Forecasting Tourist Arrivals in Prague: Google Econometrics

Zeynalov, Ayaz

University of Economics, Prague

1 December 2017

Forecasting Tourist Arrivals in Prague: Google Econometrics

Ayaz Zeynalov*

University of Economics, Prague

Abstract

It is expected that what people are searching for today is predictive of what they have done recently or will do in the near future. This study analyzes the reliability of Google search data in predicting tourist arrivals and overnight stays in Prague. Three differently weighted weekly Mixed-data sampling (MIDAS) models, ARIMA(1,1,1) with Monthly Google Trends information and model without informative Google Trends variable have been evaluated. The main objective was to assess whether Google Trends information is useful for forecasting tourist arrivals and overnight stays in Prague, as well as whether higher frequency data (weekly data) outperform same frequency data methods. The results of the study indicate an undeniable potential that Google Trends offers more accurate forecasting, particularly for tourism. The forecasting of the indicators using weekly MIDAS-Beta for tourist arrivals and weekly MIDAS-Almon models for overnight stays outperformed monthly Google Trends using ARIMA and models without Google Trends. The results confirm that predications based on Google searches are advantageous for policy makers and business operating in the tourism sector.

Keywords: Google trends, forecasting, tourism

JEL Codes: C53; E17; L83

*Corresponding author: Jan Masaryk Centre of International Studies, Faculty of International Relations, University of Economics, Prague, W. Churchilla 4, 130 67 Prague-3, Czech Republic. e-mail: ayaz.zeynalov@vse.com

1 Introduction

People reveal useful information about their needs, wants, interests, and concerns through their internet search histories. This may be the best explanation of Google's success as it has rapidly increased publicly accessible, usable information. It is a reasonable assumption that what people are searching for today is predictive of what they have done recently or will do in the near future.

Studies have focused on search for "predicts the present" and show that search queries correlated to the contemporaneous activities (Askitas and Zimmermann, 2009; Hong, 2011; Choi and Varian, 2012, etc.). In fact, Choi and Varian (2012) show how to use search engine data for nowcasting the value of economic indicators, such as unemployment claims, automobile sales, consumer confidence and travel trends.

Several studies have discovered that Google trends data is useful as an economic indicator. Researchers have tested whether the Google Trends Automotive Index improves the fit and efficiency of nowcasting models for automobile sales in Chile (Carriere-Swallow and Labbe, 2013); demonstrates strong correlations between internet searches queries and unemployment rates in Germany (Askitas and Zimmermann, 2009); forecasts the real price of oil on the basis of macroeconomic indicators and Google search data (Fantazzini and Fomichev, 2014); uses Google Flu Trends data to describe the spread of flu in the United States during 2003-2009 (Dukic et al., 2012); whether Google queries can enhance predictions of youth unemployment in France Fondeur and Karam (2013); offers significant benefits to forecasters of private consumption indicators based on search query time series provided by Google Trends (Vosen and Schmidt, 2011); uses search query volume to forecast outcomes such as unemployment levels, auto and home sales, and disease prevalence in near real time (Goel et al., 2010); analyzes factors that influence investor information demand around earnings announcements via Google searches (Drake et al., 2012); emphasizes an approach to portfolio diversification based on popularity of a stock measured by search queries using Google Trends (Kristoufek, 2013).

Tourism forecasting has been a strong interest of many studies. Studies have adjusted indicators of the inflow of tourists with a lead of almost one month of tourist arrival using Google Trends (Artola and Galn, 2012); employed modelling and forecasting for tourist arrivals to Hong Kong from China, South Korea, the UK and the USA (Song et al., 2011); evaluated the different estimation methods of forecasting tourism data, which include 366 monthly series, 427 quarterly series and 518 annual series (Athanasopoulos et al., 2011); analyzed external demand for Spanish tourist services within the framework of Structural Time Series Models which included different types of indices (Gonzalez and Moral, 1995). Claveria and Torra (2014) proposed an artificial neural network using overnight stays and tourist arrivals from different countries to Catalonia during 2001-2009. The main objective of their study was to define which method provided more accurate information on tourist numbers. They found that the Autoregressive integrated moving average (ARIMA) models outperformed Self-Exciting Threshold AutoRegressive (SETAR) and Artificial Neural Network (ANN) models. This study tests whether Google Trends information can provided more accurate forecasting for tourist arrivals in Prague.

The use of Google trend data to predict tourism is still a subject of study. Siliverstovs and Wochner (2017) find search-based tourism predictions are highly accurate approximations of Swiss tourism demand using a MincerZarnowitz-type regression model. Bangwayo-Skeete and Skeete (2015) suggest Google trend information offers significant benefits for tourist forecasting for Caribbean destinations. Rivera (2016) proposed a Dynamic Linear Model to forecast the number of hotel visitor registrations in Puerto Rico, and found strong associations between the number of hotel visitors and data from Google Trends information. The results of Önder and Gunter (2016) confirm that the forecast accuracy is improved when Google Trends data are included across source markets and forecast horizons for seasonal and seasonally adjusted data, leaning toward native language searches using Vienna as a case example. Park et al. (2017) find that Google-augmented models perform much better than the standard time-series models in terms of short-term forecasting accuracy. In particular, Google Trends models show better out-of-sample forecasting performance than in-sample forecasting.

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 GDP and provides employment for around 17% of the working population in the service sector. Therefore, accurate forecasts of tourism volume play a major role in tourism planning as they enable destinations to predict infrastructure development needs.

Google Trends provides free, large and practically real-time information, but with some disadvantages. Firstly, Google shows only absolute data, providing an index which is relative to all searches. Secondly, internet users might type similar words even if they were looking for different topics, or different words, even if they were searching for the same topic. Thirdly, web search queries are related to personal characteristics such as education, income, age, etc. Clearly, data from Google searches is imperfect, however, based on the fact that it is one of best real-time information database: it has the potential to act as a leading indicator.

Mixed-data sampling (MIDAS) is a method of estimating and forecasting the impact of high frequency variable(s) on low frequency dependent variables, which is able to ignore the traditional requirement that variables be presented at the same frequency. MIDAS uses distributed lag of polynomials to ensure parsimonious specifications for handling series sampled at different frequencies. The MIDAS method proposed by Ghysels et al. (2006), was further developed by Andreou et al. (2010) who introducing a new decomposition for the MIDAS regression. Empirical studies in the MIDAS literature have analyzed the dynamics in microstructure noise and volatility (Ghysels et al., 2007),GDP growth forecast (Ghysels and Wright, 2009; Andreou et al., 2012) nowcasting and forecasting quarterly GDP growth in the euro area (Kuzin et al., 2011), and stock market volatility and macroeconomic activity (Engle et al., 2013; Girardin and Joyeux, 2013).

