



In search of art: rapid estimates of gallery and museum visits using Google Trends

Federico Botta^{1,2*} , Tobias Preis^{1,3}  and Helen Susannah Moat^{1,3} 

*Correspondence:

federico.botta@wbs.ac.uk

¹Data Science Lab, Warwick Business School, University of Warwick, Coventry, UK

²Department of Computer Science, University of Exeter, Exeter, UK
Full list of author information is available at the end of the article

Abstract

Measuring collective human behaviour has traditionally been a time-consuming and expensive process, impairing the speed at which data can be made available to decision makers in policy. Can data generated through widespread use of online services help provide faster insights? Here, we consider an example relating to policymaking for culture and the arts: publicly funded museums and galleries in the UK. We show that data on *Google* searches for museums and galleries can be used to generate estimates of their visitor numbers. Crucially, we find that these estimates can be generated faster than traditional measurements, thus offering policymakers early insights into changes in cultural participation supported by public funds. Our findings provide further evidence that data on our use of online services can help generate timely indicators of changes in society, so that decision makers can focus on the present rather than the past.

Keywords: Nowcasting; Faster indicators; Online data; Culture; Mobility

1 Introduction

Good decisions often require accurate and up-to-date information on the current state of the world. Yet traditional approaches to measuring human behaviour and wellbeing at scale, such as surveys, can be both time-consuming and expensive. With our everyday interactions with the Internet, mobile phones and other large networked systems now generating vast volumes of data on collective human behaviour, this data landscape has however begun to change. These novel data streams have facilitated new research into human behaviour at scale [1–6], offering insights into human mobility patterns [7–15], epidemics [16–21], economic and consumer decision making [22–31] and our social rhythms and interactions [32–37].

In contrast to data from surveys, data from many of these newer sources is available with little to no delay [5, 29]. A key example is data generated through searches for information online, using search engines such as *Google* [5, 16–20, 22–29]. In this paper, we investigate whether search data can provide behavioural insights that are of value for policymakers in the domain of culture and the arts.

In the UK, access to national museums such as the *British Museum*, the *Tate Modern*, the *Natural History Museum* and the *V&A* has been free since 2001. Instead of relying on entrance fees for their permanent exhibitions, these museums and art galleries receive gov-

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

ernmental funding from the *Department for Digital, Culture, Media, and Sport* (DCMS). The goal of this policy was to boost participation in cultural activities. Ongoing evaluation of whether this policy has been successful requires continuous data, to verify that visitor numbers are still high and to alert policymakers to problems with their investments if not.

In recent years, DCMS has made data on visitor numbers available with at least a month's delay. Key policymakers receive these figures a maximum of twenty-four hours ahead of their official release. Here, we hypothesise that people who wish to visit a museum or gallery may also be likely to search on *Google* for information about the museum or gallery around the time of their visit. As data on the collective volume of *Google* searches for a given term or topics is made available publicly with near to no delay, we seek to determine whether this data might allow us to generate much faster indicators of visitor numbers, thus giving policymakers early insights into the performance of their museums and galleries. In our analyses, we aim to exploit the rapid availability of data on online searches, while bearing in mind the problems that previous research has shown can arise if this data is not treated with appropriate caution and consideration [19, 20, 27].

2 Data

We retrieve data on the numbers of visitors to museums and galleries sponsored by DCMS. This dataset is released through the official DCMS website and is freely accessible [38]. Until Spring 2019, the dataset was updated monthly on the first Thursday of each month. The dataset contains the monthly number of visitors to the museums and galleries, with data available from April 2004 onwards. Each museum and gallery records the number of visits to their site in a variety of ways, such as using sensors on the doors. The museums then provide these figures to DCMS. The Department releases the visitor numbers as official statistics on a monthly basis with a delay of one month (or more recently, one quarter). For example, figures released at the beginning of February 2019 reported on visits in December 2018. It is important to note that the official figures are occasionally subject to changes, due to further analysis performed by DCMS, or to correct for previous errors. This means that figures on the number of visits in December 2018, first released in February 2019, could be updated in new releases by DCMS even after February 2019.

We also obtain monthly time series reflecting the level of interest in each museum or gallery on *Google*, using the *Google Trends* service. *Google Trends* offers data on the volume of *Google* searches for a specific search term, such as *Tate Modern*, or, if preferred, for a related topic. While data on searches for a search term would represent the volume of queries for the exact string *Tate Modern*, data on searches for the corresponding topic may reflect searches for a variety of terms related to the *Tate Modern*, such as the names of artists exhibiting their work there. To cover this broader range of searches, here we retrieve search data for the topics relating to the museums and galleries in our analysis.

