The role of disaggregated search data in improving tourism forecasts: Evidence from Sri Lanka

Kanchana Wickramasinghe\textsuperscript{a} and Shyama Ratnasiri\textsuperscript{b}

\textsuperscript{a}Department of Accounting, Finance & Economics, Griffith Business School, Nathan Campus, Griffith University, Brisbane, Queensland, Australia; \textsuperscript{b}Department of Accounting, Finance & Economics, Griffith Business School, Gold Coast Campus, Griffith University, Brisbane, Queensland, Australia

\textbf{ABSTRACT}

Formulation of effective policies to enhance the resilience of tourism following the COVID-19 pandemic essentially requires comprehensive empirical information on changes in tourism demand and associated economic costs. The paper makes a novel contribution to tourism literature by employing regionally and temporally disaggregated tourism data and Google search data in improving the accuracy of tourism forecasts. Further, the paper adopts two timeseries variables namely tourist arrivals and guest nights in order to understand the changes due to COVID-19 in tourism demand more comprehensively. Monthly data on international tourist arrivals, guest nights and Google trends from 2004 to 2019 are used to produce regionally disaggregated (Europe, Asia, the Pacific, America, Other) monthly tourism forecasts for Sri Lanka. We find that SARMAX models outperform the other models (ARIMA, ARIMAX, SARIMA) in forecasting tourism demand following COVID-19. Interestingly, the paper makes a further step in utilizing forecasts in estimating foregone economic benefits due to COVID-19 pandemic. We find a notable difference in estimated direct economic loss depending on the variable used in estimates. The percentage loss is 40\% when arrival forecasts are used in estimates and 29\% when guest night forecasts are used in estimates. This provides important policy implications for improving post-COVID tourism.

\textbf{KEYWORDS}

Tourism forecasting; COVID-19; Google data; SARIMAX; tourist expenditure

\section{1. Background}

Forecasting is important for decision makers due to the perishable nature of tourism (Goh & Law, 2011; Khaidi et al., 2019). Accurate research-based information on future tourism demand is particularly in high demand following the COVID-19 pandemic, which has posed unprecedented challenges for tourism at global level. While all the tourism destinations are affected by the pandemic, the exact nature and degree of the effects of the COVID-19 seem to be different across destinations. Disaggregated forecasts can serve as an important policy guiding tool in understanding the varied impacts of COVID-19 on international tourism. Such forecasts are more useful in decision making than total forecasts, as they provide more detailed and diverse information on tourism demand (Song & Li, 2008). We believe that forecasts are also highly useful in understanding the foregone tourist arrivals and economic benefits. In this paper, we intend to produce geographically and temporally disaggregated forecasts, which would be useful in tourism-related policy formulation following COVID-19.

Undeniably, the reduced demand in the travel and tourism sector can have substantial downside risks on the economy (Boone et al., 2020; UNWTO, 2020). Countries which are heavily dependent on
tourism have faced significant declines in service activity, particularly in hospitality, food, entertainment, and retail services due to COVID-19 (World Bank, 2020). However, there is a lack of empirical evidence on the magnitude of negative effects of COVID-19 on tourism to guide decision making. Further, from the policy point of view, it is important to understand how the composition of COVID-19 effects in terms of declines in tourism demand from different geographical regions and resultant economic losses. In this paper, we attempt to quantify the foregone tourist expenditure due to COVID-19 pandemic with the use of geographically and temporally disaggregated forecasts.

Accurate tourism forecasts are key inputs for effective government policy planning to enable tourism development and economic growth (Jiao & Chen, 2019; Li, 2009). Number of tourist arrivals are the most commonly used indicator to measure overall changes in tourism demand (Wu et al., 2017). However, we see a limitation in using only one timeseries in forecasting tourism demand. In order to capture diverse impacts of tourism demand, we employ a second timeseries namely the number in guest nights, in addition to the number of tourist arrivals. The guest night variable captures the implications of COVID-19 for important sub-sectors of tourism including the hospitality industry. Interestingly, the number of guest nights implies both the number of arrivals and durations of stay, which is directly useful for tourism business operators. For businesses in this industry such as accommodation and hospitality providers, travel planners/trade, tour operators, adventure and attractions operators, accurate forecasts help significantly in developing their pricing and operation strategies including staff, capacity and resource management (C. A. Witt & Witt, 1995; Wu et al., 2017). Also, tourism related investment decisions by both the private and public sectors that are underpinned by accurate medium- and long-term forecasts can be more effective (Wu et al., 2017).

The specific contributions of the paper to the literature are several folds. Firstly, the paper uses two timeseries variables to denote the demand for tourism, namely, international arrivals and international guest nights. Most studies in the tourism forecasting literature only used a single time series. Tourist arrivals is the most commonly used single variable in forecasting tourism demand (Wu et al., 2017), due to the high availability of international tourism data (Haiyan Song et al., 2019). Tourism guest night forecasts on the other hand are widely used for the hospitality sub-sector as information regarding arrivals and durations of stay is embedded within this variable. We intend to test if there is a notable difference in our estimates when these two time-series variables are used in modelling.

The second empirical contribution of the paper is that it produces geographically and temporally disaggregated forecasts for total international arrivals and guest nights. We use monthly data in order to carefully understand the tourism demand changes over short run. All international arrivals to a single country are disaggregated into five major regions in this paper. The disaggregation is done considering the geographical distribution of origins and major source countries for international tourism. There is limited literature that uses disaggregated data on international tourism (Song et al., 2019; Song & Li, 2008). Turner and Witt (2001) forecast international tourist arrivals to Australia from selected countries including New Zealand, Japan, the UK and the USA. The current paper contributes to the literature by developing models to forecast disaggregated international arrivals and guest nights from all the regions of the world to a single destination.

