data (weekly data) outperform models that rely on a single data frequency. Our results highlight the strong potential of Google Trends to improve forecasting power in the case of Prague tourism flows. MIDAS allows the evaluation of series with different frequencies, such as weekly Google Trends information and monthly tourist data. The MIDAS-Beta model for tourist arrivals and the weekly MIDAS-Almon model for overnight stays outperform the models using monthly Google Trends information and the model without Google Trends information. The results confirm that using data from Google Trends enriches the information set available to policy makers and business entrepreneurs operating in the tourism sector. Moreover, our findings stress the usefulness of mixed-frequency models for tourism modeling.

The remainder of this article is organized as follows. The "Literature review" section discusses the literature on tourist arrival forecasting and Google Trends. The "Methodology and data" section discusses the methodology and data sampling. The "Results" section presents the empirical results from applying MIDAS models to tourist arrivals and overnight stays. The "Concluding remarks" section concludes the article. Robustness checks are presented in Online Appendix.

Literature review

Tourism forecasting has been the focus of many studies. Researchers have analyzed tourist demand using two price indices from origin and destination countries to evaluate performance when forecasting tourist preferences based on tourist arrivals to Spain (Gonzalez and Moral, 1995) and have developed forecasting models based on different time series methods using tourist flows from

China, South Korea, the United Kingdom, and the United States to Hong Kong (Song et al., 2011). Researchers have used different time series models to assess the determinants of tourist arrivals (Akin, 2015; Athanasopoulos et al., 2011) and have proposed ANN methods (Claveria and Torra, 2014; Hadavandi et al., 2011). The main objective of Claveria and Torra (2014)'s study was to determine which method provided the most accurate information on tourist numbers; they found that autoregressive integrated moving average (ARIMA) models outperformed self-exciting threshold autoregressive and ANN models. A meta-analysis in this literature performed by Peng et al. (2014) claims that the choice of forecasting method is the main reason for contradictory results among studies.

The usefulness of Google Trends data to predict tourism has also been examined. Bangwayo-Skeete and Skeete (2015) suggest that Google search volume provides advantages for tourism demand forecasting for Caribbean destinations. Researchers have argued that Google Trends, as a concurrent indicator, could promote more precise forecasting for Switzerland (Siliverstovs and Wochner, 2017) and that a strong correlation exists between hotel visitors and Google search queries in Puerto Rico (Rivera, 2016). Park et al. (2017) focus on the short-term forecasting of tourist outflows from South Korea to Japan. They claim that Google Trends data not only improve the precision of tourism demand forecasting but also that out-of-sample forecasting outperforms in-sample forecasting with Google Trends.

Prague is one of the most popular destinations on the European continent, with more than 6 million foreign visitors annually, accounting for up to 15 million overnight stays. Tourism makes a major contribution to Prague's economic development: It accounts for 9% of gross domestic product (GDP) and provides employment for approximately 17% of the working population in the service sector.[1] Therefore, accurate forecasts of tourism volume play a major role in tourism planning, as forecasts enable destinations to predict infrastructure development needs.

Google Trends provides free, vast, and almost real-time information but has some disadvantages. First, Google shows only absolute data, providing an index that is relevant to all searches. Second, Internet users might type similar words when searching for different topics or different words when searching for the same topic. Third, web search queries are related to personal characteristics, such as education, income, and age. Specifically, we do not know what share of tourist arrivals can be related to Internet usage. Fourth, Russia and China are examples of countries that lag in the use of Google search, as Yandex and Baidu, respectively, still maintain their dominant positions. Therefore, results based on Google Trends cannot be generalized to all countries. Clearly, data from Google searches are imperfect; however, because Google Trends provides one of the best real-time information databases, it has the potential to act as a leading indicator for many regions.

The MIDAS method proposed by Ghysels et al. (2006) was further developed by Andreou et al. (2010) who introduced a new decomposition for MIDAS regression. Empirical studies in the MIDAS literature have analyzed the dynamics of microstructure noise and volatility (Ghysels et al., 2007), GDP growth forecasting (Ghysels and Wright, 2009; Andreou et al., 2012), nowcasting and quarterly GDP growth forecasting in the euro area (Kuzin et al., 2011), and stock market volatility and macroeconomic activity (Engle et al., 2013; Girardin and Joyeux, 2013). Götz et al. (2014) developed an alternative mixed-frequency error–correction model for nonstationary variables sampled at different frequencies that are possibly co-integrated. Co-integrated MIDAS has also been introduced by Miller (2016) who focus on efficient estimation of the co-integrating vector of the model with a low-frequency and high-frequency series. MIDAS is a method for estimating and forecasting the impact of high-frequency variable(s) on low-frequency dependent variables that can avoid the traditional requirement that variables have the same frequency. MIDAS uses a distributed lag of polynomials to ensure parsimonious specifications for handling series sampled at different frequencies.

Liu et al. (2019) present a detailed scientometric analysis of tourism forecasting based on 543 articles from the Web of Science. The results suggest that a prominent emerging trend is the use of methods based on data from Web-based search engines. Jiao and Chen (2019) review 72 papers published in tourism forecasting from 2008 to 2017 and conclude that mixed-frequency, spatial regression, and hybrid models are the most prominent emerging methods in the field. Volchek et al. (2019) compare mixed-frequency data sampling with ANN models and claim that higher frequency search data enhance the performance of models used to predict tourism demand. In contrast, Jackman and Naitram (2015) find no evidence suggesting that Google Trends data add any significant information to the models. The authors use support vector regressions and Google search data to test whether Google Trends can help predict tourist arrivals in Barbados.

Business cycles play a pivotal role in tourism demand. Allowing for asymmetric income effects increases forecasting performance, as Smeral (2019) shows by considering different tourism demand elasticities varying by season across the business cycle (Smeral, 2019). Gunter et al. (2019) suggest that the number of "likes" on Facebook does not generally deliver more accurate forecasting performance. The authors claim that the models including both likes and Google Trends data typically do not outperform a simple time series model for predicting tourist arrivals to Austrian cities.

This article analyzes the eligibility of Google search data for forecasting tourist arrivals and overnight stays in Prague and reports whether weekly Google Trends data can potentially improve forecasting performance when used with MIDAS regression. First, the study investigates whether Google Trends offers significant forecasting improvements. Second, it assesses whether a higher frequency explanatory variable leads to more accurate forecasting by comparing weekly and monthly Google Trends data using MIDAS regression.

Methodology and data

Methodology

This study considers how to obtain better forecasts of tourist arrivals and overnight stays using MIDAS and aims to detect whether Google search queries can provide insight into tourism prediction for Prague tourist arrivals and overnight stays. The forecasting methodology begins with choosing a baseline model with meaningful predictive power. Then, the baseline model is run both with and without Google data to analyze whether Google can improve tourist arrival forecasting.

