Masters Programmes

Assignment Cover Sheet

Submitted by: *<2096170>*

Date Sent: 01/04/2021

Module Title: Big Data Analytics

Module Code: IB9CSB

Date/Year of Module: 2020/2021

Submission Deadline: Thursday 1st April at 20:00

Word Count: 1858

Number of Pages: 9

Question: *[e.g. question number/title, or description of assignment]*

"I declare that this work is entirely my own in accordance with the University's Regulation 11 and the WBS guidelines on plagiarism and collusion. All external references and sources are clearly acknowledged and identified within the contents.

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done it may result in me being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work."

Adaptive nowcasting of tourism statistics with Google Trends

Abstract

Tourism is an essential driver of the UK economy. However, as is the case with most government released figures, tourism reports are published with a significant time delay. Accordingly, many researchers have shown the value of search engine data in nowcasting of both offline human behaviours and economic indicators. Here, I attempt to add to this body of research by analysing if *Google Trends* data on British landmarks can provide estimates of current tourist numbers and economic activity in the UK. Utilising 191 datapoints, across a 16-year period, I find minor changes in the mean absolute percentage error (MAPE) of nowcasting models with and without *Google Trends* data. Specifically, MAPE changes between -0.12% and 0.92%. I therefore conclude that there is a lack of evidence that *Google Trends* data on British landmarks can be used to predict present tourism statistics.

1 Introduction

There are substantial time delays associated with the reporting of government figures despite their value in decision making. In comparison, data generated from everyday use of technology is readily available. As such, this emergence of “big data” has stimulated research on the value of online data in providing estimates of present human behaviours (nowcasting). For one, silos of data generated by user queries on the Google search engine have been utilised to nowcast museum visits (Botta et al, 2020), unemployment claims (Choi and Varian, 2012), and influenza outbreaks (Preis and Moat, 2014). Furthermore, strong links have been found between what people Google search online and the performance of feature films, video games and charting songs (Goel et al, 2010). Indeed, research extends beyond the realm of Google searches; Miller et al (2020) demonstrated the value of real-time aircraft statistics (ADS-B) in estimating aircraft flight volumes and current economic activity.

In this paper, I analyse the extent to which Google trends data can improve nowcasts of tourism statistics. Focus is placed on the UK tourism industry, a significant driver of the British economy that made a notable 10.9% contribution to GDP in 2019 (Knoema, 2019). I expect to find greater landmark search volume in the same month as greater tourism numbers and earnings. This is because I believe tourists conduct Google searches on British landmarks before visiting the UK. The potential value of this finding could be far reaching. Specifically, policy and investment decisions made by bodies such as the *Tourism Industry Council* have a huge impact on the UK economy, affecting a plethora of small businesses and workers. These decisions require accurate and up-to-date information, but currently

tourism statistics are accumulated and reported with significant time delays. Faster statistics could therefore aid important government decision making.

2 Method and Results

I conduct analysis on monthly datasets for the period January 2004 to November 2019, removing periods under the COVID-19 pandemic. The reason for this is to limit the analysis period to normal times, as the tourism industry was significantly impacted by the COVID-19 pandemic. I retrieve data from the *Office of National Statistics* for both monthly overseas visits to the UK and monthly UK tourism earnings (ONS, 2021). This is used in conjunction with *Google Trends* data for the 5 most popular UK tourist attractions according to TripAdvisor in 2019 (Mirror, 2019). The tourist attractions are: “Tower of London”, “Stonehenge”, “Giant’s Causeway”, “London Eye”, and “Royal Mile”. *Google Trends* search topics are chosen for the analysis over search terms, where topics include all search terms that share the same concept, in any language. This is because most visits to the UK originate from countries that do not use English as their primary language (ONS, 2015), as such topic search data is a better representation of actual online interest in British landmarks. Moreover, the *Google Trends* service does not provide absolute search volume, instead normalising the data by geography and time range. I therefore request search data for all topics simultaneously to ensure volumes are comparable.

Analysis is conducted on both the individual and collective search volume of the 5 chosen landmarks. The latter of which involves adding search volume for all 5 landmarks to form the “total searches”. This allows for better differentiation of whether individual terms or the collective search volume are responsible for patterns I discover. Furthermore, graphical inspection of the datasets indicates the presence of trends and seasonality. Accordingly, I remove these time-series features from all datasets before conducting my analysis.