Further, only a few studies augment big data in producing disaggregated forecasts capturing the total tourism demand in a single destination, though a growing number of papers are using them in aggregated forecasting. Only a few studies in the literature use Google trend data in predicting guest nights, in particular. Dinis et al. (2016, 2017) for instance find that Google trend data improves guest night forecasting in the hospitality industry. The current paper takes the approach of estimating disaggregated forecasts for two timeseries based on Google trend data. This approach captures the variations in travel patterns from different regions and helps in formulating tailor-made strategies for tourism promotion for specific regions, following the COVID-19 shock. As the internet has become a key information source for consumers, internet search patterns may be strongly correlated
with the future behaviour of consumers (Park et al., 2017). Potential tourists widely use only search engines to get information on potential destinations, cost of travel, accommodation, attractions, etc. (Jiao & Chen, 2019).

The objective of the paper contributes to the evolving post-COVID tourism literature through two aims. Firstly, the paper aims to develop accurate geographically and temporally disaggregated forecasts for international arrivals and guest nights by augmenting Google search data in modelling. Four types of timeseries models, representing univariate and multivariate approaches are utilized for this. Tourism is currently undergoing the most troublesome downturn in history due to the ongoing COVID-19 pandemic. The second aim of the paper is to employ quantitative forecasts in estimating the forgone tourist expenditure due to COVID-19 as an illustration. Thus, this paper provides a timely contribution to the literature by integrating forecasts in quantifying economic impacts of an important shock.

In order to perform a detailed analysis of temporal and geographical changes in tourism demand in response to COVID-19, we base our analysis on a single tourism destination. We use data from Sri Lanka, which is an emerging economy and an important tourism destination. International arrivals to Sri Lanka increased by around 5.2 times during the period of 2009–2018 as per the administrative data (SLTDA, 2018). In 2019, Sri Lanka was a highly ranked tourism destination by Lonely Planet (‘Best in travel in 2019’) and CNN Travel (‘20 best places to visit in 2020’). Accurate forecasting is often constrained due to the lack of data (Volchek et al., 2019). Sri Lanka’s tourism agencies maintain records on monthly tourist arrivals and guest night data by individual source countries and regions in their administrative reports. This further justifies Sri Lanka being chosen as a case study for the current research. Further, Sri Lanka is one of the world’s hard-hit tourism destination as a result of the 30-year-old civil war, recent terror attacks and the COVID-19 pandemic. The paper makes a timely contribution to the literature by producing accurate forecasts which can be utilized in future tourism development efforts. Interestingly, our approach can be easily replicated for other destinations and regions for comprehending the demand responses to global level shocks.

The remainder of the paper is organized as follows. The second section presents a brief overview of history and trends in tourism in Sri Lanka. Methodological approach details are presented in the third section. The fourth section presents the results of the quantitative analysis. The final section provides a discussion about the results and the conclusions that can be derived from the results.

![Figure 1. Annual international tourist arrivals and international tourist guest nights in Sri Lanka.](image-url)
2. Tourism trends in Sri Lanka

Sri Lanka has an organized tourism sector, which is around 5.4 decades old (Samaranayake et al., 2013). Year 2009 records a significant milestone in tourism history in Sri Lanka (Figure 1). A significant revival in international arrivals and guest nights is observed following the end of civil conflict in May 2009 (Bandara, 2019; Wickramasinghe, 2011). Tourism remained a fast-growing sector for nearly a decade until 2019 due to the post-conflict peaceful environment that prevailed in the country and the receipts from tourism also showed a steady growth during the period. Total arrivals in 2018 were over 2.3 million, which was equivalent to around 10.75% of Sri Lanka’s population in the same year.

The terror attacks that took place in April 2019 led to a sudden reduction in arrivals and guest nights in 2019. Monthly data showed that arrivals had almost recovered by December 2019. International travel accolades received by Sri Lanka from Lonely Planet and CNN Travel confirms the recovery. However, Sri Lanka has now started to go through a second major negative shock due to the ongoing COVID-19 pandemic situation which has reversed the recovering trend of Sri Lanka tourism (CBSL, 2020). Similar to the majority of countries around the world, for the first time in the history, Sri Lanka is experiencing zero tourism from April 2020.

A large percentage of international tourist arrivals to Sri Lanka has originated from European and Asian regions over the years (Figure 2). Arrivals from other Pacific, American and other regions (African and Middle East regions) are comparatively low but show a positive trend. In 2018, arrivals from European and Asian regions were 44 and 41% of total arrivals. Arrivals from the Pacific, America and other regions were 5.4, 5.9 and 3.7% of total arrivals respectively. The number of guest nights originating from the Asian region shows fluctuations due to changes in the durations of stay.

Examination of tourist arrivals by country of origin shows that a few countries dominate as sources for Sri Lanka’s tourism. More than 68% of arrivals originated from 10 individual countries (Table 1) in 2018. In relation to international guest nights, around 64% of guest nights originated from 10 source countries. The current paper intends to produce forecasts for total arrivals and guest nights disaggregated into five major regions of the world.

The top countries represent the broad regions of Europe, Asia, the Pacific and America. The paper therefore intends to produce forecasts for the said four regions, while having an ‘other’ category to capture the rest of the world. The other category, therefore, includes the countries from African and Middle Eastern regions.

3. Methodology

3.1. Data

Monthly tourist arrival data and guest night data by five regions (Europe, Asia, the Pacific, America, other) for the period of January 2004 to December 2019 were obtained from the Annual Statistical
Reports of the Sri Lanka Tourism Development Authority. The descriptive statistics are given in Table 2. We conducted stationarity tests in order to eliminate issues of spurious regression due to the presence of unit roots in the time series.

We used two explanatory variables in the multivariate time series models, namely, Google trend data and dummy variables to represent structural breaks.

3.1.1. Google trend data

We obtained Google trend data from https://trends.google.com/trends based on the selected search terms as discussed in the next paragraph. Monthly Google data was obtained as the arrival data is available on monthly basis. Google data is available from January 2004 onwards. Google data is available in the form of a normalized index, which takes positive values between 0 and 100. The index is derived by dividing the total number of searches for a given search term in a given geographical area by the total number of searches for the same area for a given period of time. The source countries which did not show a substantial number of arrivals to Sri Lanka were excluded from the calculation.