The MIDAS methodology was proposed by Ghysels et al. (2007) and developed by Andreou et al. (2010). Andreou et al. (2010) introduce a new decomposition of the conditional mean into two different parts: an aggregated term based on equal or flat weights and a nonlinear term that involves weighted, higher order differences of a high-frequency process. Götz et al. (2014) and Miller (2016) developed a mixed-frequency error–correction model. MIDAS was used to study tourism data by Bangwayo-Skeete and Skeete (2015) who emphasize that it offers substantial benefits to forecasters: MIDAS outperformed other methods using a data set containing monthly tourist arrivals at five destinations in the Caribbean from the United States, Canada, and the United Kingdom.

The methodology in this study follows Ghysels et al. (2007) and Andreou et al. (2010) and has been organized specifically for this study

$$\text{tourist}_t = \alpha + \sum_{i=1}^{n} \beta_i L^i \,\text{tourist}_t + \gamma \sum_{i=1}^{w} B(k;\theta) L^{k/w}\,\text{google}_t^{(w)} + \varepsilon_t^{(w)} \tag{1}$$

for $t = 1, \ldots, T$, where the function $B(k;\theta)$ is a polynomial specification that determines the weights for temporal aggregation. $L^{k/w}$ represents a lag operator, such as $L^{k/w}\text{google}_t = \text{google}_{t-k/w}^{(w)}$. In the model, tourist_t represents a low-frequency dependent variable and google_t represents a high-frequency independent variable. L is a polynomial lag operator. $\text{google}_t^{(w)}$ is observed w times in the same period (weekly, $w = 4$). β represents the effect of the lag values of tourist arrivals, and γ represents the effect of google_t search.

The parametrization of the weighting function is one of the main contributions of the MIDAS regression. Ghysels et al. (2007) propose two different parameterizations. The first is

$$B(k;\theta) = \frac{\varepsilon^{\theta_1 k + \cdots + \theta_Q k^Q}}{\sum_{k=1}^{w} \varepsilon^{\theta_1 k + \cdots + \theta_Q k^Q}} \tag{2}$$

which suggests an exponential *Almon* specification (Almon, 1965). Ghysels et al. (2006) uses the functional form (2) with two parameters ($\theta = [\theta_1; \theta_2]$). The specification gives equal weights when $\theta_1 = \theta_2 = 0$; otherwise, the weights can decline quickly or slowly based on the number of lags. The rate of decline determined by the number of lags is included in the model. The exponential function of weight can produce hump shapes, and a decreasing weight is guaranteed as long as $\theta_2 \leq 0$.

The second parameterization is a *Beta* formulation:

$$B(k;\theta_1,\theta_2) = \frac{f(k/w,\theta_1;\theta_2)}{\sum_{k=1}^{w} f(k/w,\theta_1;\theta_2)} \tag{3}$$

where

$$f(i,\theta_1;\theta_2) = \frac{i^{\theta_1-1}(1-i)^{(\theta_2-1)}\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} \tag{4}$$

θ_1 and θ_2 are hyperparameters governing the shape of the weighting function, and

$$\Gamma(\theta_p) = \int_0^{\infty} \varepsilon^{-i} i^{\theta_p-1}\,\mathrm{d}i \tag{5}$$

is the standard gamma function. The *Beta* specification also gives equal weights when $\theta_1 = \theta_2 = 0$. The rate of weight decline determines how the lags are included in the model, as in the Almon case. The weight slowly declines while $\theta_1 = 1$ and $\theta_2 > 1$. As θ_2 increases, the weight declines rapidly.

Evaluation of the quality of a forecast requires that forecast values be compared to actual values and to values from alternative models. The Diebold–Mariano (DM) test compares two forecasting models to determine whether they have equal predictive accuracy or if one model is more accurate. The DM test is described as

$$DM = \frac{\tilde{d}}{s_d} \tag{6}$$

where $\tilde{d}$ and s_d are the mean and sample standard deviation of d, respectively. d estimates

$$d = \varepsilon_1 - \varepsilon_2 \tag{7}$$

where ε_i represents either a squared or an absolute difference between the forecast and the actual values of two models $(i = 1, 2)$. We concentrate on the absolute values, defined as $\varepsilon_i = |\hat{y}_i - y_i|$, where $\hat{y}_i$ represents the forecast value and y_i represents the observed real value. The null hypothesis of the DM test is that both forecasts have the same accuracy; the alternative hypothesis is that model 2 (the Google Trends model) is more accurate than the baseline model (model without Google Trends).

Data and descriptive statistics

Monthly data of tourist arrivals and overnight stays from different countries to Prague from January 2010 to December 2016 were obtained from the Czech Statistical Office and Prague Immigration Department. Search volume histories related to the search terms “flights to Prague” and “hotels in Prague” were collected from Google Trends.

The weekly and monthly data series from Google Trends cover the same period. Google Trends measures how often a particular search term is entered relative to the total Google search volume across various countries (regions) and in various languages. Google Trends adjusts search data to make comparisons: each data point is divided by the total number of searches for the geography and time range. The resulting numbers are then scaled to a range of 0–100 based on the topic’s proportion of all searches on all topics.

Figures 1 and 2 show monthly tourist arrivals and overnight stays, respectively, with monthly Google search results. Visual inspection of the figures indicates a strong correlation between monthly tourist arrivals and overnight stays. Both time series show an upward trend and seasonal variation. Multiple methods are available for time series forecasting based on trends and seasonality. We used the natural logarithm of year-on-year growth to eliminate both linear trends and seasonal variation. We calculate the difference from the same month of the previous year; this differencing helps stabilize the mean of the tourist arrivals (overnight stays) by removing changes in the natural logarithm of the time series, thereby eliminating trends and seasonality. We performed a stationarity test for trend and seasonality in Table 4.

Figures 3 and 4 show monthly tourist arrivals and overnight stays, respectively, with weekly Google search results. Both overnight stays and weekly Google search results show an upward trend and seasonal variation, as well. Although a few outliers are observed, an overall close association is clear. These visual assessments provide support for investigating and developing models to analyze whether Google Trends can improve forecasting and the prediction of tourist arrivals to Prague.

Table 1 presents descriptive statistics for tourist arrivals and overnight stays in Prague by country of origin between January 2010 and December 2016. The table shows the top 10 countries for each group; these have a substantial impact on tourist arrivals and overnight stays in Prague, as they account for 64% of all tourist arrivals (Table 1, part A) and 62.5% of overnight stays (Table 1, part B). During this period, Germany, Russia, and the United States were the top three countries in