A preliminary Shapiro-Wilk test for normality is run. Adjusting with FDR correction, I find that only the offline tourism data is normally distributed (*Shapiro-Wilk, FDR* correction applied; Visits: $w = 0.996$, $p = 0.955$; Earnings: $w = 0.992$, $p = 0.493$; Tower of London: $w = 0.568$, $p < 0.001$; Stonehenge: $w = 0.962$, $p < 0.001$; Giant’s Causeway: $w = 0.903$, $p < 0.001$; London Eye: $w = 0.910$, $p < 0.001$; Royal Mile: $w = 0.895$, $p < 0.001$; Total searches: $w = 0.848$, $p < 0.001$; $df = 178$ for all).

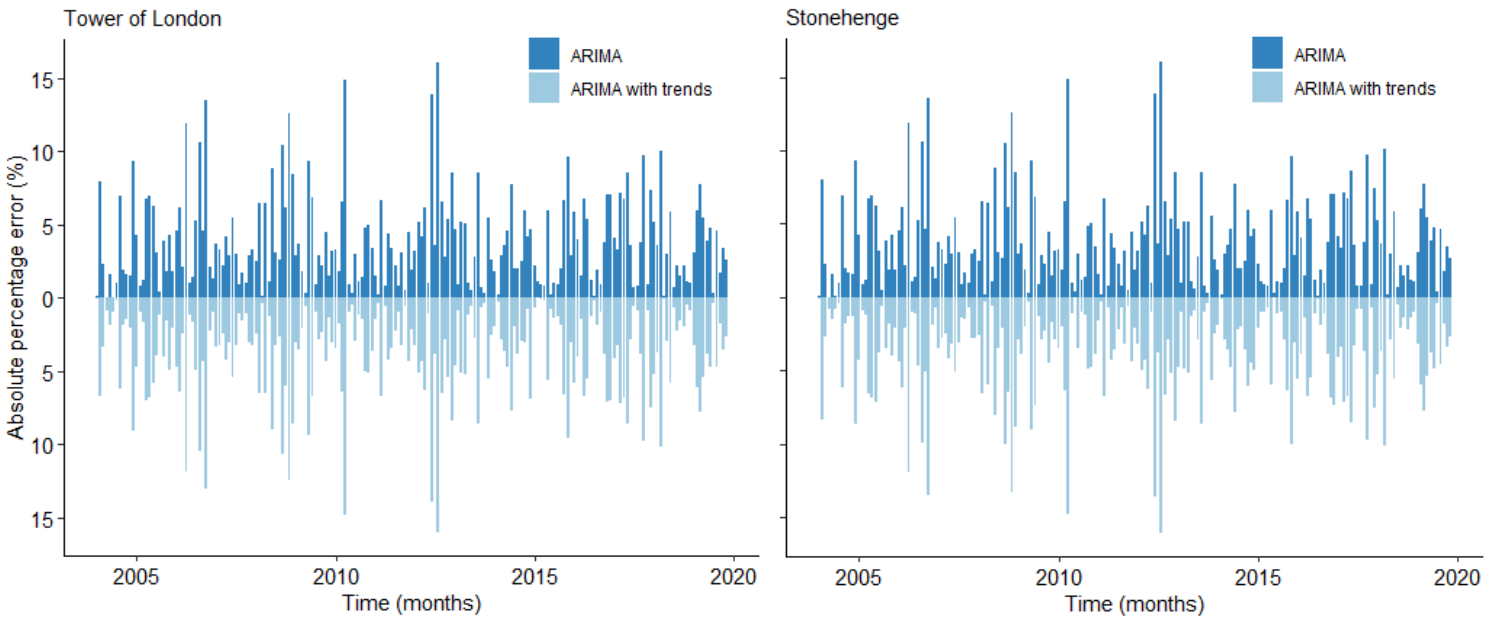
As such, I conduct a Kendall’s Tau correlation analysis, again adjusting with FDR correction. The test indicates no statistically significant correlations between offline tourism **visitor numbers** and online *Google Trends* searches for British landmarks (*Kendall Tau* for visits*, FDR* correction applied; Tower of London $\boldsymbol{\tau} = 0.019$, $p = 0.862$; Stonehenge $\boldsymbol{\tau} = -0.057$, $p = 0.822$; Giant’s Causeway $\boldsymbol{\tau} = 0.023$, $p =$