Selection of appropriate key words is an important step in augmenting search data into forecasting models (Jiao & Chen, 2019; Park et al., 2017). We selected around 10 search terms that may explain the travel decisions to visit Sri Lanka and checked the Google trends for each search term. Search terms which mostly had zeros were excluded from the analysis, as they are not explaining the information search behaviour of tourists. We selected three major search terms which correlate well with the arrivals data. They include ‘Sri Lanka visa’, ‘Sri Lanka flights’ and ‘Sri Lanka hotels’.

Figures 3 and 4 show the behaviours of arrival data and guest night data separately with Google trend data for each region.

Both current values and lagged values of Google trend data are used in forecasting models. Choosing the number of lags was done based on the travel planning information revealed by the survey of departing foreign tourists from Sri Lanka conducted by the Sri Lanka Tourism Development Authority in 2017. The survey finds that around 91% of tourists plan the trips six months ahead.

| Table 1. Percentage arrivals to Sri Lanka from top ten source markets in 2018. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Source market   | Percentage arrivals | Source market   | Percentage arrivals |
| India           | 18.2             | India           | 11.0             |
| China           | 11.3             | United Kingdom  | 10.5             |
| United Kingdom  | 10.9             | Germany         | 9.7              |
| Germany         | 6.7              | Australia       | 5.9              |
| Australia       | 4.7              | France          | 5.8              |
| France          | 4.6              | Canada          | 4.1              |
| Maldives        | 3.2              | Russia          | 3.2              |
| USA             | 3.3              | USA             | 3.1              |
| Russia          | 2.8              | Maldives        | 3.0              |
| Netherlands     | 2.4              |                 |                 |

Source: Annual Statistical Report 2018, Sri Lanka Tourism Development Authority.

| Table 2. Descriptive statistics of monthly arrivals. |
|-----------------|-----------------|-----------------|-----------------|
| Region          | Arrivals         |                            | Guest nights    |
|                 | Mean             | Standard deviation | Min             | Max             | Mean             | Standard deviation | Min             | Max             |
| Europe          | 41,286           | 28,575            | 9,012           | 135,934         | 468,514          | 358,001           | 81,108           | 1,671,988        |
| Asia            | 40,398           | 26,349            | 8,859           | 104,154         | 306,260          | 197,702           | 69,100           | 844,078          |
| Pacific         | 4,866            | 3,763             | 1,026           | 22,530          | 56,127           | 45,338            | 11,799           | 238,727          |
| America         | 5,555            | 3,640             | 1,335           | 20,943          | 61,566           | 42,823            | 15,120           | 245,033          |
| Other           | 4,889            | 4,225             | 411             | 20,001          | 44,121           | 38,109            | 3,555            | 181,009          |
| Total           | 96,804           | 62,120            | 24,739          | 253,169         | 983,144          | 672,949           | 225,125          | 2,734,225        |
Therefore, we included 6 lags for the Google trend variable. The correlation of arrivals and guest nights from different regions with current and lagged values of Google trend data from each region is statistically significant at 0.01 level (Table 3).

**Table 3.** Correlation coefficients between arrival data / guest night data and Google trend data.

| Google trend | Europe        | Asia         | The Pacific | America    | Other       | Total       |
|--------------|---------------|--------------|-------------|------------|-------------|-------------|
| Arrivals     |               |              |             |            |             |             |
| Current level| 0.5989***     | 0.3059***    | 0.5872***   | 0.5772***  | 0.5609***   | 0.4055***   |
| Lag 1        | 0.5452***     | 0.3007***    | 0.5601***   | 0.5277***  | 0.5453***   | 0.3601***   |
| Lag 2        | 0.4485***     | 0.3007***    | 0.5736***   | 0.4898***  | 0.4267***   | 0.3211***   |
| Lag 3        | 0.3518***     | 0.2938***    | 0.5210***   | 0.4879***  | 0.4137***   | 0.2539***   |
| Lag 4        | 0.3009***     | 0.2608***    | 0.5087***   | 0.4610***  | 0.4268***   | 0.2476***   |
| Lag 5        | 0.3335***     | 0.2582***    | 0.5394***   | 0.4781***  | 0.4551***   | 0.2537***   |
| Lag 6        | 0.3918***     | 0.2600***    | 0.5205***   | 0.5247***  | 0.4797***   | 0.2632***   |
| Guest nights |               |              |             |            |             |             |
| Current level| 0.6097***     | 0.2853***    | 0.5982***   | 0.5844***  | 0.5583***   | 0.4050***   |
| Lag 1        | 0.5585***     | 0.2843***    | 0.5736***   | 0.5366***  | 0.5460***   | 0.3611***   |
| Lag 2        | 0.4677***     | 0.2684***    | 0.5870***   | 0.5023***  | 0.4278***   | 0.3231***   |
| Lag 3        | 0.3762***     | 0.2341***    | 0.5472***   | 0.5014***  | 0.4145***   | 0.2577***   |
| Lag 4        | 0.3259***     | 0.2229***    | 0.5313***   | 0.4722***  | 0.4272***   | 0.2503***   |
| Lag 5        | 0.3550***     | 0.2278***    | 0.5609***   | 0.4895***  | 0.4549***   | 0.2572***   |
| Lag 6        | 0.4115***     | 0.2163***    | 0.5451***   | 0.5362***  | 0.4796***   | 0.2680***   |

***indicates 0.01 significance level.
3.1.2. Dummy variables for structural breaks

The international arrivals in Sri Lanka show notable fluctuations due to the impact of civil conflict. We checked the time series for the presence of statistically significant structural breaks using the Quandt–Andrew’s Breakpoint test and multiple breakpoint test. Dummy variables are incorporated to capture the effects of structural breaks in the multivariate models. Such effects are not considered in the univariate models.