This study analyzes the eligibility of Google search data for forecasting tourist arrivals and overnight stays in Prague. The work reports whether weekly Google Trends data can potentially improve forecasting performance when used with MIDAS regressions. Firstly, the study looks at whether Google Trends offers significant forecasting improvements. Secondly, it assesses whether a higher frequency explanatory variable is better for accurate forecasting by comparing weekly and monthly Google Trends data using MIDAS regression.

The paper is organized as follows. Section 2 discusses the methodology and data sampling. Section 3 presents the empirical results on MIDAS models performed in tourist arrivals and overnight stays. Conclusion provided in Section 4. Robustness checks are available in the Appendix.

2 Methodology and Data

2.1 Methodology

This study considers how to obtain better analyses of tourist arrivals and overnight stays by using MIDAS. The empirical estimation compares different MIDAS models with and without Google Trends information as well as comparing weekly and monthly Google data. The study aims to detect whether Google search queries can add some insight into tourism prediction for both Prague tourist arrivals and overnight stays. Forecasting literature begins by choosing a baseline model for obtaining meaningful predictive power. Afterward, the baseline model will test with and without Google data to analyze whether Google can improve forecasting of tourist arrivals.

The methodology was proposed by Ghysels et al. (2007) to present monthly tourist arrivals and overnight stays using monthly and weekly Google searches. The MIDAS methodology was developed by (Andreou et al., 2010) to introduce a new decomposition of conditional mean in two different parts: an aggregated term based on equal or flat weights and a nonlinear term which, involves weighted, higher order differences of the high frequency process. MIDAS was used to study tourism data by Bangwayo-Skeete and Skeete (2015). The authors emphasize that the Google Trends information of tourists offers significant benefits to forecasters where MIDAS outperformed other methods using a dataset containing monthly tourist arrivals from US, Canada and UK to five destinations in Caribbean.

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

$$tourist_t = \alpha + \sum_{i=1}^{n} \beta_i L^i tourist_t + \gamma \sum_{i=1}^{m} B(k;\theta) L^{k/m} google_t^{(m)} + \epsilon_t^{(m)} \tag{1}$$

for $t = 1, ..., T$, where the function $B(k;\theta)$ is a polynomial specification that determines the weights for temporal aggregation. $L^{k/m}$ represents lag operator such as $L^{k/m} google_t = google_{t-k/m}(m)$. In the model, while $tourist_t$ represents a high frequency independent variable, $google_t$ represents a low frequency dependent variable. L is a polynomial lag operator. β represents the effect of lag values of tourist, γ represents the effect of $google_t$ search.

The parameterizations of the weighting function is one of main contributions of the

MIDAS regerssion. Ghysels et al. (2007) proposes two different parametrizations. The first is

$$B(k;\theta) = \frac{\epsilon^{\theta_1 k + \ldots + \theta_Q k^Q}}{\sum_{k=1}^{m} \epsilon^{\theta_1 k + \ldots + \theta_Q k^Q}} \tag{2}$$

which suggests exponential *Almon* specification (Almon, 1965). Ghysels et al. (2006) uses functional form (2) with two parameters ($\theta = [\theta_1; \theta_2]$). The specification gives equal weight when $\theta_1 = \theta_2 = 0$, otherwise the weights can decline rapidly or slowly with the numbers of lag. 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 decline weight is guaranteed as long as $\theta_2 <= 0$.

The second parameterizations is a *Beta* formulation:

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

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

$$\Gamma(\theta_p) = \int_0^\infty \epsilon^{-i} i^{\theta_p-1} di \tag{5}$$

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

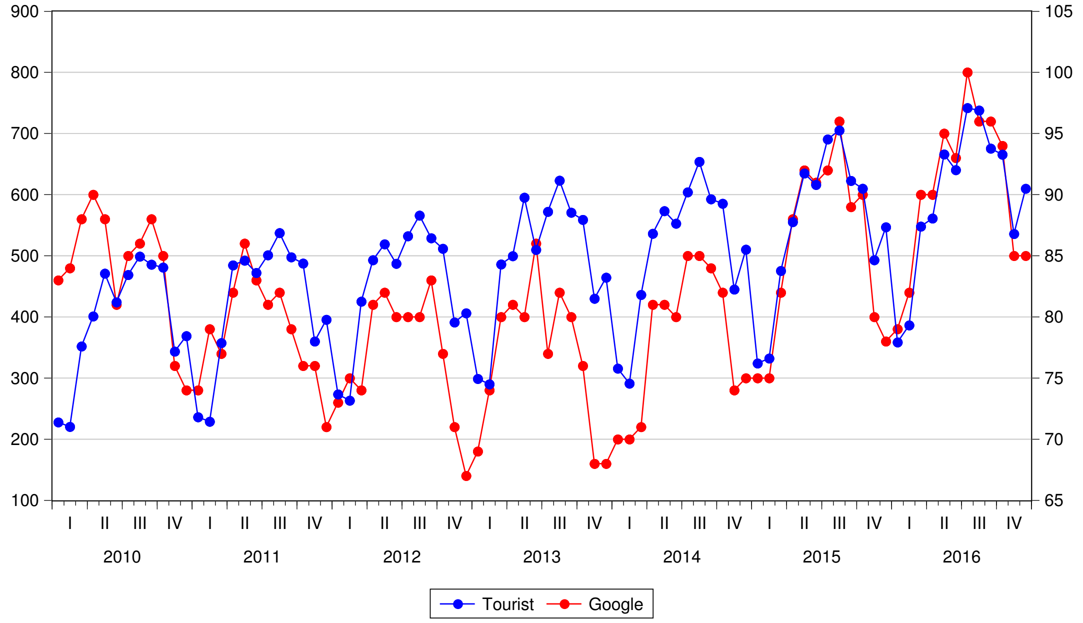
2.2 Data and descriptive statistics

Monthly data of tourist arrivals and overnight stays from different countries to Prague from January, 2010, to December, 2016, were obtained from obtained from the Prague Immigration Department. Both time series show upward trend behaviour and seasonal variations. There are multiple methods for time series forecasting based on trends as well as seasonality. Year-on-year growth has been used to eliminate both linear trends and seasonal variations.

The study also collected search volume histories related to the simple search term "flights to Prague" and "hotels in Prague" under Google Trends. Weekly and monthly data series cover the same period from Google Trends. 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. Trends adjusts search data to make comparisons: Each data point is divided by the total searches for the geography and time range, the resulting numbers are then scaled on a range of 0 to 100 based on the topic's proportion to all searches on all topics.