We restrict our *Google Trends* request to data on searches made in the United Kingdom. We retrieve data from January 2004 onwards, the earliest date for which *Google* search data is available. The *Google Trends* interface provides data for time periods of this length at monthly granularity. We request *Google Trends* data for each museum or gallery individually. *Google Trends* data is normalised, such that the highest search volume for each museum topic in the time period specified is represented as 100, and all other data points are scaled to integer values up to 100. A *Google Trends* value of 100 may therefore represent a different level of search volume for different museums. For each museum, the final

Google Trends data we retrieve provides an indication of changes in search volume for the museum over time.

We retrieve visitor number and *Google Trends* data on 16 DCMS-sponsored museum and gallery institutions or groups. For museum groups which have several museums in the UK, such as the *Science Museum* group, we consider the total number of visitors to the group of museums, and retrieve *Google Trends* data for the topic relating to the name of the group (here, *Science Museum*). In the Additional file 1, we show that we find similar results if we change this approach and only consider visitor numbers at the main *Science Museum* site in South Kensington, London (Fig. S2).

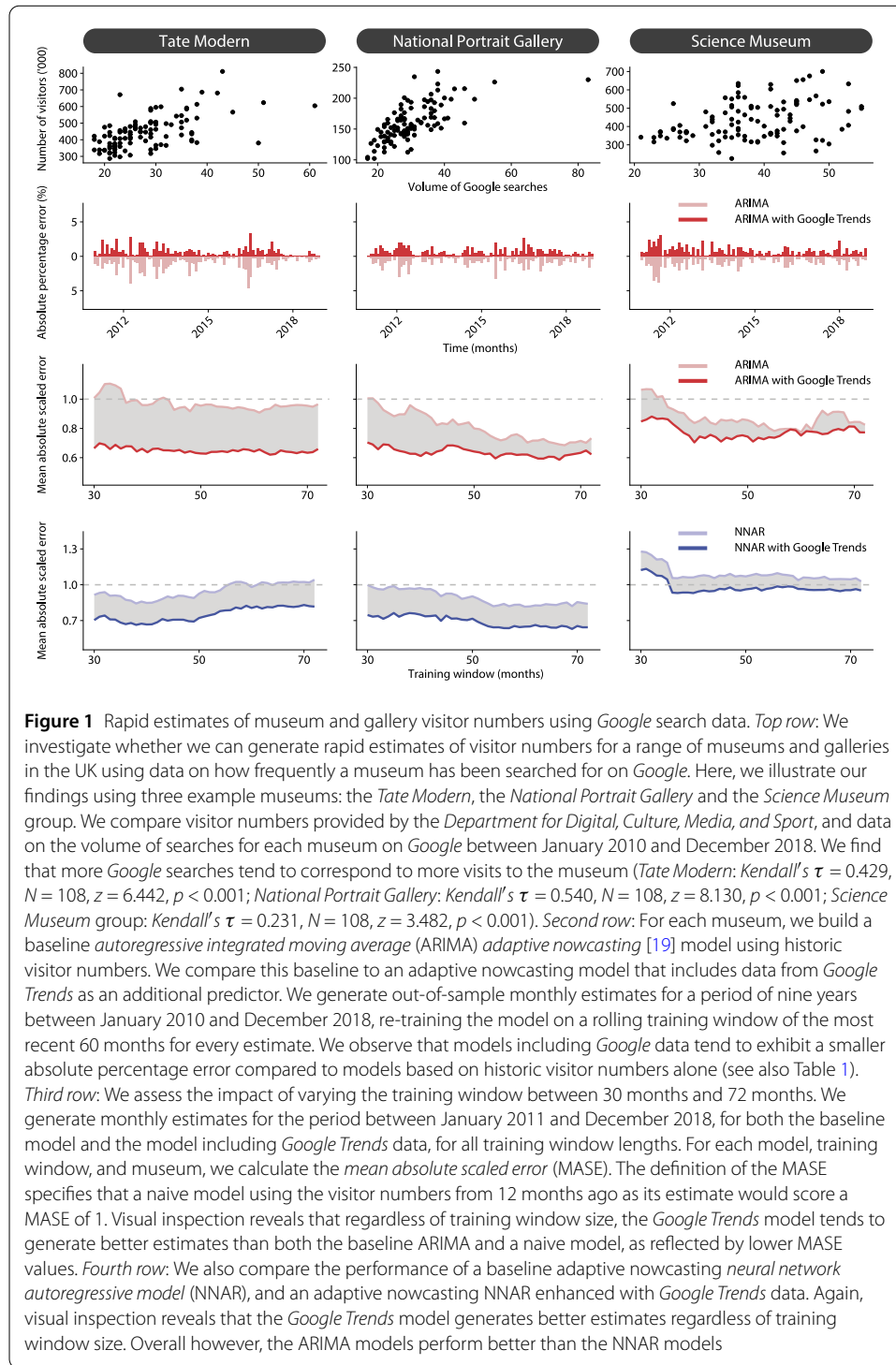
We make one exception to this approach for the *Tate Galleries* group, which is the largest group in terms of visitors, and for which there is no obvious *Google Trends* topic. There are, however, topics for each of the galleries that form part of the *Tate* group (the *Tate Britain*, the *Tate Modern*, the *Tate Liverpool* and the *Tate St Ives*), and therefore we include these galleries in our analysis as separate entities.