3.2. Models

Tourism literature clearly identifies that there is no single forecasting method that performs best in all situations (Ghalekhondabi et al., 2019; Jiao & Chen, 2019; Khaidi et al., 2019; Song et al., 2019; Song & Li, 2008; C. A. Witt & Witt, 1995). Autoregressive Integrated Moving Average (ARIMA) is the most frequently used univariate model in the tourism economics literature. The model is also widely used as a benchmark for evaluation and comparison purposes (Jiao & Chen, 2019). A review conducted by H. Song et al. (2019) finds that SARIMA models have received substantial attention in tourism forecasting studies in recent years. For Sri Lanka, Thushara et al. (2019) find that SARIMA models outperform the ARIMA models in forecasting arrivals from the top 10 source countries.

Literature points out that extensions of univariate models to include explanatory may improve forecasts. However, there is limited availability of data types that can be used as explanatory variables.
variables in forecasting models which makes implementing such models challenging (Volchek et al., 2019). Use of online big data has emerged as a solution to this in the literature (Jiao & Chen, 2019). Online search information is used by individuals to gain an overall evaluation of potential destinations using public search engines such as Google and Baidu, a search engine used in China (Jiao & Chen, 2019). Search engine data have several advantages over traditional economic variables. These real-time data are generally of high frequency, available free of charge and are sensitive to tourist behaviour (Wu et al., 2017). There is much potential for using search query data and web traffic data in tourism forecasting (Wu et al., 2017), though this approach is still in its infancy (Song et al., 2019). The current paper aims to make extensions to ARIMA and SARIMA models by employing Google search data and develop ARIMAX and SARIMAX models for forecasting.

We estimate four types of models, namely, ARMA, ARMAX, SARMA and SARMAX for the arrivals and guest nights from the five regions (Europe, Asia, the Pacific, America and other) and for their totals. For each region, the four types of models are estimated at both the level and log-transformed version for each variable. In total, eight models are estimated for each region for arrivals variable and guest night variable separately. This facilitates identification of a highly accurate forecasting model for each region for arrivals and guest nights separately.

ARMA model can be specified as,
\[ \phi(B)y_t = c + \theta(B)e_t, \]
where,
\[ \phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p \text{ and} \]
\[ \theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q \]

ARMAX model can be specified as,
\[ \phi(B)y_t = \beta x_t + \theta(B)z_t \]
where \( x_t \) is a covariate at time \( t \) and \( \beta \) is its coefficient. \( B \) is the backshift operator.

SARMA model can be specified as follows:
\[ \phi(B^{12})\Theta(B) \ln(Y_{j,t}) = \alpha + \theta(B^{12})\Theta(B)Z_{j,t} \]

SARMAX model can be specified as follows:
\[ \phi(B^{12})\Theta(B) \ln(Y_{j,t}) = \alpha + \sum_{i=0}^{p} \beta_i \ln(X_{j,t-i}) + \theta(B^{12})\Theta(B)Z_{j,t} \]
where,
\[ \phi(B^{12}) = 1 - \phi_1 B^{12} - \phi_2 B^{24} - \ldots - \phi_p B^{12p} \]
\[ \Theta(B^{12}) = 1 - \theta_1 B^{12} - \theta_2 B^{24} - \ldots - \theta_p B^{12p} \]
\[ \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \]
\[ \Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_p B^p \]

where \( \phi B^{12} \) is the seasonal operator and \( \Theta B^{12} \) is the AR operator, \( \Theta B^{12} \) is the seasonal moving average (MA) operator and \( \Theta B^{12} \) is the MA operator.

Data frequency is set as 12, as we use monthly data in the models, where \( Y_{j,t} \) is the number of arrivals from region \( j \) at time \( t \), \( x_{j,t} \) represents covariates, \( Z_{j,t} \) is the error term for region \( j \) at time \( t \), \( B \) is the Backshift operator.

The optimal models under eight types of models for each region were selected based on the Akaike information criterion (AIC). The model with the lowest AIC value was selected for forecasting.
in the next step. Eight optimal models were employed for forecasting for each region. The sub-sample from January 2004 to December 2017 was used to train the models. The sub-sample from January 2018 to April 2019 was used to evaluate the models. We intentionally excluded the period from May to December 2019 in evaluating the models, due to the terror shock and resultant downturn in tourism which occurred in late April in 2019. Forecasts were derived up to April 2020.

The accuracy of forecasts depends on how close the forecasts are to actual values and the difference between forecast values and actual values denote the forecast errors. The most commonly used measures to assess the accuracy of tourism forecasts are mean absolute percentage error (MEA), the root mean square error (RMSE), the root mean square percentage error (RMSP) and the mean absolute error (MAE) (Wu et al., 2017). Other available methods include mean square error (MSE), Theil’s U-statistics, the mean absolute deviation and the mean absolute square error. Various studies have used one or more methods to compare the accuracy of forecasts. We adopt MAPE, RMSE and MAE methods for selecting the best forecast models, as they are the widely used measures in tourism forecasting literature. They can be specified as follows.

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2}$$

$$\text{MAE} = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n}$$

where $A_t$ is the actual value at time $t$, $F_t$ is the forecast at time $t$ and $n$ = number of observations.

The forecast calculated for each region is compared against the actual arrivals recorded in the administrative reports to calculate the level of deviation. Percentage deviation is calculated using the following formula.

$$\text{Percentage deviation} = \frac{(\text{Actual values} - \text{Forecast values}) \times 100}{\text{Forecast values}}$$

Since travel reductions can be attributed to the impacts of COVID-19, the magnitude of deviation can be considered as an important indicator of the magnitude of the impact of COVID-19 on international tourism. Deviations are calculated for each region separately. Disaggregated deviations provide useful information on the magnitude of the impacts for each source region.

We use the arrival forecasts and guest night forecasts to estimate the foregone tourist expenditure to Sri Lanka. The per tourist expenditure per day data of Tourism Expenditure Survey are used to calculate the total forecast and actual expenditure for each region (SLTDA, 2017).

4. Results

4.1. Stationarity tests

Table 4 shows the results of the Dickey–Fuller generalized least squares (DF-GLS) test for estimating stationarity in timeseries models. All the variables and their levels log-transformed are stationary as per the test results. We employ the variables both at levels and log-transformed forms of each variable in our forecasting models. Order of integration for each model is thus considered as zero.