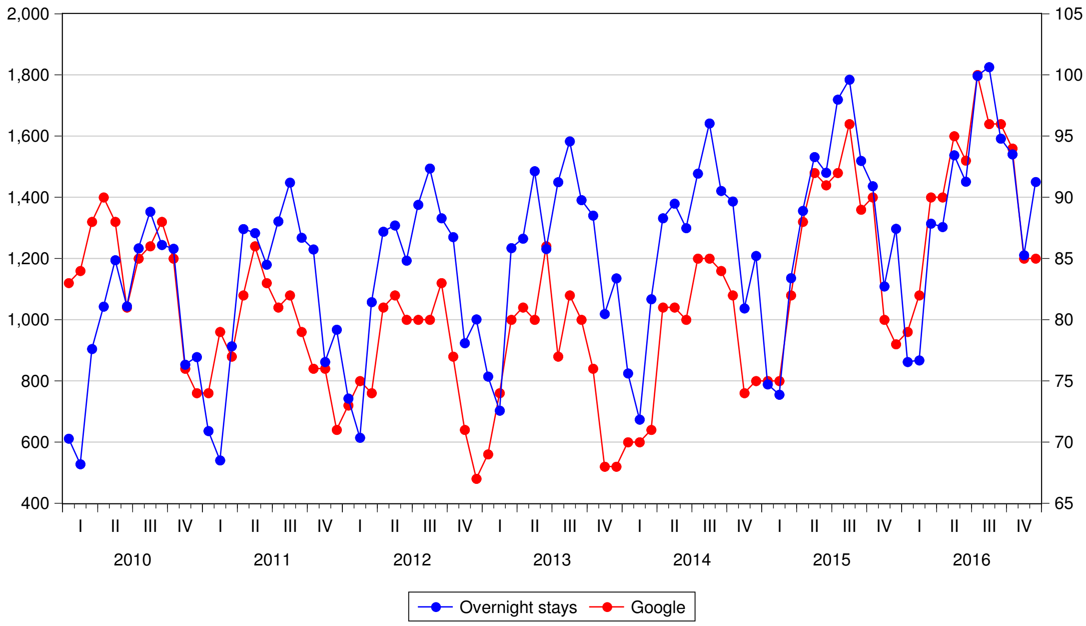
Figures 1 and 2 shows monthly tourist arrivals and overnight stays and, respectively, monthly Google search results. Visual inspection of the figures indicates a strong correlation

Figure 1: Monthly tourist arrivals to Prague and monthly Google searches for Prague



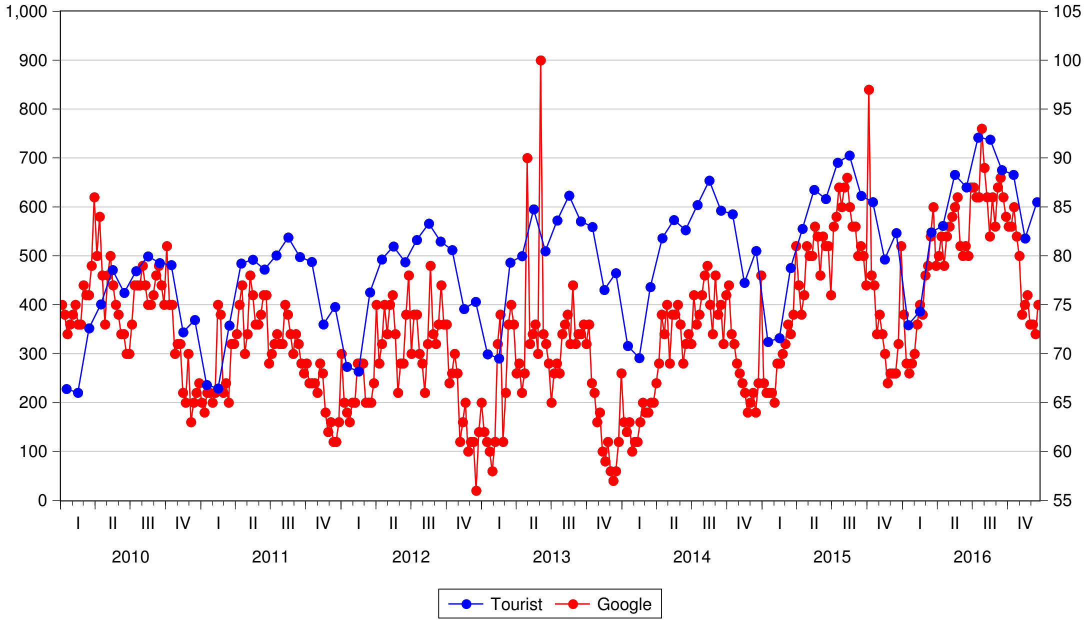
Source: Author's estimation, Google Trends and Czech Statistical Office

Figure 2: Monthly overnight stays in Prague and monthly Google searches for Prague



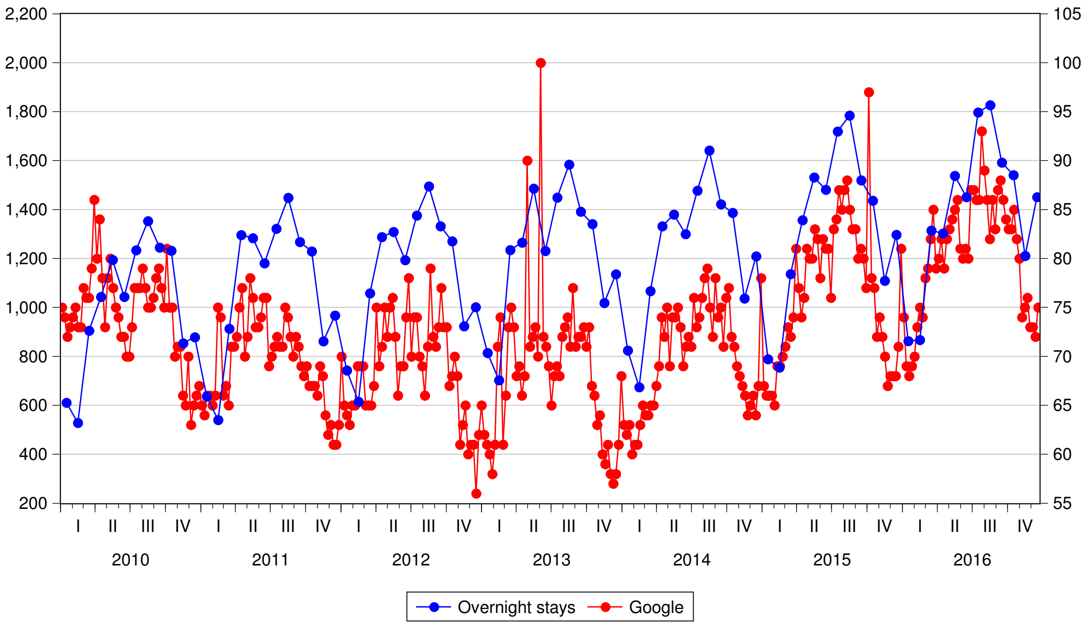
Source: Author's estimation, Google Trends and Czech Statistical Office