Our analysis does not include a few sites sponsored by DCMS. The *Geffrye Museum* closed down for refurbishment in January 2018, and severely reduced its visitor numbers during preparations for refurbishment; the *Tyne and Wear Museums* group does not have a dedicated *Google Trends* topic; and the *Museum of London* and *Museum of London Docklands* stopped being sponsored by DCMS in 2008, such that visitor numbers are no longer collated by the Department. In the first column of Table 1, we provide the complete list of museums and galleries considered in our analysis.

Figure 1 (top row) depicts a comparison between the monthly number of visitors and volume of *Google* searches between January 2010 and December 2018 for a subset of three well-known museums which are part of our analysis: the *Tate Modern*, the *National Portrait Gallery* and the *Science Museum*, where for the latter we consider visits to the full museum group. An initial correlation analysis suggests that months ranked higher in terms of *Google* search volume for a museum tend to also be months ranked higher in terms of visitor numbers (*Tate Modern*: Kendall's $\tau = 0.429$, $N = 108$, $z = 6.442$, $p < 0.001$; *National Portrait Gallery*: Kendall's $\tau = 0.540$, $N = 108$, $z = 8.130$, $p < 0.001$; *Science Museum* group: Kendall's $\tau = 0.231$, $N = 108$, $z = 3.482$, $p < 0.001$). We note that for all three museums examined here, the highest search volume index of 100 occurs between January 2004 and December 2009, and so is not depicted in this figure. This may be due to searches for these museums accounting for a larger proportion of all UK *Google* searches in *Google's* earlier days. In the Additional file 1 (Table S1), we report similar correlation analyses for all 16 museums and galleries. With the exception of the *Wallace Collection* and the *Royal Armouries*, we find that this correlation result holds across our sample.

3 Methods

We aim to generate estimates of the number of visitors to a given museum or gallery in month t at the beginning of month $t + 1$. With the release timetable for the museum and gallery official statistics that was in place until Spring 2019, this would anticipate official estimates by a month. For this task, we use the *adaptive nowcasting* approach, originally introduced by Preis and Moat [19] to improve monitoring of flu cases using *Google Trends* data. Within the adaptive nowcasting framework, we consider two families of forecasting models suitable for use with time series data: *autoregressive integrated moving average* (ARIMA) models, as used by Preis and Moat [19], and *neural network autoregressive*



(NNAR) models. In both cases, we test the performance of our models out-of-sample: that is, using data that we did not train the model on.

We first investigate the performance of the adaptive nowcasting approach when based on ARIMA models [19]. We use automatic model selection to fit the parameters of the ARIMA models [39], which for example include the number of lagged values of the time series that are used in the estimates (see [40] for further details). The ARIMA models we

build are seasonal, so that information on visitor numbers during the month of interest one year earlier can also be used to inform estimates.

We begin by building a baseline model that uses historical visitor numbers alone. This model will provide a benchmark of the quality of visitor number estimates that could be expected without using any additional data from *Google Trends* [19, 20, 27]. We train the baseline model on historical visitor numbers for a given museum over the previous 60 months. For instance, if we wish to estimate the number of visitors to the *Tate Modern* in January 2018, we train the ARIMA model on visitor numbers for the *Tate Modern* from January 2013 until December 2017. We can then generate an estimate of the number of visitors to the *Tate Modern* in January 2018. In other words, we build a one-step-ahead forecasting model training on a sliding window of 60 months. We refer to this model as the ARIMA baseline model. We return to investigate the importance of the length of the sliding window in the ARIMA baseline model subsequently.

Following Preis and Moat [19], we also build enhanced ARIMA models in which we add data on the volume of *Google* searches for the museum as a predictor. We first train the model using both historical visitor numbers and additional monthly data on the volume of search queries for the museum over the previous 60 months. We then draw on search volume data for month t , which is available at the beginning of month $t + 1$, to help generate a rapid estimate of the number of visitors in month t . For instance, to generate an estimate of the number of visitors to the *Tate Modern* in January 2018, we first train a model using data on the number of visitors to the *Tate Modern* and data on the volume of *Google* searches for the *Tate Modern* from January 2013 until December 2017. We then generate an estimate which draws on both the visits data for months previous to January 2018 and the volume of *Google* searches for the *Tate Modern* in January 2018. We hypothesise that the greater recency of the *Google* data will allow us to generate more accurate estimates of visitor numbers for that month, in comparison to generating estimates based on historical visitor numbers alone.