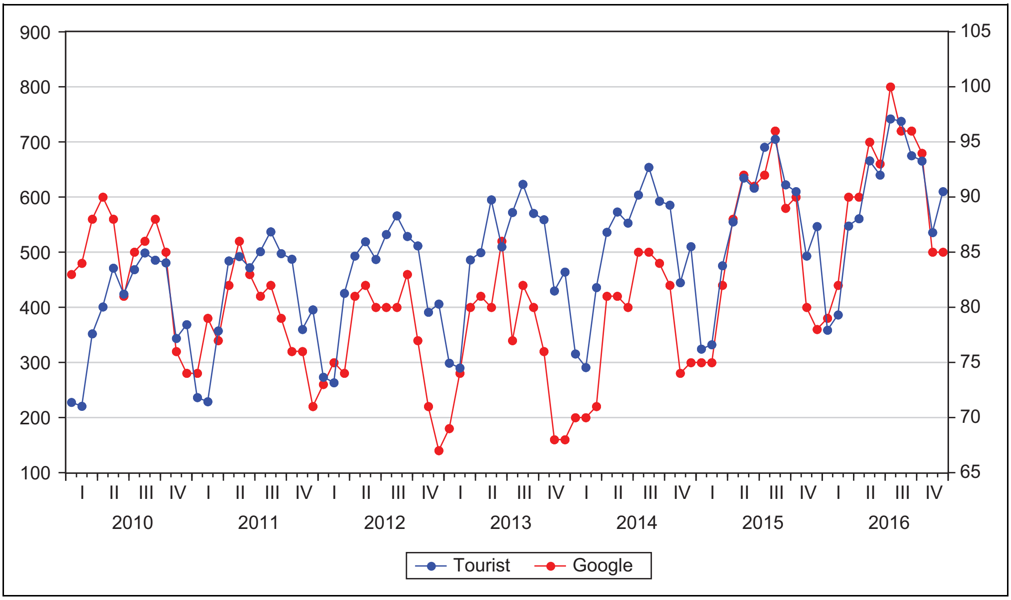

Figure 1. Monthly tourist arrivals to Prague and monthly Google searches for Prague. *Source*: Author's estimation, Google Trends and Czech Statistical Office. The left side represents the number of tourists, and the right side represents Google Trends.

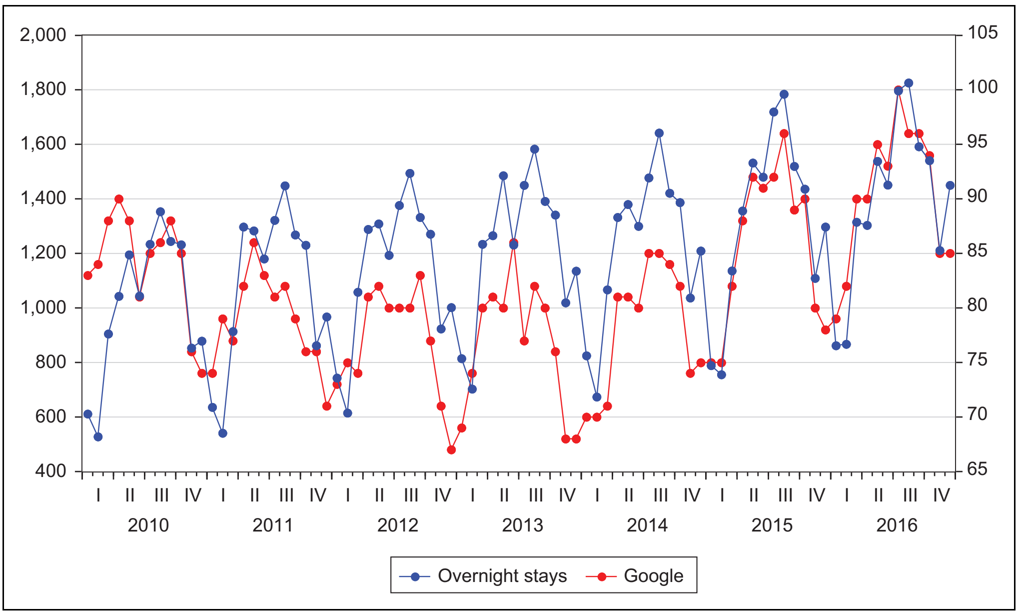

Figure 2. Monthly overnight stays in Prague and monthly Google searches for Prague. *Source:* Author's estimation, Google Trends and Czech Statistical Office. The left side represents the number of tourists, and the right side represents Google Trends.

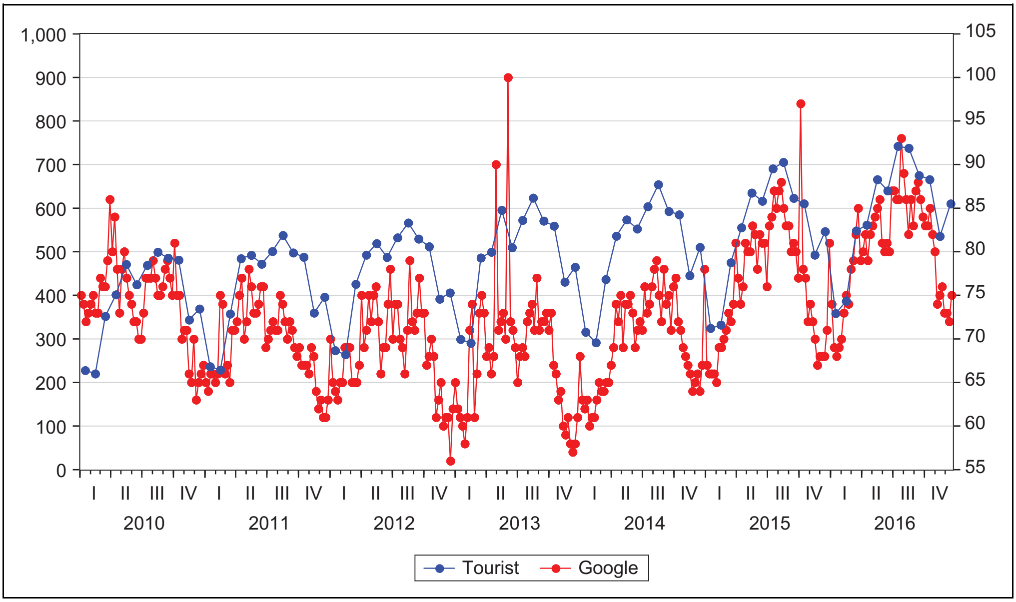

Figure 3. Monthly tourist arrivals to Prague and weekly Google searches for Prague. *Source:* Author's estimation, Google Trends and Czech Statistical Office. The left side represents the number of tourists, and the right side represents Google Trends.

