SUPPLEMENTARY INFORMATION

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
[1]Data Science Lab, Behavioural Science, Warwick Business School, University of Warwick, Scarman Road, Coventry, CV4 7AL, UK

Kendall's tau correlation analysis

In Table S1, we report the results of a Kendall's tau correlation analysis comparing visitor numbers and *Google* search volume for all museums and galleries in our analysis.

Table S1 Correlation between visitors data in the period from January 2010 until December 2018 and Google search data.

Museum	Kendall's τ	z	p
British Museum	0.294	4.311	$< \mathbf{0.01}$
Horniman Museum	0.492	7.424	$< \mathbf{0.01}$
Imperial War Museums	0.275	4.127	$< \mathbf{0.01}$
National Coal Mining Museum For England	0.615	9.275	$< \mathbf{0.01}$
National Gallery	0.168	2.474	$< \mathbf{0.05}$
National Museums Liverpool	0.448	6.797	$< \mathbf{0.01}$
National Portrait Gallery	0.540	8.130	$< \mathbf{0.01}$
Natural History Museums	0.258	3.916	$< \mathbf{0.01}$
Royal Armouries	0.098	1.474	> 0.10
Science Museum Group	0.231	3.482	$< \mathbf{0.01}$
Tate Britain	0.548	8.057	$< \mathbf{0.01}$
Tate Modern	0.429	6.442	$< \mathbf{0.01}$
Tate Liverpool	0.242	3.621	$< \mathbf{0.01}$
Tate St Ives	0.585	8.879	$< \mathbf{0.01}$
The Wallace Collection	0.001	0.011	> 0.90
Victoria and Albert Museums	0.238	3.578	$< \mathbf{0.01}$

Neural network autoregressive model

Neural network autoregressive models are forecasting methods based on neural networks [1, 2]. Figure S1 depicts an example of the kind of neural network used in our analysis. We consider networks with an input layer, where the time series data enters the model, a hidden layer, and an output node from which the model estimate is retrieved. The networks we consider receive as input the lagged values of the time series data. Each lagged value enters the network via a node in the input layer, so there is one input node per lagged value. The optimal number of lagged values to use, and therefore of input nodes in the network, is determined using AIC [3]. Seasonal components can also be included as input nodes. In the example figure, the network receives two lagged values, a seasonal value from the previous year, and additional external data $x_{ext,t}$ which in our analysis is data from *Google*. We consider networks with one hidden layer and with the number of nodes in the hidden layer equal to half the number of input nodes plus one, as detailed in [4, 5]. The input z_j to each neuron j in the hidden layer is a weighted combination of the

input layers:

$$z_j = b_j + w_{t-1,j}Y_{t-1} + w_{t-2,j}Y_{t-2} + w_{t-12,j}Y_{t-12} + w_{reg,j}x_{ext,t}$$

The parameters b_j and $w_{i,j}$ are learned from the data. To avoid the weights becoming too large, it is customary to include a *decay parameter* [3]. We set this decay parameter to be equal to 0.5. Each node in the hidden layer then transforms its input in a nonlinear fashion:

$$s(z) = \frac{1}{1 + e^{-z}}$$

The weights are initialised randomly. We therefore train each network 100 times and then average the results.

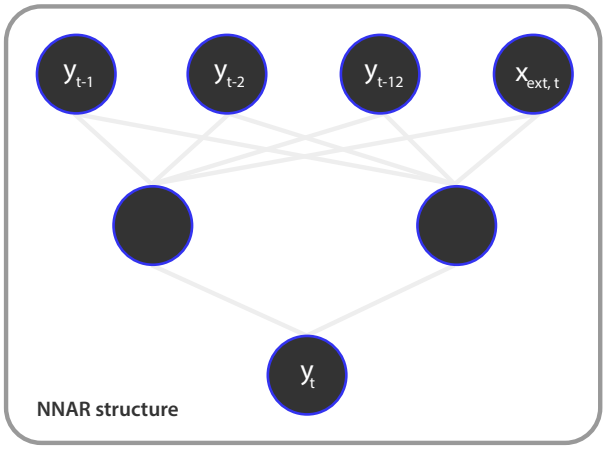

Figure S1 Forecasting with neural network autoregressive models (NNAR) | Neural networks can be used for forecasting time series data. Lagged and seasonal values of the time series are normally used as input nodes to the neural network. We automatically determine the optimal number of lags using AIC [3]. Information coming from additional external time series data can easily be incorporated in the network as an additional node in the input layer. The model used in our analysis then has one hidden layer, with the number of nodes in the hidden layer equal to half of the number of inputs plus one [4, 5]. We depict here the hypothetical case of a monthly time series with a seasonal component of one year. The network receives two lagged values from the two previous time steps, one value from the previous season (twelve months ago), and an external regressor bringing in additional information. The output layer of the network corresponds to the forecast.