4.2. Dummy variables to represent structural breaks

Tourism in Sri Lanka witnessed a notable recovery and a fast growth in tourist arrivals following the end of civil conflict in May 2009. We intended to check the impact of the event in forecasting the
arrivals and guest nights for each region. Both the multiple breakpoint test and the Quandt–Andrew’s test provided significant breakpoints which are in line with the observed pattern in arrivals, when trends and constants are incorporated (Table 5). It is interesting to note that the significant breakpoints are different across the regions. For the Asian region, the structural break point was closer to the period where the end of the conflict took place, when compared with other regions. For Europe, the Pacific and America, a significant breakpoint is recorded 6 months after the end of the conflict. Geographical distance and associated travel costs might be a reason for a delayed increase in arrivals and guest nights from such regions. Further, this could be due to an array of factors that determine the willingness to travel to destinations following conflicts. In our multivariate time series models, we incorporate separate dummy variables to capture the effect of the structural breaks for each region.

### Model estimation results

The optimal models (identified using AIC values) for each region and model type are given in Table 6. For each region, SARMA and SARMAX models outperform their univariate counterparts (ARMA and ARMAX, respectively). We therefore report only the optimal models for SARMA and SARMAX. These models were used for dynamic forecasting. As we highlight in the previous section, SARMAX models included dummy variables to represent structural breaks, Google trend data and their lags, in addition to seasonal and non-seasonal terms.

Best forecasting models for arrivals and guest nights were separately identified with the use of MAPE, RMSE and MAE values. Models with the smallest values of the three forecasting accuracy indicators were selected as the best models that can be used for forecasting. The three measures consistently identified the models with lowest errors in most of the models. However, only two measures were seen consistent with regard to arrivals from America (Table 7) and guest nights from Europe (Table 8). In forecasting arrivals from America region, both RMSE and MAE indicators revealed the SARMAX as the model with lowest error, whereas MAPE identified SARMA model as the lowest
error. Considering the values of each indicator we came to the conclusion that SARMAX is the best model for forecasting tourist arrivals from America. Same approach was adopted in identifying the best model for forecasting guest nights from Europe. The arrivals models and guest nights models with the smallest forecasting errors are highlighted in Tables 7 and 8, respectively.

For all regions, models with seasonal terms (SARMA or SARMAX) outperformed their counterparts (ARMA and ARMAX). A previous study confirmed that the use of seasonal terms improves the accuracy in forecasting tourism demand in Sri Lanka (Thushara et al., 2019). It is also clear that Google trend data and dummy variables have increased the accuracy of forecasts in 4 models out of 6 models which forecast arrivals. In the case of guest nights, we find that Google trend data has improved the accuracy of 5 models out of 6 models. This adds evidence to the existing literature on the use of Google trend data to increase forecasting accuracy.

For both total arrivals and total guest nights, SARMAX (3,0,4)(1,0,2)12 model was the best fit for log-transformed values. For arrivals from Europe region, SARMAX(3,0,4)(1,0,2)12 at levels was found to be the best fit model. For guest nights from Europe region, log SARMAX(3,0,3)(2,0,2)12 model was better than other models. Use of Google trend data and dummy variables did not help to improve forecasting of either arrivals or guest nights from the Asia region. This may be due to the fact that the Google search engine is not used in China, which is a top source country both in terms of arrivals and guest nights. For both the Pacific and America regions, SARMAX models at their levels were the best fit models, for arrivals and guest nights. Google data did not improve the accuracy of forecasting arrivals from ‘other’ regions. SARMA(4,0,3)(1,0,1)12 is the best fit model for arrivals from ‘other’ regions. With regard to guest nights from other regions, the log-transformed model of SARMAX(3,0,3)(1,0,0)12 was the best.

The best models for each region were used to derive monthly forecasts for the first quarter of 2020. The arrival forecasts and guest night forecasts for each region are presented in Tables 9

Table 6. Optimal models.

| Region  | Level/Log | SARMA         | SARMAX         | SARMA         | SARMAX         | SARMA         | SARMAX         |
|---------|-----------|---------------|---------------|---------------|---------------|---------------|---------------|
| Total   | Levels    | (4,0,4)(1,0,1)12 | (4,0,1)(1,0,0)12 | (3,0,2)(1,0,2)12 | (3,0,4)(1,0,2)12 |
|         | Log       | (3,0,4)(1,0,2)12 | (3,0,4)(1,0,2)12 | (3,0,4)(1,0,2)12 | (3,0,4)(1,0,2)12 |
| Europe  | Levels    | (4,0,4)(1,0,0)12 | (3,0,4)(1,0,2)12 | (3,0,2)(1,0,1)12 | (3,0,3)(1,0,2)12 |
|         | Log       | (4,0,2)(1,0,1)12 | (3,0,4)(1,0,2)12 | (1,0,2)(1,0,1)12 | (3,0,3)(1,0,2)12 |
| Asia    | Levels    | (4,0,4)(1,0,0)12 | (3,0,3)(1,0,0)12 | (4,0,4)(0,0,2)12 | (3,0,4)(0,0,2)12 |
|         | Log       | (3,0,3)(2,0,2)12 | (3,0,3)(1,0,2)12 | (4,0,3)(1,0,1)12 | (3,0,3)(2,0,2)12 |
| Pacific | Levels    | (3,0,3)(2,0,2)12 | (3,0,3)(2,0,2)12 | (3,0,2)(2,0,2)12 | (3,0,2)(2,0,2)12 |
|         | Log       | (1,0,1)(1,0,1)12 | (1,0,2)(1,0,1)12 | (1,0,1)(1,0,1)12 | (1,0,1)(1,0,1)12 |
| America | Levels    | (2,0,4)(1,0,2)12 | (2,0,3)(1,0,2)12 | (2,0,2)(1,0,2)12 | (2,0,2)(1,0,2)12 |
|         | Log       | (3,0,3)(1,0,2)12 | (3,0,2)(2,0,2)12 | (1,0,2)(1,0,2)12 | (2,0,2)(1,0,2)12 |
| Other   | Levels    | (4,0,3)(1,0,0)12 | (3,0,3)(1,0,2)12 | (3,0,4)(1,0,0)12 | (2,0,3)(1,0,2)12 |
|         | Log       | (4,0,3)(1,0,0)12 | (3,0,3)(1,0,2)12 | (3,0,4)(1,0,0)12 | (3,0,3)(1,0,0)12 |