Figure 3: Monthly tourist arrivals to Prague and weekly Google searches for Prague



Source: Author's estimation, Google Trends and Czech Statistical Office

Figure 4: Monthly overnight stays in Prague and weekly Google searches for Prague



Source: Author's estimation, Google Trends and Czech Statistical Office between monthly tourist arrivals and overnight stays. Figures 3 and 4 shows monthly tourist arrivals and overnight stays and, respectively, weekly Google search results. Although there are a couple of outliers, on the whole a close association is clearly visible. These visual assessments are encouraged for investigative and develop modelling for analyze whether Google Trends can improve forecasting and prediction of tourist arrivals to Prague.

Tables 1 and 2 represents descriptive statistics of tourist arrivals and overnight stays in Prague by countries of origin between January 2010 and December 2016. The tables represent the top ten countries, which have a significant impact on tourist arrivals and overnight stays in Prague. They accounting for 64% of total tourist arrivals (Table 1) and 62.5% of overnight stays (Table 2). During this period, Germany, Russia and the USA are the top three countries for both series. China and South Korea presents significant upward trends for both tourist arrivals and overnight stays in Prague.

Table 1: Descriptive analysis of monthly tourist arrivals by countries

Country	**Mean**	**SD**	**Min**	**Max**
Monthly total	487152.50	125436.20	220329	741900
Germany	59804.11	18682.81	21402	97292
Russia	32241.35	11337.70	8966	62742
USA	29904.94	15031.51	6875	61637
UK	27939.21	5735.94	14377	40716
Italy	24400.92	9174.48	11715	43163
France	18618.32	4296.31	8401	27490
Slovakia	16479.82	4981.66	6489	27600
Poland	14688.13	6105.52	4212	28246
China	10884.94	7149.15	1515	29390
South Korea	9986.29	6506.87	1528	28582
Others	175844.80	55457.10	68354	308403

Source: Author's estimation.

Table 2: Descriptive analysis of monthly overnight stays by countries

Country	**Mean**	**SD**	**Min**	**Max**
Monthly total	1199376.00	304189.60	528122	1826220
Germany	141091.61	47406.31	49201	235804
Russia	129391.30	49948.31	36216	269878
USA	73905.63	36767.10	16752	150320
UK	69880.68	15795.89	34391	107953
Italy	70228.39	30671.20	31510	136985
France	48679.69	12687.98	21125	76212
Slovakia	31344.48	9997.55	12033	59799
Poland	29115.83	12817.48	8309	61846
China	19487.88	12925.09	2834	56167
South Korea	16943.52	11191.83	2978	52099
Others	449190.00	148345.90	171762	794039

Source: Author's estimation.

Additionally, this study applies an augmented Dickey-Fuller (ADF) test, the Phillips-Perron (PP) test, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test computed to test unit root hypothesis. While ADF and PP test unit root hypothesis in true value tourist arrivals and overnight stays (Table 3), and difference value (Table 4), KPSS tests stationarity in both true and differenced values (Tale 3 and 4). As in Table 3, 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 were applied to the first difference of individual time series (Table 4), the null of nonstationarity is strongly rejected in most cases. In the case of 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 de-seasonalizing and de-trending tourist arrrivals and overnight stays before modelling and forecasting.

Table 3: Unit root tests in tourist arrivals and overnight stays - test for I(0)

	Tourist arrivals			**Overnight stays**		
Country	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**
Total	1.81	-3.73	0.89	0.08	-4.09	0.66
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

Source: Author's estimation. Estimation represents 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.

Table 4: Unit root tests in tourist arrivals and overnight stays - test for I(1)

	Tourist arrivals			**Overnight stays**		
Country	**ADF**	**PP**	**KPSS**	**ADF**	**PP**	**KPSS**
Total	-4.05	-7.72	0.35	-3.78	-6.90	0.09
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

Notes: Author's estimation. Estimation represents 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.

3 Results

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

Table 5 represents results for 3 different weighted weekly MIDAS regressions, monthly Google data, and a model without Google trends information. The results confirms that two and twelve months ahead are significantly correlated with changes in tourist arrivals. To illustrate, tourist arrivals data is 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 including, the three weeks of the corresponding month. One week ahead had significant impact on tourist arrivals, other lags have not been represented 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 performance.

Table 5: MIDAS models estimates in tourist arrivals: January, 2010 - December, 2016

	Weekly Google Search			**Monthly Google**	**Without Google**
	Beta coeff	**Exp coeff**	**Almon coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	0.066	0.042	0.135	0.079	0.114
	(0.142)	(0.139)	(0.147)	(0.127)	(0.133)
DTOURIST(-2)	0.280**	0.269**	0.262**	0.214*	0.335**
	(0.137)	(0.124)	(0.134)	(0.123)	(0.126)
DTOURIST(-3)	-0.148	-0.156	-0.160	-0.252*	-0.132
	(0.139)	(0.130)	(0.137)	(0.127)	(0.132)
DTOURIST(-12)	-0.270**	-0.276**	-0.289**	-0.252**	-0.169
	(0.129)	(0.122)	(0.130)	(0.116)	(0.122)
Weekly Google	1.049**	1.133***	1.090***		
	(0.447)	(0.401)	(0.140)		
BETA01	1.076***	-1.720	1.825**		
	(0.081)	(4.172)	(0.758)		
BETA02	20.000***	0.000	-0.808**		
	(0.002)	(0.847)	(0.379)		
BETA03	-0.037		0.074*		
	(0.086)		(0.037)		
Monthly Google (-1)				-0.738	
				(0.622)	
Monthly Google (-2)				1.783***	
				(0.632)	
CONSTANT	-40.372	-44.841	-45.708	-43.857	29.039***
	(29.899)	(27.568)	3(0.977)	(27.759)	(8.037)

Notes: The dependent variable is *tourist arrivals*; the estimated equation is $tourist_t = \alpha + \sum_{i=1}^{n} \beta_i L^i tourist_t + \gamma \sum_{i=1}^{m} W(k;\theta) L^{k/m} google_t^{(m)} + \epsilon_t^{(m)}$. While Column(2)-(4) represent weekly Google data, Column(5) represents monthly Google data. Column(6) represent ARIMA model without Google trends information. Column (2) represents MIDAS with the weight function of beta formualtion. Column (3) represents MIDAS with the weight function of Exponential formulation, Column (4) represents Almon formulation. Column (5) represents ARIMA(1,1,1) results with monthly data. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Monthly Google regression composed by ARIMA(1,1,1). The results indicate that data from two months ahead of arrivals is useful for assessing actual numbers of tourist arrivals. Monthly data offers valuable insights into the understanding of tourist arrivals to Prague. It confirms carefully identified web search activity indices like Google Trends information encompass early signals that can significantly assist in to the prediction of tourists arrivals in Prague two months ahead.