0.862; London Eye τ = 0.058, p = 0.822; Royal Mile τ = 0.010, p = 0.862; Total searches τ = -0.009, p = 0.862; N = 178 for all). Similarly, I observe no statistically significant correlations between offline tourism **earnings** and searches for British landmarks (*Kendall Tau* for earnings*, FDR* correction applied; Tower of London τ = 0.021, p = 0.870; Stonehenge τ = 0.063, p = 0.682; Giant's Causeway τ = 0.019, p = 0.870; London Eye τ = -0.018, p = 0.870; Royal Mile τ = 0.009, p = 0.870; Total searches τ = 0.067, p = 0.682; N = 178 for all).

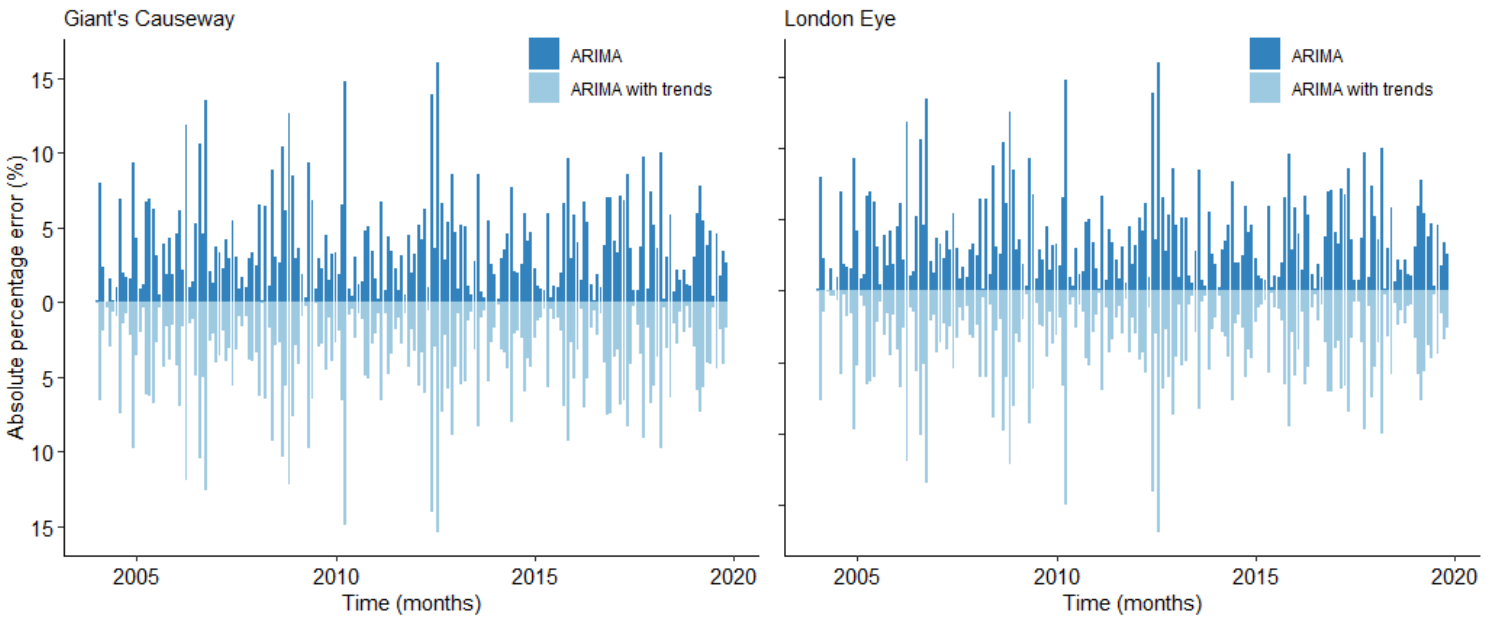
Despite the lack of correlation, I construct autoregressive integrated moving average (ARIMA) models of tourism statistics to examine if there are nuances in the data that may improve my adaptive nowcasts. I difference the data once and utilise one autoregressive term to form models individually described as ARIMA(0,1,1). I compare baseline models and 'trends-fitted' models, the latter of which involves adding *Google Trends* time series data to the respective ARIMA(0,1,1) as an external regressor.

