Tourism Management 46 (2015) 454–464

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach

Prosper F. Bangwayo-Skeete [a, *], Ryan W. Skeete [b, 1]

[a] Department of Economics, University of the West Indies, Cave Hill Campus, P.O Box 64, Bridgetown BB11000, Barbados
[b] Caribbean Tourism Organization, Baobab Tower, Warrens, St. Michael BB22026, Barbados

HIGHLIGHTS

- Unstable world tourism demand reduced predictive power of past tourist arrivals.
- Google Trends data was used to improve forecast accuracy.
- Conducted forecasting competition.
- MIDAS models using Google data outperformed conventional time series models.

ARTICLE INFO

Article history:
Received 13 November 2013
Accepted 22 July 2014
Available online

Keywords:
Tourism demand
Forecasting
Google data
MIDAS
Mixed-data frequency modeling
Caribbean
Tourist arrivals

ABSTRACT

This paper introduces a new indicator for tourism demand forecasting constructed from Google Trends' search query time series data. The indicator is based on a composite search for "hotels and flights" from three main source countries to five popular tourist destinations in the Caribbean. We uniquely test the forecasting performance of the indicator using Autoregressive Mixed-Data Sampling (AR-MIDAS) models relative to the Seasonal Autoregressive Integrated Moving Average (SARIMA) and autoregressive (AR) approach. The twelve month forecasts reveal that AR-MIDAS outperformed the alternatives in most of the out-of-sample forecasting experiments. This suggests that Google Trends information offers significant benefits to forecasters, particularly in tourism. Hence, policymakers and business practitioners especially in the Caribbean can take advantage of the forecasting capability of Google search data for their planning purposes.

1. Introduction

The past two decades witnessed an accelerated use of internet due to its abundant and timely information, resources and services. Given the treasure chest of information on online searches in predicting (current/future) consumer behavior, Google decided to publicly release the intensity search queries data (for any keyword) in 2006 drawing the attention of academic researchers and business practitioners. The emerging research utilizing Google search data are predominately found in epidemiology (see for example Ginsberg et al., 2009; Polgreen, Chen, Pennock, & Nelson, 2008; Valdivia & Monge-Corella, 2010; Wilson, 2009) and economics disciplines (see for example: Carrière-Swallow & Labbé, 2011 on automobile purchases; Vosen & Schmidt, 2011 on consumption; McLaren & Shanbhogue, 2009 on housing; Choi & Varian, 2012; McLaren & Shanbhogue, 2009; Askitas & Zimmerman, 2009; D'Amuri, 2009; D'Amuri & Marcucci, 2010; Suhoy, 2009 on unemployment/unemployment claims).

Google data is particularly pertinent to forecasters following the 2008–2009 "Global Recession" that introduced macroeconomic turbulences, (great) uncertainties and unique shocks rendering conventional historical data usually published with lags slow to reflect on-going structural changes in the economy. Lead indicators such as Google's high frequency data reflect such structural changes making it crucial in forecasting (and nowcasting) economic and financial variables such as GDP growth, unemployment, private consumption, inflation expectations and stock returns (see Da, Engelberg, & Gao, 2011; Guzmán, 2011; Joseph, Wintoki, & Zhang, 2011).

Since Choi and Varian (2012) demonstrated the use of Google Trends data to forecast tourist arrivals in Hong Kong using an autoregressive model, no other similar work was done in the tourism

* Corresponding author. Tel.: +1246 417 4306.
E-mail addresses: prosper.bangwayo-skeete@cavehill.uwi.edu (P.F. Bangwayo-Skeete), rskeete@caribtourism.com (R.W. Skeete).
[1] Tel.: +1246 427 5242.

http://dx.doi.org/10.1016/j.tourman.2014.07.014
0261-5177/

field. We build on their work with some variations. While Choi and Varian searched for "Vacation Destinations/Hong Kong" since Hong Kong is one of the vacation destinations in Google Trends, our study utilizes countries in Caribbean that are not part of Google Trend's destinations. We then construct a new indicator which measures how frequently Google searches are done for two words "hotels" and "flights" in a particular destination country (on a weekly basis), indicating the potential tourism demand. Choi and Varian (2012) aggregated their Google data for analysis, specifically using the first two weekly observations of the month to predict the total monthly visitors (i.e. discarding information for the latter two weeks of the month). Data aggregation and omission leads to information loss, hence, we employ the recently developed mixed-data sampling (MIDAS) approach to fully utilize Google's weekly data.

MIDAS is a time series regression method that allows the explanatory variables and the dependent variables to be sampled at different frequencies, while distributed lag polynomials are used to ensure parsimonious specifications. The method – introduced by Ghysels, Santa-Clara, and Valkanov (2006) – is proven useful for various forecasting applications (see a review of mixed frequency econometric models by Andreou, Ghysels, and Kourtellos (2010b)).