Table 7. Forecasting accuracy in international arrivals.

| Region  | Level/Log | MAPE | RMSE  | MAE   |
|---------|-----------|------|-------|-------|
| Total   | Levels    | 8.60 | 8.50  | 19.184| 18.198|
|         | Log       | 11.89| 6.04  | 26.833| 12.099|
| Europe  | Levels    | 14.86| 11.03 | 15.990| 11.964|
|         | Log       | 13.61| 13.19 | 14.522| 14.374|
| Asia    | Levels    | 16.94| 15.64 | 13.186| 12.179|
|         | Log       | 11.82| 13.44 | 10.956| 10.332|
| Pacific | Levels    | 19.24| 18.83 | 2.196 | 2.158 |
|         | Log       | 24.39| 23.98 | 2.710 | 2.680 |
| America | Levels    | 18.46| 18.57 | 2.787 | 2.777 |
|         | Log       | 28.24| 22.07 | 4.055 | 3.323 |
| Other   | Levels    | 28.62| 32.54 | 1.704 | 1.822 |
|         | Log       | 37.39| 42.25 | 1.923 | 2.291 |
and 10, respectively, along with the actual arrivals data obtained from administrative reports. The tables also present percentage deviation values of actual arrivals and guest nights from their forecasts. Interestingly, the tables indicate that arrivals from different regions have started to decline over the months following the onset of the global pandemic, implying heterogeneous travel decisions of tourists from different regions of the world.

When considering total arrivals, it can be seen that the number of actual arrivals is greater than the forecast arrivals in January 2020. A possible reason for this could be international travel postponements due to the Easter Sunday incident in April 2019. Several countries had issued advisories against travelling to Sri Lanka soon after the incident and it took a few months to observe a growth in arrivals following the event.

It is interesting to note that actual figures were well below forecast figures for both arrivals (27%) and guest nights (26%) from the Asia region in February 2020. Outbound travel bans imposed by China due to COVID-19 pandemic could have impacted this decline given that China is a top source country for international tourist arrivals for Sri Lanka. Tourist guest nights from Europe and the Pacific regions have continued above forecast in February 2020, despite the evolving pandemic situation in February 2020. Interestingly, both total international arrivals and guest nights were greater than the forecast values in February 2020, indicating that tourism demand to Sri Lanka had not begun to decline in February 2020 amidst the COVID-19.

Notable negative deviations between forecasts and actual values of arrivals and guest nights can be observed from March 2020. On 18th March, Sri Lanka suspended all international passenger arrivals through the international airport. Overall, there was a 59% and 56% decline in actual arrivals and guest nights, respectively, compared to forecast values in March 2020. Arrivals from ‘Other’ regions

### Table 8. Forecasting accuracy in international guest nights.

| Region | Level/Log | MAPE | RMSE | MAE |
|--------|-----------|------|------|-----|
|        | SARMA     | SARMAX | SARMA | SARMAX | SARMA | SARMAX |
| Total  | Levels    | 6.01  | 6.38 | 165,149 | 164,922 | 126,989 | 132,261 |
|        | Log       | 11.02 | 5.99 | 333,617 | 139,449 | 268,336 | 115,526 |
| Europe | Levels    | 13.65 | 13.56 | 131,019 | 134,396 | 117,359 | 122,563 |
|        | Log       | 10.25 | 8.63 | 167,726 | 142,749 | 128,546 | 107,672 |
| Asia   | Levels    | 23.45 | 29.51 | 156,754 | 192,921 | 145,278 | 181,580 |
|        | Log       | 8.89  | 9.59 | 66,230  | 65,320  | 56,033  | 56,388  |
| Pacific| Levels    | 26.73 | 26.53 | 50,922  | 50,537  | 42,506  | 42,109  |
|        | Log       | 30.94 | 31.29 | 55,406  | 55,618  | 47,708  | 48,005  |
| America| Levels    | 16.23 | 16.19 | 40,485  | 40,374  | 27,955  | 27,712  |
|        | Log       | 18.84 | 18.03 | 49,231  | 47,288  | 33,255  | 31,494  |
| Other  | Levels    | 27.83 | 35.51 | 19,361  | 19,361  | 14,781  | 16,805  |
|        | Log       | 36.04 | 16.47 | 12,953  | 12,953  | 17,166  | 10,442  |

### Table 9. Actual and forecasted international arrivals in 2020Q1.

| Month    | Value | Europe | Asia | The Pacific | America | Other | Total* |
|----------|-------|--------|------|-------------|---------|-------|--------|
| January 2020 | Forecast | 124,496 | 79,255 | 4918 | 14,323 | 5441 | 195,303 |
|          | Actual  | 113,601 | 83,009 | 9192 | 14,224 | 2008 | 228,434 |
|          | % Deviation | −9 | 5 | 87 | −1 | −63 | 17 |
| February 2020 | Forecast | 135,858 | 75,191 | 4689 | 14,837 | 4877 | 191,329 |
|          | Actual  | 135,904 | 54,754 | 9900 | 14,346 | 3709 | 207,507 |
|          | % Deviation | 0 | −27 | 111 | −3 | −24 | 8 |
| March 2020 | Forecast | 121,940 | 78,942 | 4680 | 17,735 | 5768 | 174,775 |
|          | Actual  | 42,804 | 21,404 | 2895 | 3273 | 994 | 71,370 |
|          | % Deviation | −65 | −73 | −38 | −82 | −83 | −59 |
| April 2020 | Forecast | 79,035 | 62,070 | 4594 | 12,319 | 5288 | 126,765 |
|          | Actual  | 0 | 0 | 0 | 0 | 0 | 0 |
|          | % Deviation | −100 | −100 | −100 | −100 | −100 | −100 |

*The column total denotes arrivals at aggregated level. It does not imply the total of the individual columns on arrivals from different regions.
and America showed the highest negative deviation between actual and forecast figures in March 2020. The least decline in both arrivals and guest nights compared to forecast was for the European region in that month. April 2020 recorded zero international tourist arrivals for the first time in Sri Lanka’s tourism history.