Overnight stays in Prague results are similar to tourist arrivals. Additionally, both one and two months ahead, Google monthly data turns out to convey a useful predictive content for the overnight stays. While tourist arrivals correspond to international visitors entering the country and include both tourists and same-day, non-resident visitors, overnight stays refers to the number of nights spent by non-resident tourists in accommodation establishments. Tourist arrivals concern all tourism activity, with overnight being particularly important for hotels and hostels.

Table 6: MIDAS models estimates in overnight stays: January, 2010 - December, 2016

	Weekly Google Search			**Monthly Google**	**Without Google**
	Beta coeff	**Exp coeff**	**Almon coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	0.175	0.140	0.233	0.186	0.181
	(0.128)	(0.128)	(0.144)	(0.121)	(0.128)
DTOURIST(-2)	0.319**	0.306**	0.321***	0.298**	0.335***
	(0.122)	(0.123)	(0.130)	(0.116)	(0.122)
DTOURIST(-3)	-0.162	-0.177	-0.181	-0.262**	-0.189
	(0.125)	(0.125)	(0.133)	(0.121)	(0.127)
DTOURIST(-12)	-0.333***	-0.318**	-0.323***	-0.289**	-0.254**
	(0.119)	(0.120)	(0.125)	(0.111)	(0.117)
Weekly Google	1.759	2.422**	1.843**		
	(1.179)	(1.124)	(0.951)		
BETA01	1.020***	27.609	6.030***		
	(0.067)	(29.423)	(2.220)		
BETA02	3.265	-9.458	-2.779**		
	(5.047)	(14.475)	(1.108)		
BETA03	-0.140***		0.256**		
	(0.035)		(0.108)		
Monthly Google (-1)				-3.337*	
				(1.793)	
Monthly Google (-2)				5.243***	
				(1.797)	
CONSTANT	-64.386	-110.416	-95.934	-86.803	59.022***
	(84.302)	(80.458)	(90.041)	(79.266)	(15.802)

Notes: The dependent variable is *overnight stays*; the estimated equation is $overnight_t = \alpha + \sum_{i=1}^{n} \beta_i L^i overnight_t + \gamma \sum_{i=1}^{m} W(k;\theta) L^{k/m} google_t^{(m)} + \epsilon_t^{(m)}$. While Column(2)-(4) represent weekly Google data, Column(5) represents monthly Google data. Column (2) represents MIDAS with the weight function of beta formualtion. Column (3) represents MIDAS with the weight function of Almon formulation, Column (4) represents Step formulation. Column (5) represents ARIMA results monthly data. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

In order to ensure the robustness of the MIDAS results using weekly Google trends information, the top three countries of origin for tourist arrivals and overnight stays have been selected. German tourist arrivals and overnight stays in Prague present similar results to the benchmark model result (see Appendix, Table A1). All three country models with weekly Google Trends information performed better than their corresponding baseline models at the same prediction period (see Appendix). Results from Russia and, the UK also indicate that data from one month ahead on tourist arrivals and overnight stays have significant correlation with current tourist inbound, and MIDAS weekly Google Trends model frameworks have more favourable performances than other baseline models (Table A2,Table A3).

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

The most common methods to determine forecasting accuracy are functions of forecasting error. To assess the forecasting ability of MIDAS using weekly Google Trends data Root Mean Squared Forecast Error (RMSFE) and Mean Absolute Percentage Error (MAPE) have been performed. The results are shown in Table 7. Lower MAPE and RMSE means MIDAS forecasting methods offer better forecasting performance compared to both model with monthly Google and model without Google Trends information. Thus, the usefulness of a forecasting model must be evaluated by the performance of out-of-sample forecasting. The results show that the MIDAS-Almon weekly Google model of tourist arrivals performs better than other models (Part A, Table 7). MIDAS-Almon model has lower forecasting error by all tests - RMSFE, MAPE, MAE. For overnight stay results, while MIDAS-Beta has lower RMSFE and MAE, model without Google Trends has lower value in MAPE (Part B, Table 7).

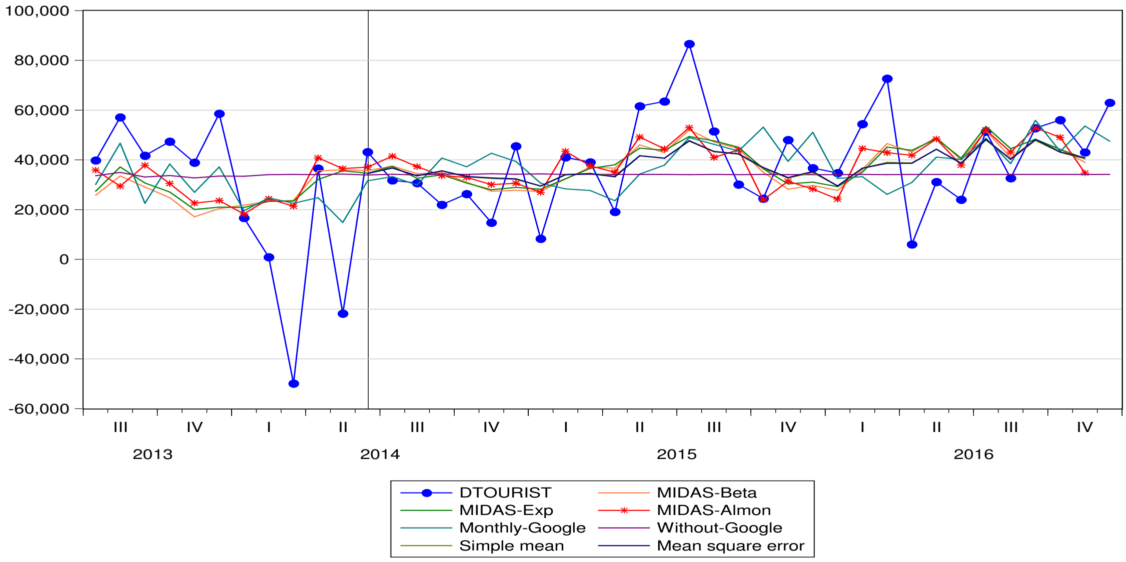
Table 7: Forecasting Evaluations of MIDAS estimates in tourist arrivals and overnight stays