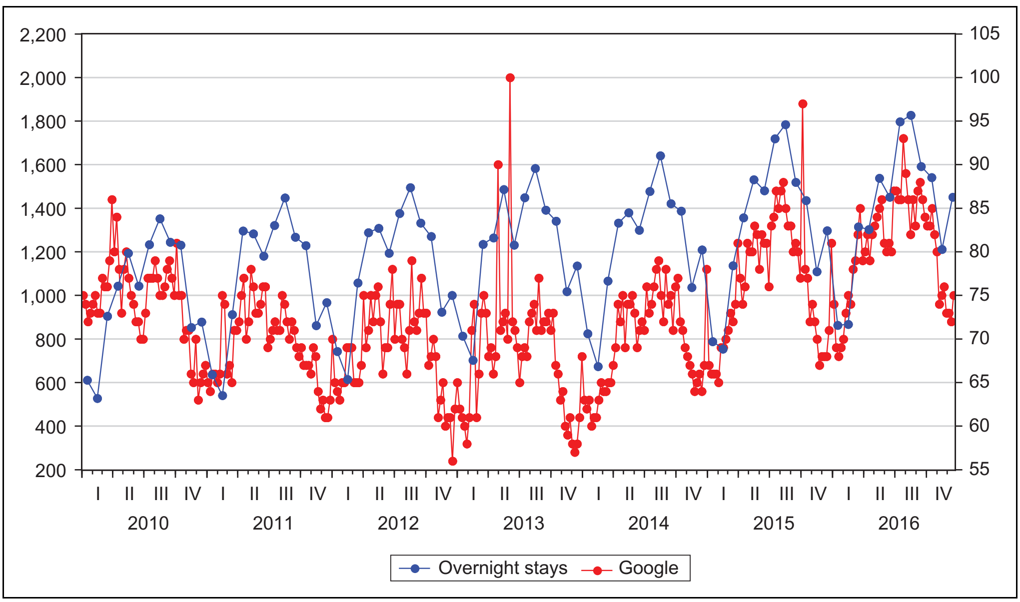

Figure 4. Monthly overnight stays in Prague and weekly Google searches for Prague. *Source:* Author's estimation, Google Trends and Czech Statistical Office. The left side represents the number of tourists, and the right side represents Google Trends.

Table 1. Descriptive analysis of monthly tourist arrivals to Prague by country.

Country	Mean	SD	Min	Max
Part A		Monthly tourist arrivals		
Germany	59,804.11	18,682.81	21,402	97,292
Russia	32,241.35	11,337.70	8966	62,742
USA	29,904.94	15,031.51	6875	61,637
UK	27,939.21	5735.94	14,377	40,716
Italy	24,400.92	9174.48	11,715	43,163
France	18,618.32	4296.31	8401	27,490
Slovakia	16,479.82	4981.66	6489	27,600
Poland	14,688.13	6105.52	4212	28,246
China	10,884.94	7149.15	1515	29,390
South Korea	9986.29	6506.87	1528	28,582
Others	175,844.80	55,457.10	68,354	308,403
Monthly total	487,152.50	125,436.20	220,329	741,900
Part B		Monthly overnight stays		
Germany	141,091.61	47,406.31	49,201	235,804
Russia	129,391.30	49,948.31	36,216	269,878
USA	73,905.63	36,767.10	16,752	150,320
UK	69,880.68	15,795.89	34,391	107,953
Italy	70,228.39	30,671.20	31,510	136,985
France	48,679.69	12,687.98	21,125	76,212
Slovakia	31,344.48	9997.55	12,033	59,799
Poland	29,115.83	12,817.48	8309	61,846
China	19,487.88	12,925.09	2834	56,167
South Korea	16,943.52	11,191.83	2978	52,099
Others	449,190.00	148,345.90	17,1762	794,039
Monthly total	1,199,376.00	304,189.60	52,8122	1,826,220

Source: Author's estimation.

both series. China and South Korea present considerable upward trends for both tourist arrivals and overnight stays in Prague.

Additionally, this study applies the augmented Dickey–Fuller (ADF) test, the Phillips–Perron (PP) test, and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test to assess the unit root hypothesis. The ADF and PP methods test the unit root hypothesis in the level values of tourist arrivals and overnight stays (Table 2, part A) and the difference value (Table 2, part B), and the KPSS method tests for stationarity in both the true and differenced values (Table 2).

In Table 2, part A, for most countries of origin, we cannot reject the null hypothesis of a unit root at the 5% level. Similar results are obtained for the KPSS test, where the null hypothesis of stationarity is rejected in most cases. When the tests are applied to the logarithmic difference of the individual time series (Table 2, part B), the null of nonstationarity is strongly rejected in most cases. For the KPSS test, we cannot reject the null hypothesis of stationarity at the 5% level for any country. These results imply that differencing is required in most cases and prove the importance of deseasonalizing and detrending tourist arrivals and overnight stays before modeling and forecasting. An adjusted MIDAS model is as follows

Table 2. Unit root tests for tourist arrivals and overnight stays.

Part A	Unit root tests for I(0)					
	Tourist arrivals			Overnight stays		
Country	ADF	PP	KPSS	ADF	PP	KPSS
Germany	1.08	−4.97	0.74	0.90	−5.08	0.56
Russia	−1.25	−4.95	0.30	−1.12	−5.59	0.33
USA	0.09	−3.77	0.56	0.15	−3.85	0.47
UK	2.13	−3.92	0.93	2.00	−4.17	0.89
Italy	−1.28	−11.70	0.36	−1.38	−13.48	0.16
France	−2.54	−6.14	0.12	−1.52	−6.83	0.06
Slovakia	1.59	−1.99	1.24	2.12	−2.46	1.19
Poland	1.71	−4.04	0.45	1.85	−4.02	0.38
China	1.13	−2.89	1.01	1.45	−2.93	0.99
South Korea	2.33	−2.27	1.06	4.56	−2.22	1.09
Others	3.20	−3.95	0.65	2.52	−4.05	0.50
Total	1.81	−3.73	0.89	0.08	−4.09	0.66
Part B	Unit root tests for I(1)					
	Tourist arrivals			Overnight stays		
Country	ADF	PP	KPSS	ADF	PP	KPSS
Germany	−4.52	−10.63	0.37	−4.16	−10.01	0.34
Russia	−2.43	−2.43	0.70	−2.46	−2.22	0.73
USA	−4.22	−4.10	0.13	−3.76	−3.76	0.14
UK	−3.79	−3.52	0.63	−2.48	−2.96	0.66
Italy	−7.28	−7.27	0.08	−7.27	−7.27	0.07
France	−2.67	−5.92	0.28	−2.77	−6.11	0.26
Slovakia	−6.33	−6.37	0.31	−5.59	−5.66	0.48
Poland	−6.83	−6.91	0.47	−6.36	−6.52	0.54
China	−4.14	−4.20	0.27	−4.31	−4.17	0.33
South Korea	−3.81	−3.84	0.73	−1.91	−3.55	0.82
Others	−7.77	−7.89	0.95	−3.72	−6.42	0.629
Total	−4.05	−7.72	0.35	−3.78	−6.90	0.09

Notes: Author's estimation. The estimation represents the monthly data for January 2010–December 2016. Tests for unit roots: ADF: augmented Dickey and Fuller (1979) test, the 5% critical value is −2.90; PP: Phillips and Perron (1988) test, the 5% critical value is −2.89. Test of stationarity: KPSS: Kwiatkowski et al. (1992) test, the 5% critical value is 0.46.

$$\Delta\log(\text{tourist}_t) = \alpha + \sum_{i=1}^{n}\beta_i L^i \Delta\log(\text{tourist}_t) + \gamma\sum_{i=1}^{w} W(k;\theta)L^{k/w}\Delta\log\left(\text{google}_t^{(w)}\right) + \varepsilon_t^{(w)} \quad (8)$$

for $t = 1, ..., T$, and $w = 1, \ldots, 4$. The dependent variable is the natural logarithm of the year-on-year change in *tourist arrivals*. The function $W(k;\theta)$ is a polynomial specification that determines the weights for temporal aggregation, such as *Beta*, *Exponential*, *or Almon*. L^i is a polynomial lag operator of *tourist arrivals*, and $L^{k/w}$ represents a lag operator of the high-frequency independent variable $\text{google}_t^{(w)}$. β represents the effect of the lag values of year-on-year change in *tourist arrivals* and γ represents the effect of the high-frequency variable $\text{google}_t^{(w)}$.