By retraining our model for each estimate, we are able to take into account that the relationship between how frequently people search for a museum and how frequently people visit the museum may change over time. For example, the arrival of an exhibition at a given gallery may prompt a large surge in *Google* searches but proportionally fewer visits, or vice versa. The adaptive nowcasting approach allows us to update our understanding of the relationship between search behaviour and visits as soon as new data comes in. In this way, we can avoid the problems that have previously been observed when models of the relationship between search behaviour and offline behaviour are only trained once and gradually become out-of-date [19, 20].

ARIMA models are widely used, but many other approaches to modelling time series exist. We therefore also consider a second family of time series models, neural network autoregressive (NNAR) models [41–44], which we plug into the adaptive nowcasting framework. In NNAR models, lagged values of the time series are used as inputs to the neural network. We consider neural networks with one hidden layer, where lagged values of the time series are combined in the hidden layer and then modified with a nonlinear transformation before the time series estimate is output via the output layer node. To add online data to an NNAR model, we add an additional input node to the neural network. We describe the NNAR models in more detail in the Additional file 1 (Fig. S1). As with the

Table 1 Estimating museum visitor numbers using *Google* search data. We aim to generate rapid estimates of the monthly number of visitors at a given museum using *adaptive nowcasting* [19] with minimum delay as soon as each month is over. We build adaptive nowcasting models using *autoregressive integrated moving average* (ARIMA) models, and *neural network autoregression* (NNAR) models. For both families of time series models, we construct a baseline model using historical visitor numbers only. We also construct a *Google Trends* model, in which data on the volume of *Google* searches for a museum is used as an additional predictor. All models reported here are trained using the previous 60 months of data. We generate monthly estimates for the period between January 2010 and December 2018, and report the *mean absolute percentage error* (MAPE). For each comparison between a baseline model and an equivalent *Google Trends* model, the smallest MAPE is highlighted in bold. We observe that nearly all models which include data derived from *Google Trends* exhibit a smaller MAPE than their counterpart baseline model based on historical visitor numbers alone

Museum	ARIMA (% error)		NNAR (% error)	
	Baseline	<i>Google Trends</i>	Baseline	<i>Google Trends</i>
British Museum	7.79	7.22	9.07	7.74
Horniman Museum	11.55	11.55	13.76	10.27
Imperial War Museums	12.49	10.60	14.43	12.09
National Coal Mining Museum For England	16.84	13.84	17.80	15.42
National Gallery	11.63	12.24	12.89	11.59
National Museums Liverpool	11.03	9.62	14.98	12.20
National Portrait Gallery	11.17	9.60	13.46	9.92
Natural History Museums	7.53	6.93	9.83	8.83
Royal Armouries	16.86	15.89	18.41	17.06
Science Museum Group	7.63	7.35	10.13	8.88
Tate Britain	21.46	13.09	24.11	14.31
Tate Modern	12.11	8.74	13.81	10.85
Tate Liverpool	15.15	12.81	17.87	14.47
Tate St Ives	58.14	52.30	61.13	52.11
The Wallace Collection	6.75	6.65	7.45	6.89
Victoria and Albert Museums	12.64	11.77	14.34	12.83

ARIMA model adaptive nowcasts, we retrain the neural networks at each time step using a sliding training window, before generating each estimate.

DCMS data is available from April 2004 onwards. So that we can work with a full calendar year of DCMS data whilst leaving space for a training window of five years (60 months), we start training our models from January 2005 and generate estimates from January 2010. For both the ARIMA and the neural network models, we produce monthly estimates for a nine year period from January 2010 until December 2018.

4 Results

In Table 1, we report the results of our analysis. Across all museums, we find that models including data from *Google Trends* exhibit a lower *mean absolute percentage error* (MAPE) than models based on historical visitor numbers alone, with the only exception being the ARIMA with *Google Trends* model for the *National Gallery*. We also note that in most cases the performance of ARIMA models is better than that of NNAR models. Figure 1 (*second row*) depicts how the absolute percentage error varies over time for the three example museums (the *Tate Modern*, the *National Portrait Gallery* and the *Science Museum* group) when using ARIMA models. Again, we observe that including data from *Google Trends* as a predictor tends to reduce the error in the estimates. In the Additional file 1, we depict the same results for the other 13 museums in our analysis (Figs. S3 to S5).

While our findings hold across nearly all of the museums considered, the improvement delivered by including *Google Trends* data does differ between museums. This could be

for a variety of reasons, including the underlying volume of search queries for the museum; the extent to which the *Google Trends* topic truly captures searches relating to the museum or museum group; the extent to which people have reason to search for the museum other than when they visit; the extent to which people visit without searching for the museum, for example because the museum is in a popular tourist area; or measurement and sampling noise in either the official visitor numbers or *Google* search data.