Tourism visitor numbers. Figure 1 depicts our adaptive nowcasting results for tourism visitor numbers. Adding *Google Trends* data does not yield any significant percentage improvements to the baseline model's mean absolute percentage error (Tower of London 0.459%; Stonehenge -0.045 %; Giant's Causeway 0.136 %; London Eye 0.620 %; Royal Mile -0.116 %; Total searches 0.035 %).

Tourism earnings. Figure 2 depicts our adaptive nowcasting results for the tourism sector's monthly earnings. Similar to the tourism case, additional *Google Trends* data does not yield significant improvements to the baseline model's mean absolute percentage error (Tower of London 0.214%; Stonehenge 0.595%; Giant's Causeway 0.076%; London Eye 0.042%; Royal Mile 0.034%; Total searches 0.917%)


Tower of London
ARIMA
ARIMA with trends
Absolute percentage error (%)
15
10
5
0
5
10
15
2005
2010
2015
2020
Time (months)

Stonehenge
ARIMA
ARIMA with trends
2005
2010
2015
2020
Time (months)


Giant's Causeway
ARIMA
ARIMA with trends
Absolute percentage error (%)
15
10
5
0
5
10
15
2005
2010
2015
2020
Time (months)

London Eye
ARIMA
ARIMA with trends
2005
2010
2015
2020
Time (months)