Results

The MIDAS models of tourist arrivals and overnight stays in Prague are presented in this section. Official statistical data for overnight stays and tourist arrivals were used to assess the forecasting performance of the weekly Google MIDAS regression models. All models were estimated using data from January 2010 to December 2016 and weekly Google Trends information.

Table 3 presents results for three different weighted weekly MIDAS regressions, monthly Google data, and a model without Google Trends information. The results confirm that data from 2 and 12 months ahead are significantly correlated with changes in tourist arrivals. To illustrate, tourist arrival data are monthly, while our Google Trends information is weekly. We use 8 lags (weeks) of Google Trends to explain each month of tourist arrivals. The estimation uses the 8 weeks up to and includes 3 weeks of the corresponding month. One week ahead has a significant impact on tourist arrivals. The other lags are not presented here. These results are comparable to those obtained by Bangwayo-Skeete and Skeete (2015), Siliverstovs and Wochner (2017), and Park et al. (2017), who found evidence that Google Trends information offers significant benefits for tourist forecasting.

The monthly Google regression is performed with an ARIMA(1,1,1) model. The results indicate that data from 2 months ahead of arrivals are useful for assessing the actual number of tourist arrivals. Monthly data provide valuable insights into the understanding of tourist arrivals to Prague. The results confirm that carefully identified web search activity indices, such as Google Trends information, represent early signals that can assist considerably in the prediction of tourist arrivals in Prague 2 months ahead. The results for overnight stays in Prague are similar to those for tourist arrivals (Table 4). Additionally, both 1 and 2 months ahead, Google monthly data convey useful predictive content for overnight stays. While tourist arrivals correspond to international visitors entering the country and include both tourists and same-day, nonresident visitors, overnight stays refer to the number of nights spent by nonresident tourists in accommodation establishments. Tourist arrivals represent all tourism activity, with overnight stays being particularly important for hotels and hostels.

The top three countries of origin for tourist arrivals and overnight stays were selected to ensure the robustness of the MIDAS results using weekly Google Trends information. German tourist arrivals and overnight stays in Prague present results similar to the benchmark model result (see Online Appendix, Table A1). All three country models with weekly Google Trends information performed better than their corresponding baseline models during the same prediction period (see Online Appendix). The results for Russia and the United Kingdom also indicate that data from 1 month ahead on tourist arrivals and overnight stays have a significant correlation with current inbound tourists, and the MIDAS weekly Google Trends model frameworks perform better than other baseline models (Online Appendix, Table A2, Table A3).

Next, an out-of-sample forecast evaluation was performed to assess the forecasting accuracy of each model. Thus, for all models, in-sample estimations were performed from January 2010 to May 2014, and out-of-sample forecasting was performed for June 2014 to December 2016.

The most common methods used to determine forecasting accuracy are functions of the forecasting error. The root mean squared forecast error (RMSFE) and mean absolute percentage error

Table 3. MIDAS model estimates of tourist arrivals: January 2010–December 2016.

	Weekly Google search			Monthly Google	Without Google
	Beta coeff.	Almon coeff.	Exp coeff.	ARIMA	ARIMA
DLTOURIST(−1)	0.189 (0.134)	0.253** (0.129)	0.226* (0.130)	0.532** (0.127)	0.597** (0.217)
DLTOURIST(−2)	0.212** (0.102)	0.226** (0.124)	0.223** (0.125)	0.371** (0.123)	0.353** (0.167)
DLTOURIST(−3)	−0.256 (0.133)	−0.296** (0.130)	−0.296** (0.126)	−0.267 (0.130)	−0.261 (0.186)
DLTOURIST(−12)	−0.225* (0.126)	−0.279** (0.122)	−0.258** (0.121)	−0.182 (0.121)	−0.212** (0.088)
Weekly Google	1.049** (0.447)	1.133*** (0.401)	1.090*** (0.140)		
BETA01	0.999*** (0.205)	0.004** (0.002)	0.182** (0.075)		
BETA02	19.997*** (0.004)	−0.002** (0.001)	−0.001** (0.000)		
BETA03	−0.128*** (0.007)	0.001** (0.000)	0.001* (0.001)		
Monthly Google (−1)				−0.002 (0.002)	
Monthly Google (−2)				0.003* (0.002)	
Constant	−0.403 (0.298)	−0.448 (0.275)	−0.457 (0.977)	−0.438 (0.279)	0.026* (0.014)

Notes: The dependent variable is the natural logarithm of the year-on-year change in *tourist arrivals*; the estimated equation is $\Delta\log(\text{tourist}_t) = \alpha + \sum_{i=1}^{n}\beta_i L^i \Delta\log(\text{tourist}_t) + \gamma\sum_{i=1}^{m} W(k;\theta)L^{k/m}\Delta\log\left(\text{google}_t^{(m)}\right) + \varepsilon_t^{(m)}$. Columns (2) to (4) represent weekly Google data and column (5) represents monthly Google data. Column (6) represents the ARIMA model without Google Trends information. Column (2) represents MIDAS with a beta weight function. Column (3) represents MIDAS with an exponential weight function. Column (4) represents the Almon formulation. Column (5) represents the ARIMA(1,1,1) results with monthly data. MIDAS: mixed-data sampling; ARIMA: autoregressive integrated moving average.

***Statistical significance at the 1% level.

**Statistical significance at the 5% level.

*Statistical significance at the 10% level.

Table 4. MIDAS model estimates of overnight stays: January 2010–December 2016.

	Weekly Google Search			Monthly Google ARIMA	Without Google ARIMA
	Beta coeff.	Almon coeff.	Exp coeff.		
DLTOURIST(−1)	0.267* (0.136)	0.340*** (0.125)	0.292** (0.144)	0.281** (0.127)	0.269** (0.127)
DLTOURIST(−2)	0.225* (0.115)	0.230** (0.123)	0.242** (0.122)	0.263** (0.116)	0.231* (0.123)
DLTOURIST(−3)	−0.209 (0.133)	−0.219 (0.125)	−0.237 (0.125)	−0.243** (0.121)	−0.219 (0.127)
DLTOURIST(−12)	−0.251** (0.129)	−0.314*** (0.120)	−0.282** (0.118)	−0.245** (0.114)	−0.235** (0.119)
Weekly Google	0.175** (0.087)	0.242** (0.112)	1.843** (0.951)		
BETA01	1.661* (0.671)	0.006** (0.002)	0.009* (0.004)		
BETA02	0.169* (0.077)	−0.003*** (0.001)	−0.002** (0.000)		
BETA03	−0.180* (0.085)	0.003*** (0.000)	0.009 (0.006)		
Monthly Google (−1)				−0.003** (0.002)	
Monthly Google (−2)				0.004*** (0.002)	
Constant	0.076 (0.081)	0.039 (0.075)	0.076 (0.067)	0.034 (0.073)	0.049*** (0.015)