We perform further validation of our results with the modified *Diebold-Mariano* test, which compares errors from time series models to check whether two different models exhibit a statistically significant difference in forecast accuracy [45, 46]. This test can be used with a range of forecast error measures, but it has been shown that the *mean absolute scaled error* (MASE) satisfies all the required assumptions of the test, such as the asymptotic normality of the forecast errors, whereas other common measures such as the MAPE may not [47]. The MASE is a scale invariant measure of the accuracy of forecasts [48], making it possible to directly compare forecast errors for museums with very different visitor numbers. The MASE is also symmetric [48], such that it results in an equal penalty for underestimates and overestimates of the number of visitors.

The MASE compares the absolute error of a forecast with the error that would be expected from a naive forecast. For seasonal data, the naive forecast is that each value will be equal to the value observed one season ago. For monthly data with annual seasonality, this is therefore the value of the time series twelve months ago.

The MASE for seasonal time series is hence defined as:

$$\text{MASE} = \frac{\sum_{t=1}^T |e_t|}{\frac{T}{T-m} \sum_{t=m+1}^T |Y_t - Y_{t-m}|},$$

where e_t is the forecast error, defined as the actual value Y_t minus the value forecast by the model undergoing testing; m is the seasonal period, which is 12 for our analyses of monthly data with annual seasonality; and Y_{t-m} is the naive forecast estimate. By definition, the naive seasonal forecast model would score a MASE of 1. Values of the MASE lower than 1 imply that the model undergoing testing performs better than the naive forecast model.

We generate monthly estimates from January 2011 until December 2018 using the same procedure described above, varying the training window between 30 months and 72 months. Bearing in mind once again that we start training our models with data from January 2005, our analysis here generates estimates from January 2011 onwards, rather than January 2010, to allow us to explore how performance differs when we use a longer training window of six years (72 months).

Figure 1 (*third and fourth rows*) depicts the value of the MASE for the three example museums in our analysis (the *Tate Modern*, the *National Portrait Gallery* and the *Science Museum* group). Again, visual inspection suggests that regardless of training window size, models including data from *Google Trends* tend to exhibit a lower MASE than baseline models based on historical visitor numbers alone. This holds both for ARIMA and NNAR models. In the Additional file 1, we provide similar illustrations of the results for the other 13 museums analysed here (Figs. S3, S4 and S5). While we find a greater boost to performance from *Google Trends* data for some museums and much worse performance for others, we tend to observe the same broad pattern across our sample of museums.

We then perform the Diebold-Mariano test as follows. For a given training window size, for each model, we calculate the MASE for estimates made across all museums. In our

analysis, there are four different models: baseline ARIMA, ARIMA with *Google Trends*, baseline NNAR, and NNAR with *Google Trends*. To compare all four models with all other models, we therefore need to carry out six different pairwise comparisons. To correct for multiple hypothesis testing, we adjust the p -values returned by the Diebold-Mariano test using the *false discovery rate* correction [49]. Across all training windows, we find that models including data from *Google Trends* have a statistically significantly lower MASE compared to models based on historical visitor numbers alone. We report further details of this analysis in the Additional file 1 (Tables S2, S3 and S4).

To complement the Diebold-Mariano analysis, as a further check, we build a regression model of the mean absolute scaled errors to investigate whether the type of adaptive now-casting model used is a key predictor of the size of the error once the museum, month and training window size are all taken into account. We fit a generalised linear model using a gamma distribution, a logarithmic link function and robust standard errors, with the *model*, *museum*, *month* and *training window* as predictors. With 4 different models, 16 museums, 96 months of data and 43 training window lengths, our regression model is fit on 264 192 observations in total.

