Box 2.1: Can the Volume of Queries on Search Engines Forecast Singapore's Visitor Arrivals?

Technological advancements have led to a surge in data generation and storage in recent years. As a result of this Big Data revolution, a slew of supplementary indicators is now available and can be used to estimate the performance of key economic activities. In particular, real-time Big Data can provide immediate insights into consumer behaviour, especially when more traditional economic indicators are typically released with a time lag. This form of contemporaneous forecasting or 'nowcasting'[1] is of particular interest as it can inform economic surveillance by providing timely estimates of key economic indicators. For instance, the volume of queries on search engines which are available on a real-time basis can be a proxy for consumer interest and in turn, be used as an indicator to forecast the growth of consumer-oriented sectors.

Within the tourism industry, travellers are increasingly using the internet when planning their itineraries. These online travel searches leave behind a digital footprint and present a potential data mine to assess visitor arrivals to a particular destination. For example, the volume of online searches for Singapore-related tourism queries originating from Australia may reasonably be seen as a proxy for the interest for travel demand from Australia to Singapore.

Building on from this hypothesis and adapting from an earlier study by Choi and Varian (2011), we constructed a time series econometrics model to assess whether predictions of visitor arrivals to Singapore can be improved with data on tourism-related volume searches on Google. Specifically, we focused on visitor arrivals to Singapore originating from six different source markets, viz, Australia, Canada, Germany, Malaysia, US and UK. These six source markets were selected based on the availability of data from Google, specifically, the Google Trends index[2]. We ran the model on monthly data for the period January 2004 to March 2013.

In particular, the following auto-regressive moving average (ARIMAX) model equation was estimated for each source market[3]:

$$VA_t = \beta_1 VA_{t-1} + \beta_2 VA_{t-12} + \beta_3 Trends_{t-i} + \beta_4 REER_{t-j} + AR(m) + MA(n)$$

Where:

VA is the visitor arrivals to Singapore from country k. The subscript $t\text{-}1$ and $t\text{-}12$ denote visitor arrivals in the preceding month and visitor arrivals in the same month a year ago respectively. This structure is known in the literature as a seasonal autoregressive model.

$Trends_{t-i}$ is the Google Trends index of tourism-related queries on Singapore originating from country k and lagged by i periods based on the cross-correlation between visitor arrivals and Trends.

$REER_{t-j}$ is the real effective exchange rate[4] of country k and lagged by j periods based on the cross-correlation between visitor arrivals and REER.

[1] 'Nowcast' is a contraction for 'now' and 'forecasting' which can provide a good prediction for the near future.

[2] Google Trends provides an index of the volume of Google queries categorised by geographic location and category. The Google Trends index reports a query index rather than the raw level of queries for a given search term. First, the total query volume for the search term is normalised by the total number of queries in that region at a point in time. Second, the series is then indexed to the time period where the maximum normalised search volume is observed. Hence, the Google Trends index at each date refers to the percentage deviation from the maximum normalised search volume.

[3] Prior to estimating this equation, we first tested for unit roots to ensure the regression would not generate spurious results. Next, we ran cross-correlations between visitor arrivals and the explanatory variables to ascertain the appropriate lead-lag relationship between the dependent variable and the explanatory variables.

[4] The real effective exchange rate is defined as $\frac{CPI_{SG}}{CPI_k} \times NEER_k$, where CPI_{SG} refers to the consumer price index (CPI) of Singapore, CPI_k refers to the CPI of country k and $NEER_k$ is the nominal effective exchange rate of country k.

$AR(m)$ and $MA(n)$ refer to autoregressive and moving average terms which are selected based on the correlogram of the residuals and the Akaike Information Criteria.

Google Trends was positively correlated with Singapore's visitor arrivals

From the results in Exhibit 1, we observe that Google Trends is positively correlated with visitor arrivals from several countries. Depending on the country of origin, Google Trends may be a coincident indicator or a leading indicator. For instance, the Google Trends index two months ago is a reasonably good indicator of visitor arrivals originating from the UK.