Notes: The dependent variable is the natural logarithm of the year-on-year change in *overnight stays*; the estimated equation is $\Delta\log(\text{overnight}_t) = \alpha + \sum_{i=1}^{n}\beta_i L^i \Delta\log(\text{overnight}_t) + \gamma\sum_{i=1}^{w} W(k;\theta)L^{k/w}\Delta\log\left(\text{google}_t^{(w)}\right) + \varepsilon_t^{(w)}$. Columns (2) to (4) represent weekly Google data and column (5) represents monthly Google data. Column (2) represents MIDAS with a beta weight function. Column (3) represents MIDAS with an Almon weight function, and column (4) represents the step formulation. Column (5) represents ARIMA results for monthly data. MIDAS: mixed-data sampling; ARIMA: autoregressive integrated moving average.
***Statistical significance at the 1% level.
**Statistical significance at the 5% level.
*Statistical significance at the 10% level.

Table 5. Forecasting evaluations of MIDAS estimates of tourist arrivals and overnight stays.

	Tourist arrivals				
Part A	RMSFE	MAE	MAPE	SMAPE	Theil's U
MIDAS-Beta	15,718.42	13,011.24	58.24	36.90	0.19
MIDAS-Almon	15,077.63*	12,270.19*	55.87*	35.08*	0.18*
MIDAS-Exp.	16,142.47	13,223.80	59.43	37.09	0.19
Monthly Google	18,426.94	14,859.77	57.95	40.39	0.22
Without-Google	19,368.91	15,272.02	57.05	41.43	0.25
Mean	16,129.36	13,166.85	56.85	37.02	0.20
MSE	16,125.15	13,131.82	56.75	36.94	0.20
	Overnight stays				
Part B	RMSFE	MAE	MAPE	SMAPE	Theil's U
MIDAS-Beta	57,650.78*	—	124.11	64.41*	0.34
MIDAS-Almon	58,197.61	45,517.45	124.78	66.57	0.33*
MIDAS-Exp	59,185.03	45,020.40	123.65	63.37	0.34
Monthly Google	63,678.66	48,874.18	111.31	66.78	0.36
Without-Google	65,173.31	48,782.67	103.44*	67.88	0.40
Mean	58,850.05	44,027.4o	115.13	62.68	0.35
MSE	58,857.74	44,035.83	115.08	62.68	0.35

Notes: The MIDAS models represent weekly Google data with different weighting functions. The monthly Google model represents regressions with monthly Google data, and the without-Google model represents the result without Google Trends information. Column (2) represents the RMSFE results, column (3) represents the MAE results, column (4) represents the MAPE results, column (5) represents the symmetric MAPE results, and column(6) represents Theil's *U* statistics. MIDAS: mixed-data sampling; ARIMA: autoregressive integrated moving average; RMSFE: root mean squared forecast error; MAE: mean absolute error; MAPE: mean absolute percentage error; MSE: mean standard error.
*The most accurate forecasting model.

(MAPE) were used to assess the forecasting ability of MIDAS using weekly Google Trends data. The results are presented in Table 5. Lower MAPE and RMSE values indicate that weekly MIDAS forecasting methods offer better forecasting performance than the model with monthly Google Trends information and the model without Google Trends information. The usefulness of a forecasting model must be evaluated based on the out-of-sample forecasting performance. The results show that the MIDAS-Almon weekly Google model of tourist arrivals performs better than the other models (Table 5, part A). The MIDAS-Almon model has the lowest forecasting error as measured by all metrics—RMSFE, MAPE, and MAE. For the overnight stay results, while MIDAS-Beta has the lowest RMSFE and MAE, the model without Google Trends has a lower MAPE (Table 5, part B).

Figures 5 and 6 show the forecasting evaluations using different MIDAS regressions for tourist arrivals and overnight stays. For tourist arrivals, MIDAS-Almon is the best forecasting model (see Figure 5), whereas MIDAS-Beta is the best forecasting model for overnight stays (see Figure 6).

In the next step, we focus on the test of equal predictive accuracy by Diebold et al. (1991), which is widely used to compare two forecasting models. A modified version of the DM test purposed by Harvey et al. (1997) corrects for the tendency of the DM statistic to be oversized in

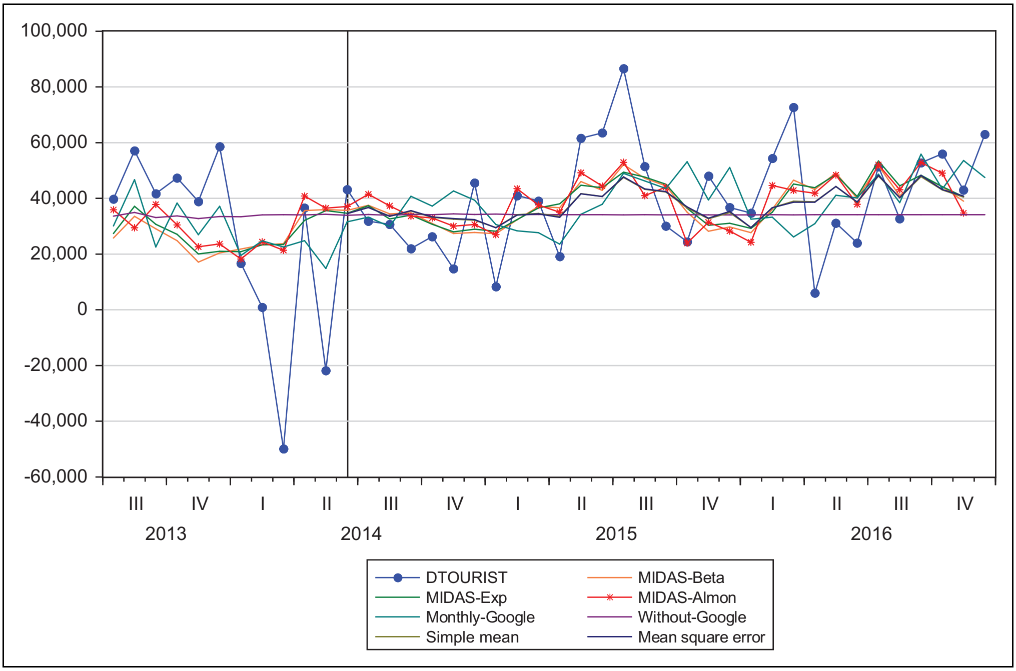

Figure 5. Forecasting tourist arrivals in Prague by MIDAS estimates: January 2012–December 2016. Lines represent the forecasting results from different models. DTOURIST represents the change in tourist arrivals. The most accurate forecasting method is MIDAS-Almon.

small samples due to bias in the estimated variance; this version was further developed by Clark and McCracken (2001) for the models in which the competing forecasts are nested due to the asymptotic distribution of the DM statistic and is not standard normal. The authors claim that with nested models, the forecasts are asymptotically the same under the null hypothesis, which leads to a nonstandard asymptotic distribution. We used the classic DM test, and the values of the test are based on the absolute values for the out-of-sample period of July 2014–December 2016. The positive significant values indicate that the MIDAS forecasting models are significantly more accurate than the competing models without Google Trends. The null hypothesis is that both forecasts have the same accuracy, but model 2 (the Google Trends model) is more accurate than the baseline model (model without Google Trends). All models reject the null hypothesis; therefore, the Google Trends models are more accurate than the baseline model (Table 6).