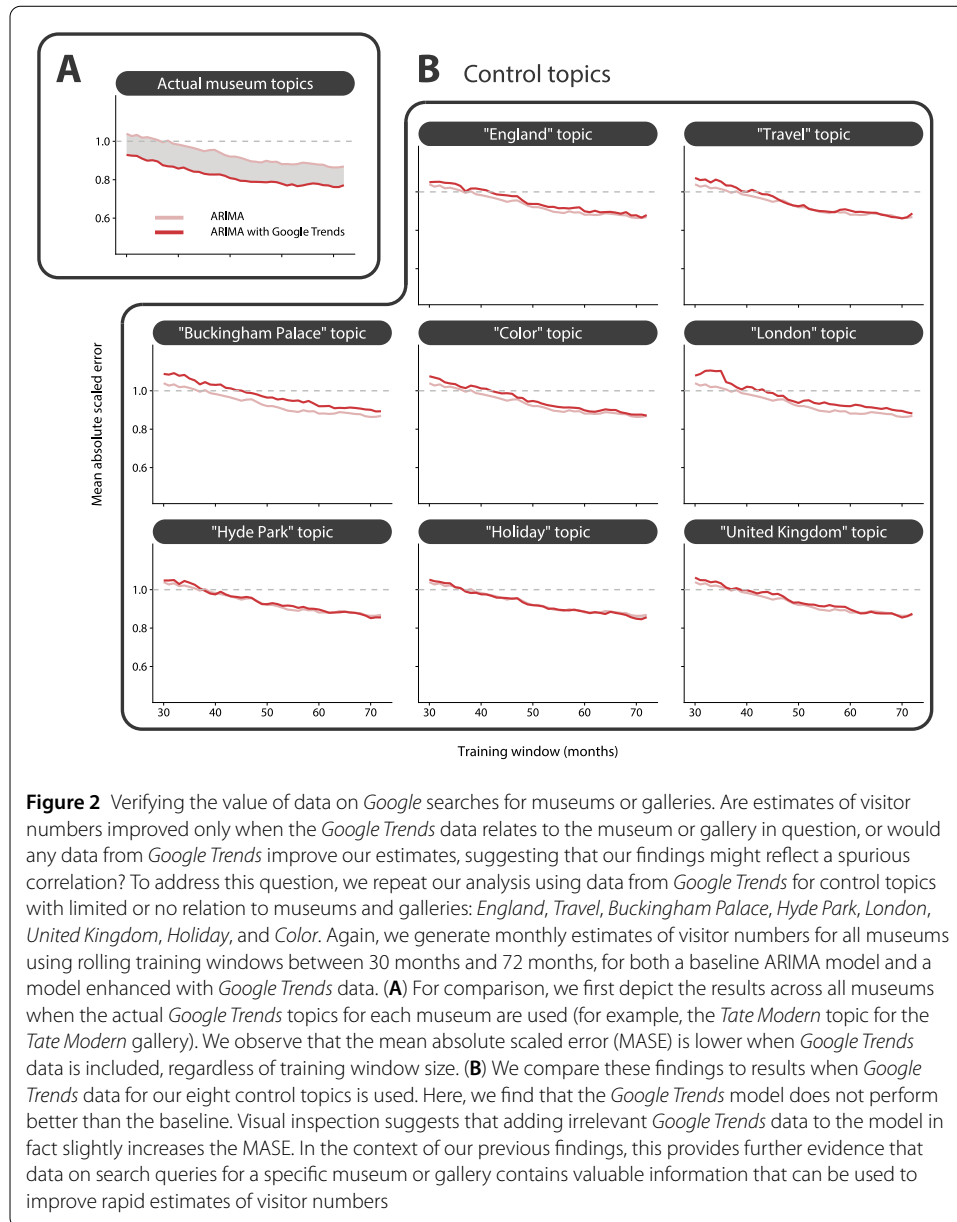
Each independent variable enters the model as a categorical variable. For the *model* variable, the four categories correspond to the different models: ARIMA, NNAR, ARIMA with *Google Trends*, and NNAR with *Google Trends*. We use the baseline ARIMA model as our reference level in the regression.

We are particularly interested in the coefficients of the *model* dummy variables, since these indicate whether models using *Google Trends* data have lower errors than the baseline ARIMA model. The fitted coefficient of the *model* dummy variable corresponding to the ARIMA with *Google Trends* model is statistically significant and negative (-0.132 , $p < 0.001$). Similarly, the coefficient for the NNAR with *Google Trends* model is also statistically significant and negative (-0.078 , $p < 0.001$), whereas the coefficient for the NNAR model with no *Google Trends* data is statistically significant and positive (0.078 , $p < 0.01$; for all regression results, see sample size information above). Both these results suggest that models which include *Google Trends* data result in smaller errors than their baseline counterparts, for ARIMA and NNAR models alike. More details of this analysis are presented in the Additional file 1 (Table S5).

Our analysis so far has shown that models which use data from *Google Trends* perform better than models based on historical visitor numbers alone. However, one question remains: must the *Google Trends* data relate to the museum or gallery in question, or would any data from *Google Trends* appear to improve our estimates? If *Google Trends* data from unrelated topics were to significantly improve our estimates, this would suggest that our findings might be the result of a spurious correlation between search data and visitor numbers data. To address this final question, we repeat our analysis using data from *Google Trends* for control topics with limited or no relation to museums and galleries.

For our control topics, we choose: *England*, *Travel*, *Buckingham Palace*, *Hyde Park*, *London*, *United Kingdom*, *Holiday*, and *Color*. Again, we restrict our *Google Trends* request to data on searches made in the United Kingdom. We generate estimates for all museums in our analysis for the time period between January 2011 and December 2018, and calculate the MASE.

Figure 2 depicts our results for ARIMA models. For simplicity, we present the results averaged across all museums. We again see lower MASEs for estimates generated using data



on *Google* searches for the museums and galleries in comparison to estimates generated using historical visitor numbers alone (Fig. 2(A)). In contrast, data on *Google* searches for unrelated topics makes near to no difference to estimates of visitor numbers when compared to estimates based on historical visitor numbers alone (Fig. 2(B)). In fact, visual inspection suggests that models that draw on *Google Trends* data on irrelevant control topics tend to perform a little worse than the baseline overall.