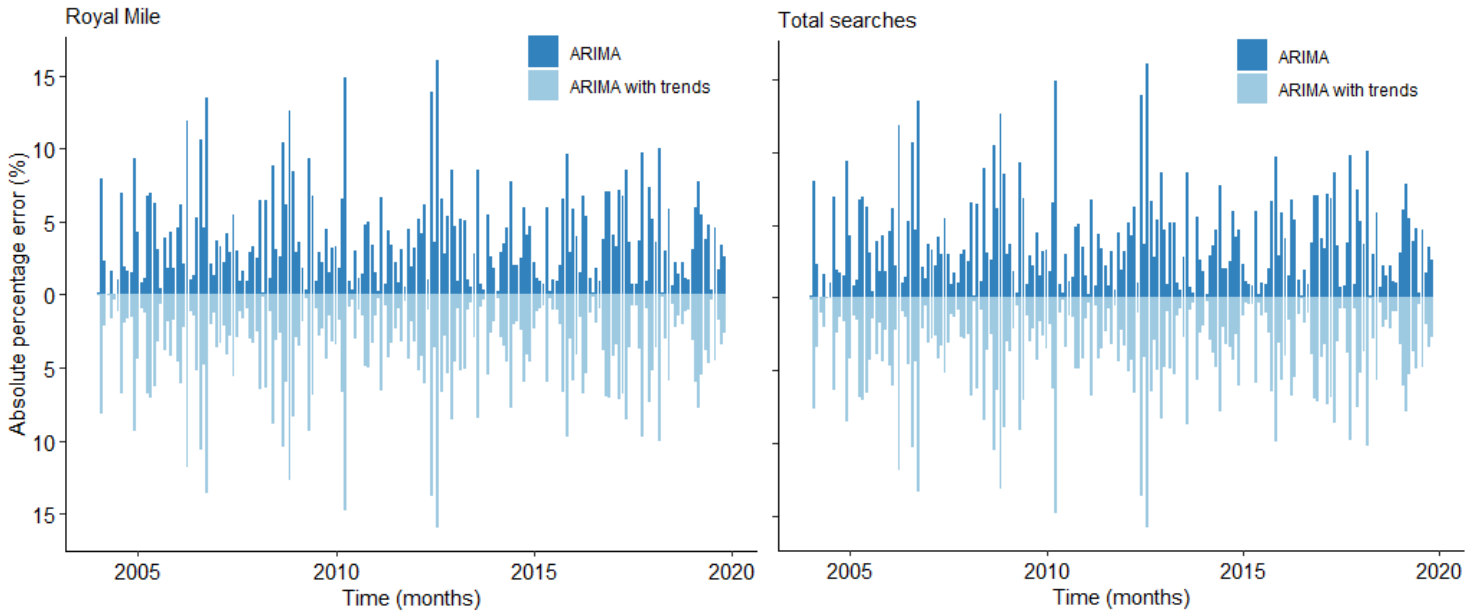
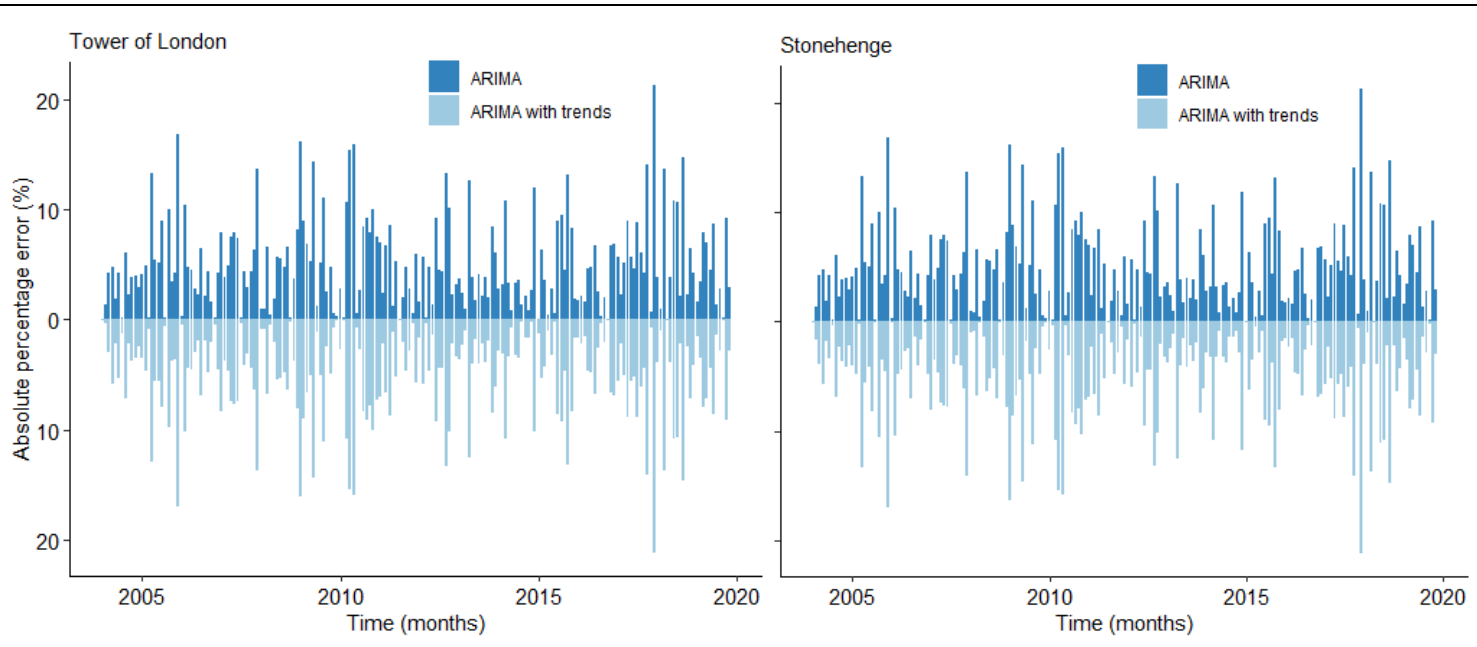


Figure 1. Rapid estimates of monthly overseas visits to the UK. Absolute percentage error for visitor ARIMA models with (light blue) and without (dark blue) *Google Trends* data as an additional predictor. For all months, both ARIMA models display similar absolute percentage errors, indicating that adding trends data does not improve the adaptive nowcasts. Mean absolute percentage error changes are indicative of this fact (Tower of London 0.459%; Stonehenge -0.045 %; Giant's Causeway 0.136 %; London Eye 0.620 %; Royal Mile -0.116 %; Total searches 0.035 %).