Diebold-Mariano test

In our analysis of forecast accuracy, we perform a Diebold-Mariano (DM) test to compare four different models: ARIMA, NNAR, ARIMA with *Google Trends*, and NNAR with *Google Trends*. Since the test can only be performed between two models at a time, we have to perform six pairwise Diebold-Mariano tests to compare all models to each other. For this reason, we correct the resulting p-values of the tests to account for multiple comparisons, using the *false discovery rate* [6] correction method. We then perform this analysis for 43 different values of the training window used to calibrate the model. This results in 258 test statistics. For the

comparison between models with and without data from *Google Trends*, we find that all Diebold-Mariano statistics are larger than 2.592 (all $p < 0.05$)). We report here three examples of the full output of the Diebold-Mariano test for three specific training windows:

- 30 months, which is the shortest training window we used in our analysis
- 72 months, which is the longest training window we used in our analysis
- 54 months, which is the median of the range of training windows considered in our analysis

Tables S2, S3 and S4 report the values of the DM statistic, and the corresponding levels of significance. The *p-values* in the tables have already been adjusted using *false discovery rate* correction [6]. For each of these training window lengths, we find that the models which include data from *Google Trends* deliver statistically significant improvements over their baseline counterpart.

Table S2 Diebold-Mariano test statistics to compare models trained on 30 months of data. Bold entries indicate the key comparisons between a model which includes *Google Trends* data and the corresponding baseline model that is based on historical visitor numbers alone. Positive values of the DM statistic reflect lower error rates for the column model in comparison to the row model.

		Baseline		*Google Trends*	
		ARIMA	NNAR	ARIMA	NNAR
Baseline	ARIMA	-	-2.1 ($p > 0.05$)	**3.28 ($\mathbf{p < 0.01}$)**	0.82 ($p > 0.40$)
	NNAR		-	4.02 ($p < 0.001$)	**8.97 ($\mathbf{p < 0.0001}$)**
Google Trends	ARIMA			-	-1.42 ($p > 0.1$)
	NNAR				-

Table S3 Diebold-Mariano test statistics to compare models trained on 54 months of data. Bold entries indicate the key comparisons between a model which includes *Google Trends* data and the corresponding baseline model that is based on historical visitor numbers alone. Positive values of the DM statistic reflect lower error rates for the column model in comparison to the row model.

		Baseline		*Google Trends*	
		ARIMA	NNAR	ARIMA	NNAR
Baseline	ARIMA	-	-4.65 ($p < 0.001$)	**5.68 ($\mathbf{p < 0.001}$)**	-0.48 ($p > 0.5$)
	NNAR		-	7.12 ($p < 0.001$)	**7.78 ($\mathbf{p < 0.001}$)**
Google Trends	ARIMA			-	-4.11 ($p < 0.001$)
	NNAR				-

Table S4 Diebold-Mariano test statistics to compare models trained on 72 months of data. Bold entries indicate the key comparisons between a model which includes *Google Trends* data and the corresponding baseline model that is based on historical visitor numbers alone. Positive values of the DM statistic reflect lower error rates for the column model in comparison to the row model.

		Baseline		*Google Trends*	
		ARIMA	NNAR	ARIMA	NNAR
Baseline	ARIMA	-	-5.07 ($p < 0.001$)	**4.26 ($\mathbf{p < 0.001}$)**	-1.13 ($p > 0.2$)
	NNAR		-	6.03 ($p < 0.001$)	**6.17 ($\mathbf{p < 0.001}$)**
Google Trends	ARIMA			-	-3.98 ($p < 0.001$)
	NNAR				-

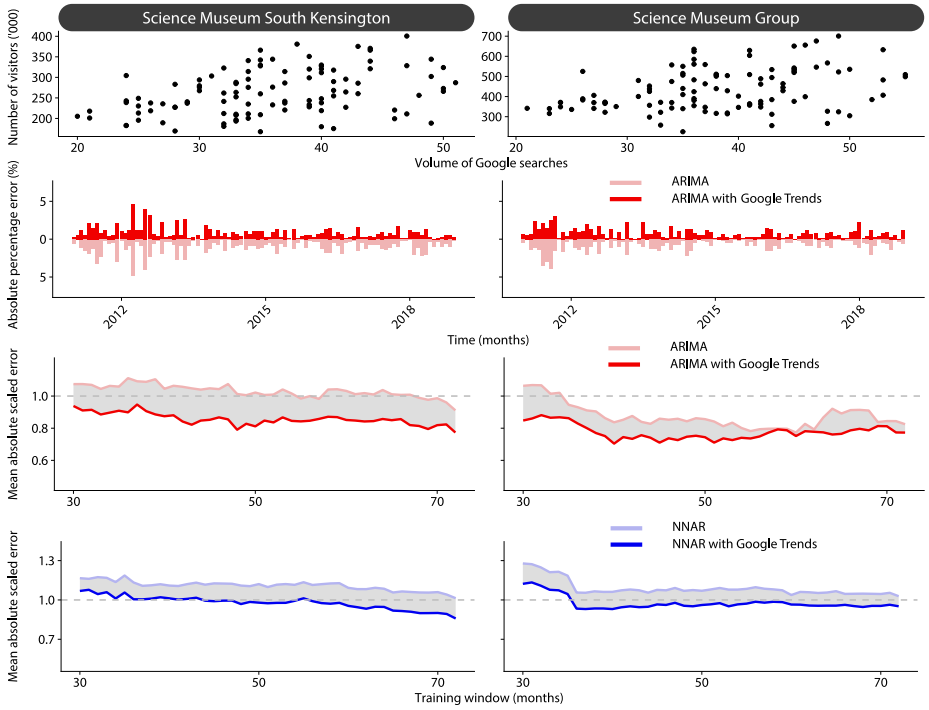

Figure S2 Investigating differences in performance when using visitor numbers for the main Science Museum site in South Kensington and for the whole Science Museum group.