To verify that *Google Trends* data for control topics does not improve estimates of visitor numbers, we investigate the performance of both ARIMA and NNAR models. We again build a regression model of the mean absolute scaled errors, using a gamma distribution, a logarithmic link function and robust standard errors, with the *model*, *museum*, *month* and *training window* as predictors. We build one such regression model for each control topic. We find that the fitted coefficient of the *model* dummy variable for the ARIMA with

Google Trends is positive for all control topics, and statistically significantly so in the vast majority of cases. For the NNAR with *Google Trends* model, the coefficient is statistically significantly larger than the coefficient for the NNAR baseline for two of the control topics (both differences > 0.01 , both $ps < 0.025$), with no significant difference for the other six control topics (all absolute differences < 0.0007 , all $ps > 0.24$). We report on this analysis in further detail in the Additional file 1 (Tables S6–S13).

Overall, the results therefore indicate that the MASE either remains roughly the same or increases when irrelevant *Google Trends* data is fed into the model. We conclude that trying to improve estimates of visitor numbers with data on *Google* searches for unrelated control topics does not work, and may make the estimates worse. In the light of our previous results, this provides further evidence that data on search queries for a specific museum or gallery truly does contain valuable information on the number of people visiting those sites.

5 Discussion

We have shown that publicly available data from *Google Trends* on our collective online interest in a museum can be used to generate rapid estimates of the number of visitors to that museum before the official visitor numbers are released. These results hold for a range of museums and galleries sponsored by the *Department of Digital, Culture, Media and Sport* (DCMS). In particular, we have seen that models which draw on *Google* search query data outperform models based on historical visitor numbers alone. We note that none of the models in this analysis rely solely upon *Google* search data. Our findings provide evidence that historic visitor numbers are also valuable in estimating recent visitor numbers, and so we do not suggest that this traditional data should be discarded when available [19, 20, 27]. Our use of an adaptive nowcasting approach [19], where the model is retrained as new data comes in, should also help reduce the impact of spikes in search queries during which the relationship between online interest in the museum and visitor numbers changes, for example due to an exhibition which receives widespread news coverage.

Our analysis of course has a variety of limitations. Historical data released by DCMS is subject to updates whenever an error is found in the data collection. Such errors in historical visitor numbers would impact our estimates too, although our approach would allow us to integrate revised data as it arrives to improve future estimates. Estimates generated for different museums exhibit varying levels of accuracy, with some museums performing better than others. Use of these estimates for individual museums would therefore require consideration of the level of accuracy required. We also note that the models could almost certainly be improved by drawing on further information sources, such as data on tourist visits to the UK or London, derived either from official data or potentially from other online sources such as online photographs [10–12]; or data on the number of visits to the museums' and galleries' own websites.

Our findings provide further evidence that rapidly available data on collective online behaviour can help generate faster insights into changes in society as they occur. Out-of-date data impacts decision making across a wide range of policy areas. Reducing delays in data provision whilst managing the quality of the data provided is a challenge. However, we suggest that appropriately careful analyses of online data can help us work towards this goal, and thereby provide decision makers with a greater understanding of the present state of the world.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-020-00232-z>.

Additional file 1. Supplementary information. (PDF 529 kB)

Acknowledgements

We thank Gwion Aprhobad and Niall Goulding from the Department for Digital, Culture, Media and Sport for comments and the Office for National Statistics Data Science Campus for support.

Funding

TP, FB and HSM were supported by the University of Warwick and the ESRC via grant ES/M500434/1. TP and HSM are also grateful for support from The Alan Turing Institute under the EPSRC grant EP/N510129/1 via Turing awards TU/B/000008 (TP) and TU/B/000006 (HSM).

Availability of data and materials

All data sets are publicly available through the DCMS website [38] and *Google Trends* [50].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally. All authors read and approved the final manuscript.

Author details

¹Data Science Lab, Warwick Business School, University of Warwick, Coventry, UK. ²Department of Computer Science, University of Exeter, Exeter, UK. ³The Alan Turing Institute, British Library, London, UK.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 20 December 2019 Accepted: 24 May 2020 Published online: 05 June 2020