Exhibit 1: Regression Results for the ARIMAX model with Google Trends as an explanatory variable

Specification	(1)	(2)	(3)	(4)	(5)	(6)
Countries	Australia	Germany	UK	Canada	Malaysia	US
$Trends_{t-i}$	170.8**	43.59*	60.55**	24.51*	395.3**	116.9**
VA_{t-1}	-0.0250	-0.0484	0.0308	-0.0020	-0.0702*	-0.0281
VA_{t-12}	1.048**	0.9789**	0.9542**	0.9312**	0.9559**	0.9153**
$REER_{t-j}$	99.53	51.94	-62.49	-5.623	-61.52	-32.92
Number of lags:						
$Trends_{t-i}$	0	2	2	0	0	0
$REER_{t-j}$	0	2	1	0	2	0
AR and MA Structure:						
AR(1)	No	No	No	Yes	Yes	Yes
AR(12)	Yes	Yes	Yes	No	No	No
MA(1)	Yes	No	Yes	Yes	Yes	Yes
Observations[†]	92	92	92	91	91	91

*** P-value<0.01, ** P-value<0.05, * P-value<0.10

[†] Note: The number of observations differs as specifications (1)-(3) takes the first difference of visitor arrivals, while specifications (4)-(6) takes the second difference of visitor arrivals based on the results of the unit root tests.

The inclusion of Google Trends improved forecasting accuracy

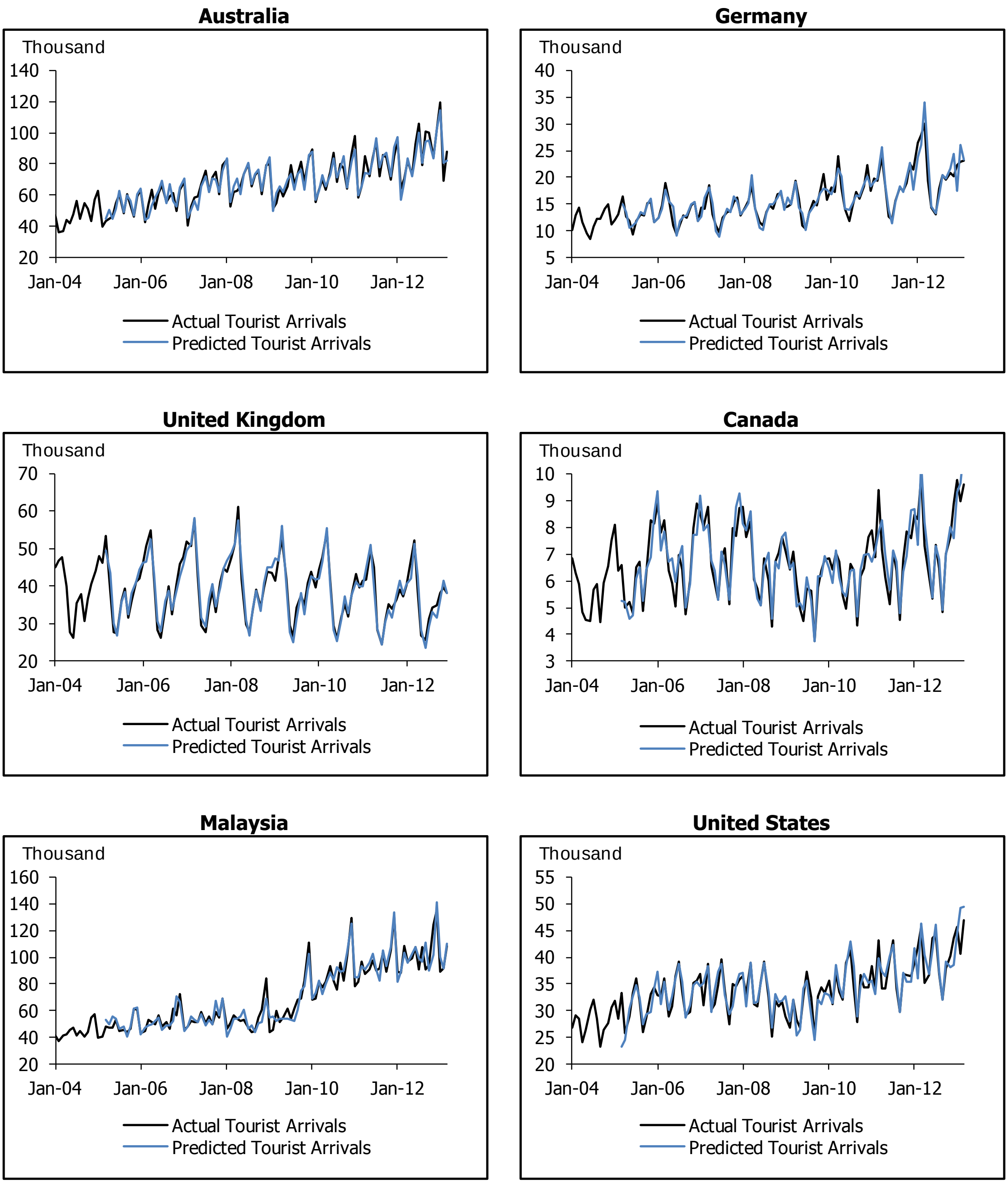
Next, to evaluate the ability of the ARIMAX model to nowcast visitor arrivals to Singapore, we used data from the period January 2004 to March 2012 to estimate the coefficients of the model, both with and without the Google Trends variable. We conducted in-sample forecasting for visitor arrivals from January 2005 to March 2012, and then one-step-ahead out-of-sample forecasting for visitor arrivals from April 2012 to March 2013. The graphs of the actual and predicted visitor arrivals over this period for the model with Google Trends are plotted in Exhibit 2. In addition, Exhibit 3 gives a quantitative evaluation of whether the inclusion of Google Trends improves forecasting performance. Specifically, we calculated the deviation of predicted tourist arrivals ($\overline{VA}_t$) from actual tourist arrivals with two measures. First, the mean absolute prediction error (MAPE) was calculated based on the following definition:

$$MAPE = \frac{\sum |\overline{VA}_t - VA_t|}{n}$$

Second, the root mean squared prediction error (RMSPE) was computed. The RMSPE is defined as:

$$RMSPE = \sqrt{\frac{\sum(\overline{VA}_t - VA_t)^2}{n}}$$

Exhibit 2: One-Step Ahead Forecasts of the ARIMAX Model with Google Trends



Source: Singapore Tourism Board

Exhibit 3: Error of ARIMAX models with and without Google Trends as an explanatory variable

Country	MAPE		RMSPE (%)	
	With Google Trends	Without Google Trends	With Google Trends	Without Google Trends
Australia	36.3	37.1	0.70	0.73
Germany	11.9	12.0	0.94	0.95
UK	18.8	19.1	0.56	0.57
Canada	5.12	5.26	0.87	0.91
Malaysia	49.2	50.2	0.92	0.95
US	18.0	18.7	0.71	0.73

Overall, we find that the inclusion of Google Trends as an explanatory variable lowered both the MAPE and the RMSPE, indicating that forecasting accuracy is improved.

While we have established that Google Trends is informative in nowcasting visitor arrivals from several countries, there are several limitations on its use. First, the Google Trends index only provides an indication of searches on Google. In countries which mainly use search engines other than Google, such as China, Japan and South Korea, Google Trends may be a poor proxy for consumer sentiments. Second, below a certain threshold, the Google Trends index is censored to zero. As such, if the share of the search term 'Singapore' is a relatively small share of the overall search volume from a certain country, the Google Trends index may not be available. Third, Google Trends is an index normalised by the total volume of searches originating from a region. Hence, a decline in the Google Trend index may be due to an increase in the overall search volume instead of a fall in travel interest to Singapore. A more accurate forecast may be computed if the actual volume of search queries is known.

Big Data may be able to provide timely indicators for key socio-economic indicators

This study is an attempt to highlight the use of publicly available information to estimate key economic variables in Singapore. With the rise of Big Data, more information would increasingly be made available, with some providing useful content that can support economic surveillance efforts. For instance, Choi and Varian (2011) find that Google Trends can be used to forecast several other consumer-related indicators such as retail sales, motor vehicle sales and home sales. Besides online Google searches, Dzielinski and Hasseltoft (2012) use firm-specific news items to forecast the aggregate stock returns and volatility of firms. These are but a few examples of the practical applications of Big Data.

Contributed by:
Goh Sze Lyn, Research Assistant
Leong Chi Hoong, Economist
Economics Division
Ministry of Trade and Industry

REFERENCES

Choi, H. and Varian, H. (2011). "Predicting the Present with Google Trends." Google.

Dzielinski, M. and Hasseltoft, H. (2012). "Aggregate News Tone, Stock Returns and Volatility." S.S.R.N.