We employed the estimates from the tourist expenditure survey to calculate the loss in tourist expenditure based on the differences between the forecast and actual values. The loss of tourism expenditure based on arrivals forecasts and guest night forecasts are presented in Table 11. It can be seen that there is a loss in tourist expenditure in January and February, when the estimation is done based on arrival forecasts (Column 2). This can be attributed to the decline in arrivals from Europe compared to forecast values in January 2020. The decline in arrivals from Asia compared to forecast due to COVID-19 might likely explain much of the loss of tourist expenditure in February 2020. When tourist guest night forecasts are considered for the estimations, we see a gain in tourist expenditure both in January and February (Column 3), which contrasts with the results for arrivals discussed previously. There are two likely explanations for this. Firstly, guest nights from the European region show a positive deviation from the forecast in both the months (Table 10) and this may explain the gain in tourist expenditure. For March, both the calculations indicate a significant loss of tourist expenditure due to the pandemic.

As we highlighted previously, the guest night series encompasses information relating to both arrivals and durations of stay. The different impacts of timeseries variables in estimating the affects in this case justify the benefits of using more than one timeseries variable in forecasting tourism demand. Overall, the estimated loss in tourist expenditure to the economy of Sri Lanka is 40 percent in the first quarter of 2020 when the estimation is based on arrival forecasts. When guest night forecasts are employed in estimations, the overall loss in tourist expenditure is 28% in the first quarter of 2020. This provides important implications relating to resilience in the hospitality sector towards shocks and how tourism demand from different regions responded to the pandemic.

| Table 10. Actual and forecasted international guest nights in 2020Q1. |
|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Month             | Value           | Europe          | Asia            | The Pacific     | America         | Other           | Total*          |
| January 2020      | Forecast        | 1,284,652       | 572,672         | 159,967         | 160,808         | 41,909          | 2,033,957       |
|                   | Actual          | 1,397,292       | 619,801         | 129,607         | 166,421         | 17,871          | 2,467,087       |
|                   | % Deviation     | 9               | 8               | −19             | 3               | −57             | 21              |
| February 2020     | Forecast        | 1,335,409       | 551,479         | 125,329         | 181,151         | 30,566          | 1,953,218       |
|                   | Actual          | 1,671,619       | 408,830         | 139,590         | 167,848         | 33,010          | 2,241,076       |
|                   | % Deviation     | 25              | −26             | 11              | −7              | 8               | 15              |
| March 2020        | Forecast        | 1,091,711       | 553,996         | 164,871         | 205,938         | 40,004          | 1,744,501       |
|                   | Actual          | 526,489         | 159,817         | 40,820          | 38,294          | 8847            | 770,796         |
|                   | % Deviation     | −52             | −71             | −75             | −81             | −78             | −56             |
| April 2020        | Forecast        | 638,650         | 480,634         | 168,822         | 141,034         | 37,986          | 1,237,507       |
|                   | Actual          | 0               | 0               | 0               | 0               | 0               | 0               |
|                   | % Deviation     | −100            | −100            | −100            | −100            | −100            | −100            |

*The column total denotes arrivals at aggregated level. It does not imply the total of the individual columns on arrivals from different regions.

| Table 11. Estimates of tourist expenditure losses based on forecasted values for 2020Q1. |
|-------------------------|---------------------------------|---------------------------------|---------------------------------|
| Month                   | Percentage in tourist expenditure loss based on arrival forecasts for each month | Percentage in expenditure loss based on guest night forecasts for each month |
| January                 | −1                              | 1                              |
| February                | −2                              | 4                              |
| March                   | −18                             | −16                            |
| April                   | −19                             | −18                            |
| Total loss in expenditure in the quarter | −40                             | −29                            |
5. Discussion and conclusion

The study aimed at producing disaggregated forecasts for international tourist arrivals and guest nights in Sri Lanka by incorporating Google search data into traditional forecasting methods. Regionally disaggregated data were incorporated in ARMA, ARMAX, SARMA and SARMAX models. The paper adds empirical evidence to the growing view that Google trend data can improve tourism forecasting accuracy, both with regard to arrivals and guest nights. Inclusion of Google has increased the forecasting accuracy in 4 regions out of 6 for the arrival models. Google data improved the accuracy of 5 out of 6 models which forecast guest nights.

With the help of estimated forecasts, the paper makes an important attempt in quantitatively estimating the foregone international tourist expenditure due to COVID-19 pandemic in Sri Lanka. The results point out to a conclusion that foregone tourist expenditure is different between the two timeseries used for forecasting. When tourist arrivals are used in estimation, the foregone tourist expenditure is 40% when compared with forecasts. When the guest nights are used for the estimates, it indicates a gain in tourist expenditure in January and February and overall lost amounts to only 28%.

The findings of the results point out the importance of employing disaggregated data in forecasting. Geographically disaggregated forecasts essentially reveal the magnitudes of changes in tourism demand trends in specific regions. In our paper we use this information effectively in understanding the geographically different responses to a global shock based on a single country context. Further, we employ monthly data in timeseries models. This has provided diverse time-dependent information on tourism demand.

Use of multiple series for forecasting reveals important information on tourism demand which may be of varying importance to different stakeholders. Our estimates showed that the guest night variable captures the economic effects of COVID-19 shock more comprehensively, when compared with the international arrivals. The guest night variable interestingly incorporates duration of stay, in addition to number of arrivals and has helped in capturing the dynamics during different periods of stays from different regions. For future research, we strongly recommend deviating away from using single timeseries to include two or more timeseries to capture different aspects of tourism demand. Further, we suggest the use of geographically and temporally disaggregated data to reveal diverse information on tourism demand, which are more useful for decision makers at all levels.