References

1. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D et al (2009) Computational social science. *Science* 323:721–723
2. Vespignani A (2009) Predicting the behavior of techno-social systems. *Science* 325:425–428
3. King G (2011) Ensuring the data-rich future of the social sciences. *Science* 331:719–721
4. Giles J (2012) Computational social science: making the links. *Nature* 488:448–450
5. Moat HS, Preis T, Olivola CY, Liu C, Chater N (2014) Using big data to predict collective behavior in the real world. *Behav Brain Sci* 37:92–93
6. Conte R, Gilbert N, Bonelli G, Cioffi-Revilla C, Deffuant G, Kertesz J et al (2012) Manifesto of computational social science. *Eur Phys J Spec Top* 214:325–346
7. González M, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. *Nature* 453:779–782
8. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. *J Stat Mech Theory Exp* 2009:L07003
9. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. *EPJ Data Sci* 4:10
10. Barchiesi D, Moat HS, Alis C, Bishop S, Preis T (2015) Quantifying international travel flows using Flickr. *PLoS ONE* 10:e0128470
11. Preis T, Botta F, Moat HS (2020) Sensing global tourism numbers with millions of publicly shared online photographs. *Environ Plan A* 52:471–477
12. Barchiesi D, Preis T, Bishop S, Moat HS (2015) Modelling human mobility patterns using photographic data shared online. *R Soc Open Sci* 2:150046
13. Botta F, Moat HS, Preis T (2015) Quantifying crowd size with mobile phone and Twitter data. *R Soc Open Sci* 2:150162
14. Douglass RW, Meyer DA, Ram M, Rideout D, Song D (2015) High resolution population estimates from telecommunications data. *EPJ Data Sci* 4:4
15. Botta F, Moat HS, Preis T (2019) Measuring the size of a crowd using Instagram. *Environ Plan B*, published online ahead of print
16. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. *Nature* 457:1012–1014
17. Brownstein JS, Freifeld CC, Madoff LC (2009) Digital disease detection-harnessing the Web for public health surveillance. *N Engl J Med* 360:2153–2157
18. Althouse B, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, Baumbach J et al (2015) Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Sci* 4:17
19. Preis T, Moat HS (2014) Adaptive nowcasting of influenza outbreaks using Google searches. *R Soc Open Sci* 1:140095
20. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: traps in big data analysis. *Science* 343:1203–1205

21. Hickmann KS, Fairchild G, Friedhorsky R, Generous N, Hyman JM, Deshpande A et al (2015) Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput Biol* 11:e1004239
22. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends. *Sci Rep* 3:1684
23. Curme C, Preis T, Stanley HE, Moat HS (2014) Quantifying the semantics of search behavior before stock market moves. *Proc Natl Acad Sci USA* 111:11600–11605
24. Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, Weber I (2012) Web search queries can predict stock market volumes. *PLoS ONE* 7:e40014
25. Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with Google searches: evidence from the Visegrad Group countries. *PLoS ONE* 10:e0127084
26. Kristoufek L (2013) BitCoin meets Google Trends and Wikipedia: quantifying the relationship between phenomena of the Internet era. *Sci Rep* 3:3415
27. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with Web search. *Proc Natl Acad Sci USA* 107:17486–17490
28. Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. *Sci Rep* 2:350
29. Choi H, Varian H (2012) Predicting the present with Google Trends. *Econ Rec* 88:2–9
30. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T (2013) Quantifying Wikipedia usage patterns before stock market moves. *Sci Rep* 3:1801
31. Miller S, Moat HS, Preis T (2020) Using aircraft location data to estimate current economic activity. *Sci Rep* 10:7576
32. Botta F, del Genio CI (2017) Analysis of the communities of an urban mobile phone network. *PLoS ONE* 12(3):e0174198
33. Aledavood T, López E, Roberts SG, Reed-Tsochas F, Moro E, Dunbar RI et al (2015) Daily rhythms in mobile telephone communication. *PLoS ONE* 10:e0138098
34. Saramäki J, Moro E (2015) From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *Eur Phys J B* 88:164
35. Yasseri T, Sumi R, Kertész J (2012) Circadian patterns of Wikipedia editorial activity: a demographic analysis. *PLoS ONE* 7:e30091
36. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J (2012) Dynamics of conflicts in Wikipedia. *PLOS ONE* 7:e38869
37. Samoilenko A, Yasseri T (2014) The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Sci* 3:1
38. Department for Digital, Culture, Media and Sports. Museums and galleries monthly visits. <https://www.gov.uk/government/statistical-data-sets/museums-and-galleries-monthly-visits>
39. Makridakis S, Wheelwright SC, Hyndman RJ (2008) Forecasting methods and applications. Wiley, New York
40. Stock JH, Watson M (2011) Introduction to econometrics, 3rd edn. Pearson Education, Harlow
41. Crone SF, Hibon M, Nikolopoulos K (2011) Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *Int J Forecast* 27:635–660
42. Hornik K, Leisch F (2001) Neural network models. In: *A Course in Time Series Analysis*, 348–362
43. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M et al (2018) forecast: Forecasting functions for time series and linear models. R package version 8.4. Available from. <http://pkg.robjhyndman.com/forecast>
44. Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. *J Stat Softw* 26:1–22
45. Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J Bus Econ Stat* 13:253–263
46. Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared errors. *Int J Forecast* 13:281–291
47. Franses PH (2016) A note on the mean absolute scaled error. *Int J Forecast* 32:20–22
48. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *Int J Forecast* 22:679–688
49. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300
50. Google Trends. <http://trends.google.com>

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)