			Tourist Arrivals		
Part A	RMSFE	MAE	MAPE	SMAPE	Theil's U
MIDAS-Beta	15718.42	13011.24	58.24	36.90	0.19
MIDAS-Exp	16142.47	13223.80	59.43	37.09	0.19
MIDAS-Almon	15077.63*	12270.19*	55.87*	35.08*	0.18*
Monthly-Google	18426.94	14859.77	57.95	40.39	0.22
Without-Google	19368.91	15272.02	57.05	41.43	0.25
Mean	16129.36	13166.85	56.85	37.02	0.20
MSE	16125.15	13131.82	56.75	36.94	0.20
			Overnight stays		
Part B	RMSFE	MAE	MAPE	SMAPE	Theil's U
MIDAS-Beta	57650.78*	44641.69*	124.11	64.41*	0.34
MIDAS-Exp	59185.03	45020.40	123.65	63.37	0.34
MIDAS-Almon	58197.61	45517.45	124.78	66.57	0.33*
Monthly-Google	63678.66	48874.18	111.31	66.78	0.36
Without-Google	65173.31	48782.67	103.44*	67.88	0.40
Mean	58850.05	44027.4o	115.13	62.68	0.35
MSE	58857.74	44035.83	115.08	62.68	0.35

Notes: MIDAS models represent weekly Google data with different weighting functions. While Monthly Google model represents regressions with monthly Google data, last model repsrest result without Google trends information. While Column (2) represents results from Root Mean Squared Forecast Error (RMSFE), Column(3) represents Mean Absolute Error (MAE) represnt, Column(4) represents Mean Absolute Percentage Errors (MAPE), Column (5)Symmetric MAPE, Column(6) represents Theil's U Statistics. MSE represents Mean Standard errors. * denotes best accurate forecasting models.

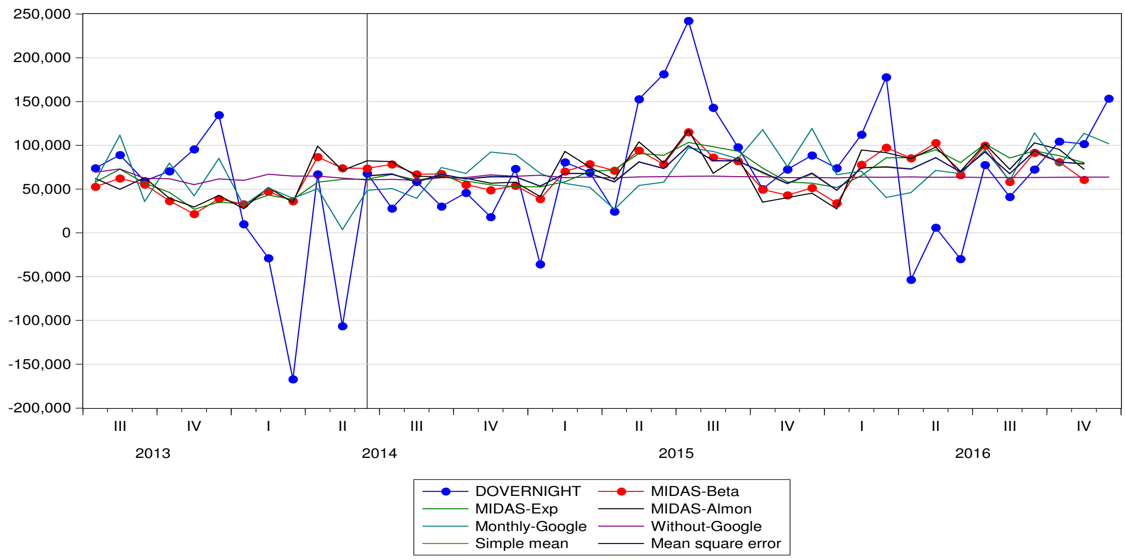
Figures 5 and 6 show forecasting evaluations for tourist arrivals and overnight stays using different MIDAS regressions. For tourist arrivals MIDAS-Almon is the best fitted forecasting

Figure 5: Forecasting tourist arrivals in Prague by MIDAS estimates: Jan, 2012 - Dec, 2016



Notes: Lines represent forecasting results from different models. DTOURIST represents change of tourist arrivals using a blue line. The most accurate forecasting method (MIDAS-Almon) represents with is represented by a red line.

Figure 6: Forecasting overnight stays in Prague by MIDAS estimates: Jan, 2012 - Dec, 2016



Notes: Lines represent forecasting results from different models. DOVERNIGHT represents change on overnight stays using a blue line. The most accurate forecasting method (MIDAS-Beta) is represented by a red line.

model (Figure 5). When analysing the forecast performance for overnight stays, MIDAS-Beta is the best fitted forecasting (Figure 6).

In summary, the comparison forecast performance of several time-series models with weekly and monthly Google augmented models, and model without Google Trends, model for inbound tourism demand in Prague confirm that weekly Google augmented models performed much better forecasting performance than monthly Google and models without Google trends information. Therefore, we can conclude that weekly Google data increasing forecasting performance for both tourist arrivals and overnight stays inbound of Prague tourism demand.

4 Concluding Remarks

The main objective of this study is perform accurate nowcasting and forecasting of tourist arrivals and overnight stays in Prague. The accurate forecasting of tourism trends is important due to the rapidly growing impact of global tourism. Internet searches play an increasingly important role in tourism and on assessing tourism consumption dynamics. This has inspired my evaluation of the performance of Google Trends searches on Prague tourist arrivals and overnight stays using MIDAS, ignoring same frequency assumptions.

Three different weigthed MIDAS models using weekly data, ARIMA(1,1,1) with Monthly Google Trends information and a model without informative variable have been evaluated. The main objective was to assess whether Google Trends information brings significant benefits to the evaluation and forecasting of tourist arrivals and overnight stays in Prague, as well as, whether higher frequency data (weekly data) outperform same frequency data methods.