In summary, the model with weekly Google Trends information performed better than the models with monthly Google Trends information or the models without Google Trends information. Therefore, we can conclude that weekly Google data improve the forecasting performance for both tourist arrivals and overnight stays in Prague.

A robustness check

As a robustness check, we introduce an extended model using the consumer price index (CPI) and consumer sentiment indicator (CSI) to check whether Google Trends offers explanatory power on

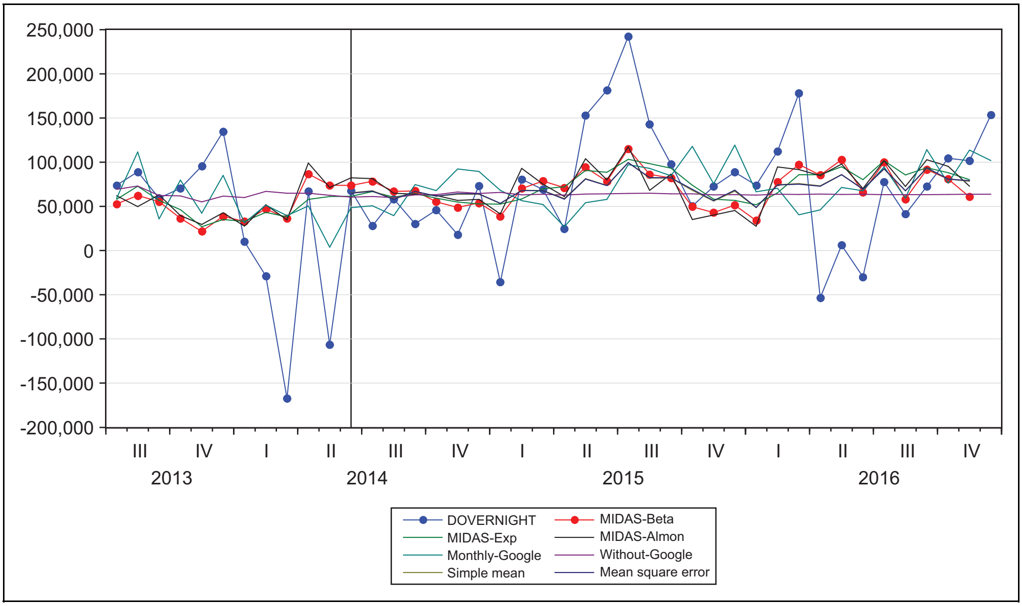

Figure 6. Forecasting overnight stays in Prague by MIDAS estimates: January 2012–December 2016. Lines represent the forecasting results from different models. DTOURIST represents the change in tourist arrivals. The most accurate forecasting method is MIDAS-Almon.

Table 6. Forecasting evaluations—Diebold and Mariano test.

Without Google Trends ARIMA	Model with Google MIDAS-Beta	MIDAS-Almon	MIDAS-Exp.	Monthly Google ARIMA
Tourist arrivals	4.61***	4.83***	4.63***	4.32***
Overnight stays	4.45***	5.01***	4.83***	4.07***

Notes: The training sample is from the period of January 2012–June 2014, and the evaluation sample is from July 2014 to December 2016. The Diebold–Mariano test compares the baseline model without Google Trends with the Google Trends models. The null hypothesis is that both forecasts have the same accuracy; the alternative hypothesis is that the Google Trends model is more accurate than the model without Google Trends information. MIDAS: mixed-data sampling; ARIMA: autoregressive integrated moving average.
***, ** and * denote statistical significance at 1%, 5% and 10% level respectively.

top of other potential explanatory variables. It is natural to expect that these two variables might influence tourism flows because they reflect the relative prices in Prague (compared to those in the home countries of potential tourists) and also the economic situation in the tourists' home countries. Both variables are available at a monthly frequency, normalized and seasonally adjusted.

Data on CPI are obtained from the Federal Reserve database. We define CPI as the weighted CPI of the top 10 originating countries of tourist arrivals to Prague relative to the Czech Republic's CPI: The variable takes the value of one when there is no difference between origin countries and Czech Republic in terms of the CPI. When the origin countries have a higher price index compared to the

Czech Republic, our CPI variables take a value above one, and they are less than one if the origin countries have a lower price index compared to the Czech Republic. CSI is defined based on consumer confidence indicators for each origin country and provides an indication of the future development of households' consumption and savings based upon answers regarding their expected financial situation and their sentiment about the general economic situation, unemployment, and ability to save, which is provided by OECD (2019) in its consumer confidence index. CSI is also estimated as a weighted indicator of the top 10 origin countries based on tourist arrivals to Prague.

The results of the extended model are consistent with our original findings and are provided in Online Appendix Table A4. We compare models with Google Trends to the model without Google Trends in Online Appendix Table A5. Significant values imply that the forecasting results when Google Trends are included provide statistically better predictions than the competing model without Google Trends. Additionally, mixed-frequency models perform better than same-frequency models. In particular, the MIDAS-Almon model performs better than all other MIDAS models, the models with monthly Google Trends, and the models without Google Trends information.

Concluding remarks

The main contribution of our study is that it combines state-of-the-art mixed-frequency modeling (MIDAS) with Google Trends data for forecasting tourist arrivals. Data on tourist arrivals are typically available at a monthly frequency, while Google Trends data are also available at a weekly frequency (i.e. they allow the practitioner to use fresher information at the time of estimation). In this article, we show that, at least in the case of Prague, one of the most frequently visited European cities, the mixed-frequency model significantly outperforms all common alternatives: a model with monthly Google Trends data and a pure time series ARIMA model without Google Trends data.

The specification of our model is simple and can be readily used by practitioners: for example, hotel chain management or municipal authorities, who need to know how many people will visit their city during a specific time period. In particular, the model only requires two time series: tourism arrivals (which are, of course, necessary for any such exercise) and search data from Google Trends (which are readily available online and free of charge). The model is also designed for real-time forecasting, as all the data are published in time for the estimation period. Extensions of our model suggest that other variables might further improve the forecasting power of the exercise, but such variables (e.g. a measure of price competitiveness related to the weighted average of the tourists' home countries and consumer sentiment indices in the tourists' home countries) are more difficult to obtain and properly construct.