Unlike the voluminous literature on forecasting tourism that employ various econometric methods[2] which primarily relies on historical arrivals data (see for example Athanasopoulos & de Silva, 2012; Lorde & Moore, 2008; Jackman & Greenidge, 2010; Song & Li, 2008 for a review on tourism forecasting) and uses same data frequency; we uniquely explore the application of Google data and the MIDAS approach to (international) tourism demand. The significance of international tourism is demonstrated by its significant contribution to total trade in services, rapid growth in global travelers and its sizable contribution to foreign exchange revenue in developing countries. In 2006, the industry's constituted 27 percent of global trade in services as global tourist arrivals shot to 1035 million in 2012 from 277 million in 1980 (UNWTO, 2007, 2013). International tourism is the principal foreign exchange earner for about 83 percent of developing countries (Roe, Ashley, Page, & Meyer, 2004).

As for the Caribbean, tourism industry is the cornerstone of the region's economic development contributing from 20 to 80% of gross domestic product (GDP) which is disproportionately higher than any other region. Moreover, tourism generates about 25–90% of employment in the region and creates significant foreign exchange earnings, with the proportion ranging between 20 and 85% (Oxford Economics, 2010). Therefore, tourism forecasting using timely information is not only important for business planning, growth strategies and operations of travel and tourism companies but in assessing and predicting the region's overall economic activity. In addition, the enormous consequences of various crises and disasters as well as the delayed publication time render historical data insufficient to accurately plan for items such as inventories and employment levels. In particular, the recent "Global Recession", occasional natural disasters (e.g. hurricanes), terrorist incidents, crime and other exogenous factors cause great instabilities in world tourism demand such that past tourist arrivals data (and/other macroeconomic variables) lost some of their predictive power. Against this background, it is crucial for researchers to develop forecasting methods using lead indicators, Google data in our case, which can accommodate unexpected events in predicting the potential impacts of such one-off events (Song & Li, 2008).

The increased popularity of internet allows millions of persons, to a larger extend in developed countries, to utilize the web search engines to plan their trips or search for travel destinations.[3] The resultant data about tourist destination search queries –an indicator of tourism demand – could thus be highly correlated to actual tourist arrivals in a recipient country. Accordingly, our study examines the usefulness of "hotels", "flights" and "destination country" search indicators to forecasters by testing to what extent the Google search queries data improves the simple Autoregressive – AR method and the Seasonal Autoregressive Integrated Moving Average – SARIMA (using historical tourist arrivals data) compared to an alternative specification, AR-MIDAS (which augments the AR with high frequency Google data). We conduct out-of-sample forecasting experiments using weekly search query data to predict monthly overnight tourist arrivals from the region's three leading source countries, United States (U. S.), United Kingdom and Canada to five Caribbean countries. The popular Caribbean tourist destination countries of Jamaica, Dominican Republic, Bahamas, St. Lucia and Cayman Islands are used with data spanning January 2004–December 2012.

The remainder of the paper describes the MIDAS approach in Section 2. Section 3 presents the data and descriptive statistics. This is followed by the results in Section 4 while Section 5 concludes.

2. MIDAS regression approach

As stated earlier we utilize MIDAS method proposed by Ghysels et al. (2006) and Andreou, Ghysels, and Kourtellos (2010a) to predict monthly tourist arrivals using weekly Google search indicators. This paper is the first to apply the MIDAS method to tourism forecasting. Traditionally, studies predominantly employed the method to forecast financial variables – stock returns and stock return volatility (Ghysels et al., 2006) and macroeconomic time series – GDP growth, unemployment, private consumption and inflation expectations (Armesto, Engemann, & Owyang, 2010; Kuzin, Marcellino, & Schumacher, 2011; Tay, 2006). This regression method deals with a common problem of data sampled at different frequencies. For example macroeconomic data is usually sampled at monthly or quarterly frequencies while financial time series are usually sampled daily or higher frequencies. In such instances, the high frequency financial variables are used to forecast the low frequency macroeconomic variables.

The customary approach to estimation involves aggregating all variables to the same (low) frequency using an equal weighting scheme. However, the equal weighting scheme does not fully exploit the information available in high frequency data which may lead to information loss and thereby to inefficient and/or biased estimates (Armesto et al., 2010). In contrast, the MIDAS regression model projects the high frequency process onto the low frequency process with a parsimonious weighting scheme. Consider the following MIDAS model:

$$Y_t = \alpha + \sum_{i=1}^{p} \beta_i L^i Y_t + \gamma \sum_{k=1}^{m} \Phi\left(k;\Theta\right) L_{HF}^{k} X_t + \varepsilon_t \qquad (1)$$

where the function $\Phi(k;\Theta)$ is a polynomial that determines the weights for temporal aggregation. X_t is a high frequency explanatory variable and Y_t is a low frequency dependent variable. L is the

[2] The econometric forecasting methods employed in tourism demand analysis include: structural time series approach, univariate and integrated autoregressive moving-average (ARIMA) models, the Generalized Autoregressive Conditional Heteroskedastic (GARCH) framework, autoregressive distributed lag model (ADLM), the error correction model (ECM), the vector autoregressive (VAR) model, and the time varying parameter (TVP) models.

[3] As a result, ticket bookings have been predominantly done online.