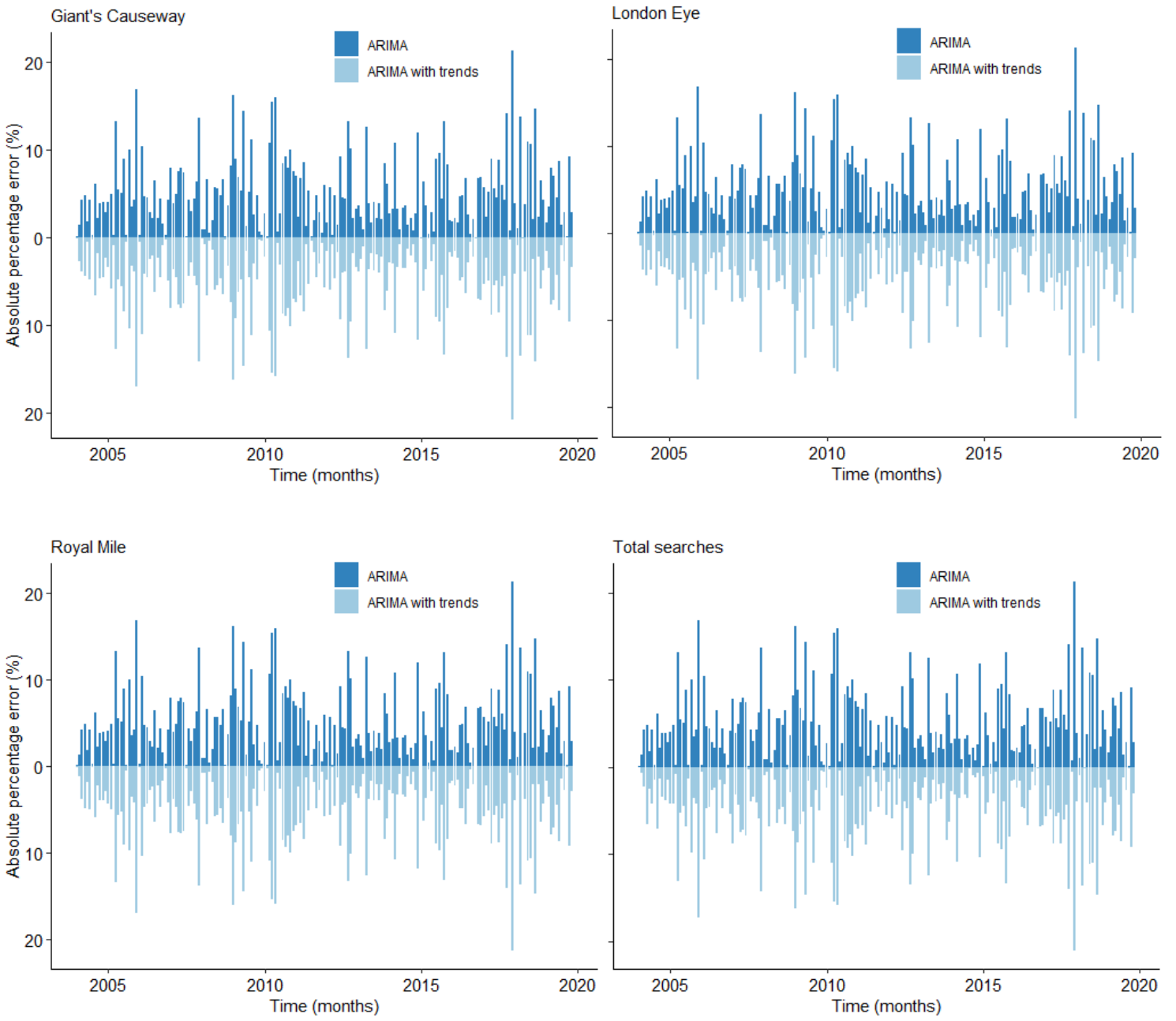


Figure 2. Rapid estimates of monthly UK tourism earnings. Absolute percentage error for ARIMA earnings models with (light blue) and without (dark blue) *Google Trends* data as an additional predictor. For all months, both ARIMA models display similar absolute percentage errors, indicating that adding trends data does not improve the adaptive nowcasts. Mean absolute percentage error changes are indicative of this (Tower of London 0.214%; Stonehenge 0.595%; Giant's Causeway 0.076%; London Eye 0.042%; Royal Mile 0.034%; Total searches 0.917%)

3 Discussion

To summarise, I attempted to answer the question: can Google Trends data on British landmarks provide estimates of current tourist numbers and economic activity in the UK? This was motivated by a plethora of research that utilises online search engine data in nowcasting of offline human behaviours. My research attempted to apply key themes from this body of research to the UK tourism industry, an industry that contributes significantly to the British economy but is still victim to the substantial time delays associated with governmental reporting. Accordingly, nowcasting models for tourism earnings and visits were produced, with and without Google Trends data as an external regressor. It was clear that Google trends data for the 5 chosen landmarks did not significantly improve the adaptive nowcasting models. Specifically, small changes in the mean absolute percentage error (MAE) of less than 1% were observed.