There are several limitations of this paper, which can potentially be overcome by future tourism forecasting research. Firstly, our analysis does not take into account search data from Baidu search engine and we only focus on Google search data. Google is identified as the most widespread search engine at global scale. However, this might have led to omit important travel decision-making behaviours of Chinese tourists, as China accounts for nearly 11% of international arrivals to Sri Lanka as in 2018 (SLTDA, 2018). Follow-up research ideally can look into the ways of integrating both Google and Baidu search engine data when predicting tourism demand from different regions of the world. Secondly, the current analysis is confined to timeseries econometric methods, with Google search data and a dummy variable to represent structural changes as explanatory variables. We had to omit other important socio-demographic and economic variables due to the absence of monthly data for the variables. However, methodological advancement with the use of such variables as explanatory variables is possible for destinations where comprehensive economic and tourism-related data are available. Further, machine learning approaches can be supplemented with econometric approaches to reveal their performance in improving forecasting accuracy based on temporally and geographically disaggregated data. Finally, our findings are based on Sri Lanka as a test case and much potential exists for assessing the differential outcomes in predicting tourism demand in multiple tourism regions and destinations using the proposed methodology.
Acknowledgements

We thankfully acknowledge the proofreading support provided by the Griffith Institute of Tourism (GIFT), Griffith University, Australia.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Bandara, J. S. (2019). Tourism in post-conflict development: Making use of new opportunities in Sri Lanka. In D. Weerakoon & S. Jayasooriya (Eds.), Managing domestic and international challenges and opportunities in post-conflict development: Lessons from Sri Lanka (pp. 173–194). Springer Nature Singapore Pte Ltd.

Boone, L. (2020). Tackling the fallout from COVID-19. In R. Baldwin & B. W. D. Mauro (Eds.), Economics in the time of COVID-19 (pp. 37–44). CEPR Press.

CBSL. (2020). Annual report 2019. Central Bank of Sri Lanka.

Dinis, G., Costa, C., & Pacheco, O. (2016). The use of Google trends data as proxy of foreign tourist inflows to Portugal. International Journal of Cultural and Digital Tourism, 3(1), 2241–9705.

Dinis, G., Costa, C., & Pacheco, O. (2017). Similarities and correlation between resident tourist overnights and Google trends information in Portugal and its tourism regions. Tourism and Management Studies, 13(3), 15–22. https://doi.org/10.18089/tms.2017.13302

Ghalehkhondabi, I., Ardjomand, E., Young, W. A., & Weckman, G. R. (2019). A review of demand forecasting models and methodological developments within tourism and passenger transportation industry. Journal of Tourism Futures, 5(1), 75–93. https://doi.org/10.1108/JTF-10-2018-0061

Goh, C., & Law, R. (2011). The methodological progress of tourism demand forecasting: A review of related literature. Journal of Travel & Tourism Marketing, 28(3), 296–317. https://doi.org/10.1080/10548408.2011.562856

Jiao, E. X., & Chen, J. L. (2019). Tourism forecasting: A review of methodological developments over the last decade. Tourism Economics, 25(3), 469–492. https://doi.org/10.1177/1354816618812588

Khaidi, S. M., Abu, N., & Muhammad, N. (2019). Tourism demand forecasting—a review on the variables and models. Paper presented at the Journal of Physics: Conference Series.

Li, G. (2009). Tourism demand modeling and forecasting: A review of literature related to Greater China. Journal of China Tourism Research, 5(1), 2–40. https://doi.org/10.1080/19388160802711386

Park, S., Lee, J., & 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. https://doi.org/10.1080/10548408.2016.1170651

Samaranayake, H. M. S., Lantra, N., & Jayawardena, C. (2013). Forty six years of organised tourism in Sri Lanka (1966–2012). Worldwide Hospitality and Tourism Themes, 5(1), 423–444. https://doi.org/10.1108/WHATT-05-2013-0026

SLTDA. (2017). Survey of departing foreign tourists from Sri Lanka. Sri Lanka Tourism Development Authority.

SLTDA. (2018). Annual Statistical report 2018. Sri Lanka Tourism Development Authority.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting: A review of recent research. Tourism Management, 29(2), 203–220. https://doi.org/10.1016/j.tourman.2007.07.016

Song, H., Qiu, R. T., & Park, J. (2019). A review of research on tourism demand forecasting. Annals of Tourism Research, 75, 338–362. https://doi.org/10.1016/j.annals.2018.12.001

Thushara, S. C., Su, J. J., & Bandara, J. S. (2019). Forecasting international tourist arrivals in formulating tourism strategies and planning: The case of Sri Lanka. Cogent Economics and Finance, 7(1). https://doi.org/10.1080/23322039.2019.1699884

Turner, L. W., & Witt, S. F. (2001). Forecasting tourism using univariate and multivariate structural time series models. Tourism Economics, 7(2), 135–147. https://doi.org/10.5367/000000001101297775

UNWTO. (2020, March 13). Tourism and COVID-19 [Press release].

Volchek, K., Liu, A., Song, H., & Buhaldis, D. (2019). Forecasting tourist arrivals at attractions: Search engine empowered methodologies. Tourism Economics, 25(3), 425–447. https://doi.org/10.1177/1354816618811558

Wickramasinghe, K. (2011). Post-conflict tourism development in Sri Lanka. In Sri Lanka state of the economy 2011: Post-conflict growth and recovery (pp. 151–157). Institute of Policy Studies of Sri Lanka.

Witt, C. A., & Witt, S. F. (1995). Forecasting tourism demand: A review of empirical research. International Journal of Forecasting, 11(3), 447–475. https://doi.org/10.1016/0169-2070(95)00591-7

World Bank. (2020, June). Global economic Prospects. https://doi.org/10.1596/978-1-4648-1553-9. License: Creative Commons Attribution CC BY 3.0 IGO.

Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. International Journal of Contemporary Hospitality Management, 29(1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249