Three caveats of our results are in order, and these caveats offer opportunities for future research. First, we use data for a single city, Prague. The results might not be representative for cities and regions where Google is not the primary search engine (most prominently, China). Second, the MIDAS algorithm used for mixed-frequency modeling is relatively new, and there is certainly scope for optimization to improve forecasting performance. Third, as we have noted, other explanatory variables in addition to Google Trends data, for example those related to price competitiveness and consumer sentiment, might matter for tourism arrivals, and it will be valuable to examine these influences in more detail.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Zeynalov acknowledges support from the University of Economics, Prague - IGA (grant #IG212039). Havranek acknowledges support from the Czech Science Foundation (grant #19-26812X).

ORCID iD

Ayaz Zeynalov https://orcid.org/0000-0001-5371-9797

Supplemental material

Supplemental material for this article is available online.

Note

1. The statistical data are from the Czech Statistical Office: Public database, Tourist Figures in 2016, https://www.czso.cz/csu/czso/tourism.ekon.

References

Akin M (2015) A novel approach to model selection in tourism demand modeling. *Tourism Management* 48(C): 64–72.

Almon S (1965) The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society* 33: 178–196.

Andreou E, Ghysels E and Kourtellos A (2010) Regression models with mixed sampling frequencies. *Journal of Econometrics* 158(2): 246–261.

Andreou E, Ghysels E and Kourtellos A (2012) Forecasting with mixed-frequency data. In: *The Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

Athanasopoulos G, Hyndman R J, Song H, et al. (2011) The tourism forecasting competition. *International Journal of Forecasting* 27(3): 822–844.

Bangwayo-Skeete PF and Skeete RW (2015) Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management* 46(C): 454–464.

Bokelmann B and Lessmann S (2019) Spurious patterns in Google Trends data-an analysis of the effects on tourism demand forecasting in Germany. *Tourism Management* 75: 1–12.

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

Clark TE and McCracken MW (2001) Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105(1): 85–110.

Clark M, Wilkins EJ, Dagan DT, et al. (2019) Bringing forecasting into the future: using Google to predict visitation in US national parks. *Journal of Environmental Management* 243: 88–94.

Claveria O and Torra S (2014) Forecasting tourism demand to Catalonia: neural networks vs. time series models. *Economic Modelling* 36(C): 220–228.

Dickey DA and Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74(366): 427–431.

Diebold FX and Mariano RS (1991) *Comparing Predictive Accuracy I: An Asymptotic Test*. Institute for Empirical Macroeconomics, Federal Reserve Bank of Minneapolis.

Engle RF, Ghysels E and Sohn B (2013) Stock market volatility and macroeconomic fundamentals. *The Review of Economics and Statistics* 95(3): 776–797.

Ghysels E and Wright JH (2009) Forecasting professional forecasters. *Journal of Business & Economic Statistics* 27(4): 504–516.

Ghysels E, Santa-Clara P and Valkanov R (2006) Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131(1): 59–95.

Ghysels E, Sinko A and Valkanov R (2007) Midas regressions: further results and new directions. *Econometric Reviews* 26(1): 53–90.
Girardin E and Joyeux R (2013) Macro fundamentals as a source of stock market volatility in China: a GARCH-MIDAS approach. *Economic Modelling* 34(supplement C): 59–68.
Gonzalez P and Moral P (1995) An analysis of the international tourism demand in Spain. *International Journal of Forecasting* 11(2): 233–251.
Götz TB, Hecq A and Urbain JP (2014) Forecasting mixed-frequency time series with ECM-MIDAS models. *Journal of Forecasting* 33(3): 198–213.
Gunter U, Önder I and Gindl S (2019) Exploring the predictive ability of likes of posts on the Facebook pages of four major city DMOS in Austria. *Tourism Economics* 25(3): 375–401.
Hadavandi E, Ghanbari A, Shahanaghi K, et al. (2011) Tourist arrival forecasting by evolutionary fuzzy systems. *Tourism Management* 32(5): 1196–1203.
Harvey D, Leybourne S and Newbold P (1997) Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13(2): 281–291.
Jackman M and Naitram S (2015) Research note: nowcasting tourist arrivals in Barbados—just Google it! *Tourism Economics* 21(6): 1309–1313.
Jiao EX and Chen JL (2019) Tourism forecasting: a review of methodological developments over the last decade. *Tourism Economics* 25(3): 469–492.
Kuzin V, Marcellino M and Schumacher C (2011) MIDAS vs. mixed-frequency VAR: nowcasting GDP in the euro area. *International Journal of Forecasting* 27(2): 529–542.
Kwiatkowski D, Phillips PC, Schmidt P, et al. (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54(1): 159–178.
Law R, Li G, Fong DKC, et al. (2019) Tourism demand forecasting: a deep learning approach. *Annals of Tourism Research* 75: 410–423.
Li X and Law R (2020) Forecasting tourism demand with decomposed search cycles. *Journal of Travel Research* 59(1): 52–68.
Liu H, Liu Y, Wang Y, et al. (2019) Hot topics and emerging trends in tourism forecasting research: a scientometric review. *Tourism Economics* 25(3): 448–468.
Miller JI (2016) Conditionally efficient estimation of long-run relationships using mixed-frequency time series. *Econometric Reviews* 35(6): 1142–1171.
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.
Peng B, Song H and Crouch GI (2014) A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management* 45(C): 181–193.
Phillips PCB and Perron P (1988) Testing for a unit root in time series regression. *Biometrika* 75(2): 335–346.
Rivera R (2016) A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management* 57(C): 12–20.
Siliverstovs BA and Wochner DS (2017) Google Trends and reality: Do the proportions match? Appraising the informational value of online search behavior: evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization* 145: 1–23.
Smeral E (2019) Seasonal forecasting performance considering varying income elasticities in tourism demand. *Tourism Economics* 25(3): 355–374.
Song H, Li G, Witt SF, et al. (2011) Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting* 27(3): 855–869.
Volchek K, Liu A, Song H, et al. (2019) Forecasting tourist arrivals at attractions: search engine empowered methodologies. *Tourism Economics* 25(3): 425–447.
Wen L, Liu C and Song H (2019) Forecasting tourism demand using search query data: a hybrid modelling approach. *Tourism Economics* 25(3): 309–329.

Author biographies

Tomas Havranek is an associate professor at Charles University, Prague. Between 2016 and 2019, he served as an advisor to the Board of the Czech National Bank, which is the central bank of the Czech Republic. In his research he focuses on international trade, monetary policy, and meta-analysis. His research has been published, among other journals, in the *Journal of the European Economic Association*, *Journal of International Economics*, and *Review of Economic Dynamics*. Currently he is the most cited economist of Czech origin. More information about him is available at www.meta-analysis.cz.

Ayaz Zeynalov is an assistant professor in the University of Economics in Prague. He completed his PhD in economics at IES, Charles University in Prague in April 2017. His research interests are macroeconomic analyses of economic development, natural resource economics, econometric modelling, international trade, and so on. He has published his research papers in peer-reviewed journals including World Development, Resource Policy, and International Economics and Economic Policy. He has reviewed couple manuscripts for journals including the *Economic Modeling*, *Resources Policy*, *Economic System*, *Prague Economic Papers*, and *International Journal of Applied Behavioral Economics*.