This is an unexpected result for two reasons. Firstly, I had initially hypothesised that tourists conduct *Google* searches for landmarks before deciding to holiday in the UK. As such greater search volume would indicate both larger tourist numbers and earnings for the tourism sector. Secondly, Choi and Varian (2012) have previously found a link between Hong Kong's search category query index and visitor arrival data to the country. Thus, I believed that extending this research by focusing on a particular area of tourism would bring about similar, interesting insights. This appears to not be the case. *Google* search volume for a particular subset of the tourism industry may not be as good a proxy for UK tourism statistics. This could be attributed to one of many reasons. For one, the extent to which people are searching for landmarks but not visiting the UK. For example, the Big Ben clock renovations would likely cause "Big Ben" search volume to surge. Axiomatically, we would not expect tourists to flock to the UK after searching for this. However, events of this type are rare and should not have significant impact on general patterns but rather explain large spikes in the nowcast errors. Instead consider the extent to which people are visiting the UK and not searching for landmarks before their visits. This could be due to those on business trips or recurring visits. Visitors of this nature will likely not have interest in British landmarks. Indeed, other proxies such as UK hotel or airport searches may be better estimates of business and recurring visitors. Finally, considerations must be made of the noise in the *Office of National Statistics* and *Google Trends* data.

Going forward, I suggest further investigation into this application of nowcasting. Given the prevalence of tourism in the UK economy, valuable insights could be discovered if one conducts analysis beyond the bounds of this investigation. Indeed, my results provide a lack of evidence but do not disprove my initial hypothesis. Potential ways in which the investigation could be improved are adding more landmarks to the analysis, implementing sliding training windows into the ARIMA models, and considering business or recurring visitors which I previously outlined above.

References

Botta, F., Preis, T. & Moat, H.S. (2020) 'In search of art: rapid estimates of gallery and museum visits using Google Trends'. EPJ Data Science 9(1), article number: 4. DOI: 10.1140/epjds/s13688-020-00232-z

Choi, H & Varian, H. (2012) 'Predicting the Present with Google Trends'. *The Economic Record*, 88, pp. 2-9. DOI: 10.1111/j.1475-4932.2012.00809.x

Delahaye, J. (2019) 'UK's top 10 most popular attractions from 2019 according to TripAdvisor'. *Mirror* [online] 17 December. Available from <https://www.mirror.co.uk/travel/uk-ireland/uks-top-10-most-popular-21115577> [23 February 2021]

Goel, S., Hofman, M.J., Lahaie, S., Pennock, M.D. & Watts, J.D. (2010) 'Predicting consumer behavior with Web search'. *Proceedings of the National Academy of Sciences* 107(41) pp.17486-17490. DOI: 10.1073/pnas.1005962107

Google trends [online]. Available from <https://trends.google.com> [20 March 2021]

Knoema. (2019) *United Kingdom - Contribution of travel and tourism to GDP as a share of GDP* [online]. Available from: <https://knoema.com/atlas/United-Kingdom/topics/Tourism/Travel-and-Tourism-Total-Contribution-to-GDP/Contribution-of-travel-and-tourism-to-GDP-percent-of-GDP> [25 February 2021]

Miller, S., Moat, H.S. & Preis, T. (2020) Using aircraft location data to estimate current economic activity. *Scientific Reports* 10(1), article number: 7576. DOI: 10.1038/s41598-020-63734-w

Office for National Statistics (2016) 'Top 50 countries visited by UK residents and top 50 overseas countries who visited the UK' [online]. Available from <https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/adhocs/006046top50countriesvisitedbyukresidentsandtop50overseascountrieswhovisitedtheuk> [23 March 2021]

Office for National Statistics (2020a) 'OS visits to the UK: Thousands - SA' [online]. Available from <https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/timeseries/gmat/ott> [23 February 2021]

Office for National Statistics (2020b) 'OS visits to UK:Earnings: £ Millions-(Cur.Price-SA)' [online]. Available from

<https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/timeseries/gmaz/ott> [23 February 2021]

Preis T. & Moat H.S. (2014) ‘Adaptive nowcasting of influenza outbreaks using *Google* searches’. *Royal Society Open Science* 1(2), article number: 140095. DOI: 10.1098/rsos.140095