The results show an undeniable potential for Google Trends to improve evaluation and forecasting in tourism. MIDAS allowed evaluation of different frequencies series like weekly Google Trends information and monthly tourist data. The forecasting performance of the indicators using weekly MIDAS-Beta for tourist arrivals and weekly MIDAS-Almon for overnight stays outperformed monthly Google trends using ARIMA and a model without Google trends. The results confirms that using data from Google searches enriches information available for policy makers and business entrepreneurs operating in the tourism sector. The accurate forecasting of tourist arrivals and overnight stays plays a vital role due to their enormous impact on economic growth in tourism-dependent destinations.

The MIDAS approach has only recently been introduced and is still in the development stage. A challenging question to be considered in future research is whether optimizing MIDAS may improve forecasting performance for different frequencied data series.

References

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 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*.

Artola, C. and Galn, E. (2012). Tracking the future on the web: construction of leading indicators using internet searches. Banco de Espania Occasional Papers 1203, Banco de Espania.

Askitas, N. and Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly (formerly: Konjunkturpolitik), Duncker & Humblot, Berlin*, 55(2):107–120.

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

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(C):454 – 464.

Carriere-Swallow, Y. and Labbe, F. (2013). Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting*, 32(4):289–298.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *The Economic Record*, 88(1):2–9.

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

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431.

Drake, M. S., Roulstone, D. T., and Thornock, J. R. (2012). Investor Information Demand: Evidence from Google Searches Around Earnings Announcements. *Journal of Accounting Research*, 50(4):1001–1040.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking Epidemics With Google Flu Trends Data and a State-Space SEIR Model. *Journal of the American Statistical Association*, 107(500):1410–1426.

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

Fantazzini, D. and Fomichev, N. (2014). Forecasting the real price of oil using online search data. *International Journal of Computational Economics and Econometrics*, 4(1/2):4–31.

Fondeur, Y. and Karam, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30(C):117–125.

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.

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

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.

Goel, S., Hofman, J., Lehaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):486490.

Gonzalez, P. and Moral, P. (1995). An analysis of the international tourism demand in Spain. *International Journal of Forecasting*, 11(2):233–251.

Hong, W.-C. (2011). Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm. *Energy*, 36(9):5568–5578.

Kristoufek, L. (2013). Can Google Trends search queries contribute to risk diversification? *Scietific Reports*, (2713).

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, P. C., Schmidt, P., and Shin, Y. (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.

Önder, I. and Gunter, U. (2016). Forecasting tourism demand with google trends for a major european city destination. *Tourism Analysis*, 21(2-3):203–220.

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.

Phillips, P. C. B. 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(Supplement C):12 – 20.

Siliverstovs, B. A. and Wochner, D. S. (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*.

Song, H., Li, G., Witt, S. F., and Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting*, 27(3):855–869.

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

Appendix

Table A1: MIDAS models estimates in tourism inbound from GERMANY to PRAGUE

	Weekly Google Search			**Monthly Google**	**Without Google**
Tourist arrivals	**Beta coeff**	**Almon coeff**	**Step coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	-0.151	-0.216	-0.117	-0.161	-0.167
	(0.129)	(0.131)	(0.131)	(0.144)	(0.132)
DTOURIST(-2)	0.342**	0.350**	0.324**	0.405***	0.453***
	(0.124)	(0.133)	(0.125)	(0.135)	(0.122)
DTOURIST(-3)	0.137	0.127	0.108	0.155	0.180
	(0.127)	(0.135)	(0.129)	(0.138)	(0.132)
DTOURIST(-12)	-0.355**	-0.275**	-0.344***	-0.280**	-0.254**
	(0.115)	(0.118)	(0.115)	(0.124)	(0.120)
Weekly Google	144.374*	151.835**	89.874***		
	(72.546)	(69.269)	(26.944)		
BETA01	0.977***	-20.274	-54.677**		
	(0.042)	(36.201)	(27.228)		
BETA02	3.107	0.001	-7.410		
	(3.142)	(0.002)	(7.909)		
BETA03	-0.080***				
	(0.024)				
Monthly Google (-1)				-23.571	
				(110.468)	
Monthly Google (-2)				72.660	
				(107.626)	
CONSTANT	-5164.326	-5363.140	-4257.535	-4142.889	2988.717
	(4172.183)	(4022.456)	(4177.083)	(4395.591)	(1215.660)

	Weekly Google Search			**Monthly Google**	**Without Google**
Overnight Stays	**Beta coeff**	**Almon coeff**	**Step coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	-0.082	-0.134	-0.182	-0.125	-0.117
	(0.140)	(0.121)	(0.136)	(0.138)	(0.129)
DTOURIST(-2)	0.386***	0.396***	0.488***	0.454***	0.500***
	(0.135)	(0.117)	(0.127)	(0.128)	(0.116)
DTOURIST(-3)	0.142	0.103	0.126	0.164	0.188
	(0.135)	0(.125)	(0.140)	(0.134)	(0.129)
DTOURIST(-12)	-0.358***	-0.393***	52.723**	-0.327**	-0.303**
	(0.124)	(0.116)	(86.212)	(0.123)	(0.119)
Weekly Google	266.271	501.238***	77.588		
	(199.440)	(155.273)	(118.454)		
BETA01	-0.535	-25.017	-297.355		
	2.757	55.769	407.039		
BETA02	-0.507	0.012	-24.453		
	2.719	13.651	47.032		
BETA03	-0.084				
	0.036				
Monthly Google (-1)				-251.534	
				(281.237)	
Monthly Google (-2)				153.515	
				(273.631)	
CONSTANT	-8859.927***	-23079.250***	-6735.677	-4079.070	5863.350**
	(1669.720)	(9398.583)	(11721.840)	(11461.080)	(2723.377)

Notes: The dependent variables are *tourist arrivals* and *overnight stays*; While Column(2)-(4) represent weekly Google data, Column(5) represents monthly Google data. Column(6) represent ARIMA model without Google trends information. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A2: MIDAS models estimates in tourism inbound from RUSSIA to PRAGUE

	Weekly Google Search			Monthly Google	Without Google
Tourist arrivals	**Beta coeff**	**Almon coeff**	**Step coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	0.420***	0.456***	0.453***	0.373***	0.583***
	(0.131)	(0.133)	(0.128)	(0.125)	(0.129)
DTOURIST(-2)	0.019	0.122	0.027	0.137	0.196
	(0.151)	(0.148)	(0.144)	(0.133)	(0.150)
DTOURIST(-3)	0.190	0.188	0.181	0.136	0.233*
	(0.127)	(0.134)	(0.124)	(0.122)	(0.137)
DRTOURIST(-12)	-0.422***	-0.322***	-0.406***	-0.439***	-0.199**
	(0.100)	(0.096)	(0.098)	(0.096)	(0.085)
Weekly Google	468.134***	284.092***	105.806***		
	(123.086)	(102.596)	(37.937)		
BETA01	-0.438	-2.021	-47.214		
	(24.451)	(4.196)	(49.837)		
BETA02	-0.413	0.000	133.611**		
	(24.451)	(0.001)	(64.874)		
BETA03	0.028				
	(0.077)				
Monthly Google (-1)				483.969**	
				(184.484)	
Monthly Google (-2)				284.899	
				(183.975)	
CONSTANT	-23681.190***	-14408.300***	-22514.870***	-20197.410***	87.338
	(6319.816)	(5284.226)	(6118.155)	(4837.935)	(711.263)
	Weekly Google Search			**Monthly Google**	**Without Google**
Overnight Stays	**Beta coeff**	**Almon coeff**	**Step coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	0.466***	0.485***	0.495***	0.392***	0.623***
	(0.138)	(0.135)	(0.134)	(0.129)	(0.132)
DTOURIST(-2)	0.168	0.166	0.062	0.168	0.230
	(0.152)	(0.154)	(0.150)	(0.136)	(0.155)
DTOURIST(-3)	0.086	0.087	0.093	0.051	0.118
	(0.144)	(0.140)	(0.129)	(0.125)	(0.142)
DRTOURIST(-12)	-0.280**	-0.271***	-0.343***	-0.412***	-0.152*
	(0.122)	(0.095)	(0.098)	(0.099)	(0.087)
Weekly Google	1269.535**	1201.694***	417.137**		
	(601.374)	(413.059)	(165.583)		
BETA01	1.000***	26.993	-210.580		
	(0.255)	(99.930)	(220.677)		
BETA02	20.000***	-9.139	572.430**		
	(0.006)	(32.644)	(285.244)		
BETA03	0.198				
	(1.727)				
Monthly Google (-1)				1976.908**	
				(791.296)	
Monthly Google (-2)				1395.206*	
				(810.783)	
CONSTANT	-65890.280**	-62140.300***	-91117.800***	-90139.430***	-1640.273
	(32272.940)	(21442.330)	(25937.850)	(21083.430)	(3154.173)

Notes: The dependent variables are *tourist arrivals* and *overnight stays*; the estimated equation is $tourist_t = \alpha + \sum_{i=1}^{n} \beta_i L^i tourist_t + \gamma \sum_{i=1}^{m} W(k;\theta) L^{k/m} google_t^{(m)} + \epsilon_t^{(m)}$. While Column(2)-(4) represent weekly Google data, Column(5) represents monthly Google data. Column(6) represent ARIMA model without Google trends information. Column (2) represents MIDAS with the weight function of beta formualtion. Column (3) represents MIDAS with the weight function of Exponential formulation, Column (4) represents Almon formulation. Column (5) represents ARIMA(1,1,1) results with monthly data. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A3: MIDAS models estimates in tourism inbound from UK to PRAGUE

	Weekly Google Search			Monthly Google	Without Google
Tourist arrivals	**Beta coeff**	**Almon coeff**	**Step coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	0.231*	0.215	0.226*	0.207	0.266**
	(0.137)	(0.136)	(0.134)	(0.132)	(0.132)
DTOURIST(-2)	-0.080	-0.063	-0.060	-0.063	-0.025
	(0.146)	(0.140)	(0.144)	(0.134)	(0.137)
DTOURIST(-3)	0.101	0.141	0.095	0.096	0.161
	(0.116)	(0.112)	(0.116)	(0.112)	(0.110)
DRTOURIST(-12)	-0.161*	-0.166*	-0.163*	-0.180**	-0.086
	(0.090)	(0.088)	(0.090)	(0.088)	(0.079)
Weekly Google	95.894*	80.856*	7.385		
	(51.801)	(41.438)	(15.318)		
BETA01	3.246	-23.578	23.063		
	(3.065)	(16.098)	(17.742)		
BETA02	3.885	0.000	-35.266		
	(3.899)	(0.001)	(75.574)		
BETA03	-0.185**				
	(0.076)				
Monthly Google (-1)				41.498	
				(36.177)	
Monthly Google (-2)				23.888	
				(35.955)	
CONSTANT	-2714.248	-2009.320	-2296.525	-2714.019	1469.235***
	(2275.306)	(1827.674)	(2284.350)	(1971.266)	(436.050)
	Weekly Google Search			**Monthly Google**	**Without Google**
Overnight Stays	**Beta coeff**	**Almon coeff**	**Step coeff**	**ARIMA**	**ARIMA**
DTOURIST(-1)	0.336**	0.306**	0.332**	0.319**	0.347***
	(0.137)	(0.137)	(0.134)	(0.132)	(0.130)
DTOURIST(-2)	0.166	0.151	0.158	0.144	0.186
	(0.152)	(0.150)	(0.149)	(0.139)	(0.137)
DTOURIST(-3)	0.138	0.166	0.117	0.118	0.165
	(0.116)	(0.114)	(0.118)	(0.115)	(0.111)
DRTOURIST(-12)	-0.088*	-0.122	-0.098	-0.123	-0.042
	(0.095)	(0.089)	(0.093)	(0.090)	(0.072)
Weekly Google	163.171**	183.293*	44.949**		
	(60.695)	(101.691)	(21.118)		
BETA01	8.400**	-22.710	58.500		
	(3.836)	(60319.490)	(47.676)		
BETA02	19.999***	0.000	-97.531		
	(0.002)	(1141.485)	(201.535)		
BETA03	-0.175*				
	(0.101)				
Monthly Google (-1)				103.521	
				(100.012)	
Monthly Google (-2)				37.673	
				(98.787)	
CONSTANT	-5031.880	-5826.782	-4710.630	-6958.963	2124.010**
	(6997.911)	(5328.457)	(7183.803)	(6162.040)	(987.520)

Notes: The dependent variables are *tourist arrivals* and *overnight stays*; the estimated equation is $tourist_t = \alpha + \sum_{i=1}^{n} \beta_i L^i tourist_t + \gamma \sum_{i=1}^{m} W(k;\theta) L^{k/m} google_t^{(m)} + \epsilon_t^{(m)}$. While Column(2)-(4) represent weekly Google data, Column(5) represents monthly Google data. Column(6) represent ARIMA model without Google trends information. Column (2) represents MIDAS with the weight function of beta formualtion. Column (3) represents MIDAS with the weight function of Exponential formulation, Column (4) represents Almon formulation. Column (5) represents ARIMA(1,1,1) results with monthly data. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.