Comparison between visitor number estimates for the *Science Museum* group and for the *Science Museum* site in South Kensington

In the main text, we present the results of our analysis for the *Science Museum* group, when we consider the number of visitors to museums across the whole group. However, we note that this resulted in one of the lowest correlations between the number of visitors and search query data. We hypothesise that this is because *Science Museum* topic data from *Google Trends* relates mainly to the *Science Museum*

in South Kensington, London. Here, we carry out the same analysis using only the number of visitors to the South Kensington site of the *Science Museum*. We find a stronger correlation when training on and estimating visitor numbers for the *Science Museum* South Kensington site (*Kendall's* $\tau = 0.274$, $N = 108$, $z = 4.095$, $p < 0.001$), in comparison to training on and estimating visitor numbers for all museums in the *Science Museum* group (*Kendall's* $\tau = 0.231$, $N = 108$, $z = 3.482$, $p < 0.001$; Fig. S2). However, we also note that the overall value of the MASE is comparable across the two analyses (Fig. S2).

Generalised linear model of errors

In our subsequent analysis of the forecast accuracy, 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. Each predictor enters the model as a categorical variable. For the *model* variable, the four categories correspond to the different models. We use the baseline ARIMA model as our reference level. 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.

As described in the main text, we are particularly interested in the coefficients corresponding to the *model* variable, since these indicate whether models using *Google Trends* data have lower errors than the baseline ARIMA model. Table S5 reports the results of the regression fit for the variables of interest. 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 ($0.078, p < 0.01$). Both these results suggest that ARIMA and NNAR models which include *Google Trends* data produce smaller errors than their baseline counterparts.

Tables S6 – S13 report analogous results for unrelated *Google Trends* control topics. 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). This suggests that adding *Google Trends* data for topics of limited relevance does not improve our estimates, and may in fact worsen the accuracy.

Table S5 Results of the linear regression analysis of the errors of the different models.

Model	Estimated coefficient	Standard error	z	p
ARIMA with *Google Trends*	-0.132	0.005	-25.279	<0.001 ***
NNAR with *Google Trends*	-0.078	0.005	-14.982	<0.001 ***
NNAR	0.078	0.005	15.191	<0.001 ***

Table S6 Results of the linear regression analysis of the errors of the different models when using *England* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.020	0.005	3.890	<0.001 ***
NNAR with *Google Trends*	0.094	0.005	18.315	<0.001 ***
NNAR	0.081	0.005	15.649	<0.001 ***

Table S7 Results of the linear regression analysis of the errors of the different models when using *Travel* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.016	0.005	3.140	<0.01 **
NNAR with *Google Trends*	0.076	0.005	14.697	<0.001 ***
NNAR	0.082	0.005	15.955	<0.001 ***

Table S8 Results of the linear regression analysis of the errors of the different models when using *Buckingham Palace* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.036	0.005	7.125	<0.001 ***
NNAR with *Google Trends*	0.095	0.005	18.641	<0.001 ***
NNAR	0.083	0.005	16.098	<0.001 ***

Table S9 Results of the linear regression analysis of the errors of the different models when using *Hyde Park* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.007	0.005	1.350	>0.10
NNAR with *Google Trends*	0.076	0.005	14.826	<0.001 ***
NNAR	0.081	0.005	15.736	<0.001 ***

Table S10 Results of the linear regression analysis of the errors of the different models when using *London* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.040	0.005	7.934	<0.001 ***
NNAR with *Google Trends*	0.083	0.005	16.182	<0.001 ***
NNAR	0.081	0.005	15.639	<0.001 ***

Table S11 Results of the linear regression analysis of the errors of the different models when using *United Kingdom* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.011	0.005	2.082	<0.05 *
NNAR with *Google Trends*	0.081	0.005	15.763	<0.001 ***
NNAR	0.081	0.005	15.758	<0.001 ***

Table S12 Results of the linear regression analysis of the errors of the different models when using *Holiday* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.002	0.005	0.308	>0.70
NNAR with *Google Trends*	0.082	0.005	15.971	<0.001 ***
NNAR	0.082	0.005	15.867	<0.001 ***

Table S13 Results of the linear regression analysis of the errors of the different models when using *Color* as the *Google Trends* topic.

Model	Estimated coefficient	Standard error	*z*	*p*
ARIMA with *Google Trends*	0.019	0.005	3.625	<0.001 ***
NNAR with *Google Trends*	0.082	0.005	15.996	<0.001 ***
NNAR	0.081	0.005	15.728	<0.001 ***

Results for all museums

In Figs. S3, S4 and S5, we depict the results shown in Fig. 1 in the main text for the remaining 13 museums in our analysis.

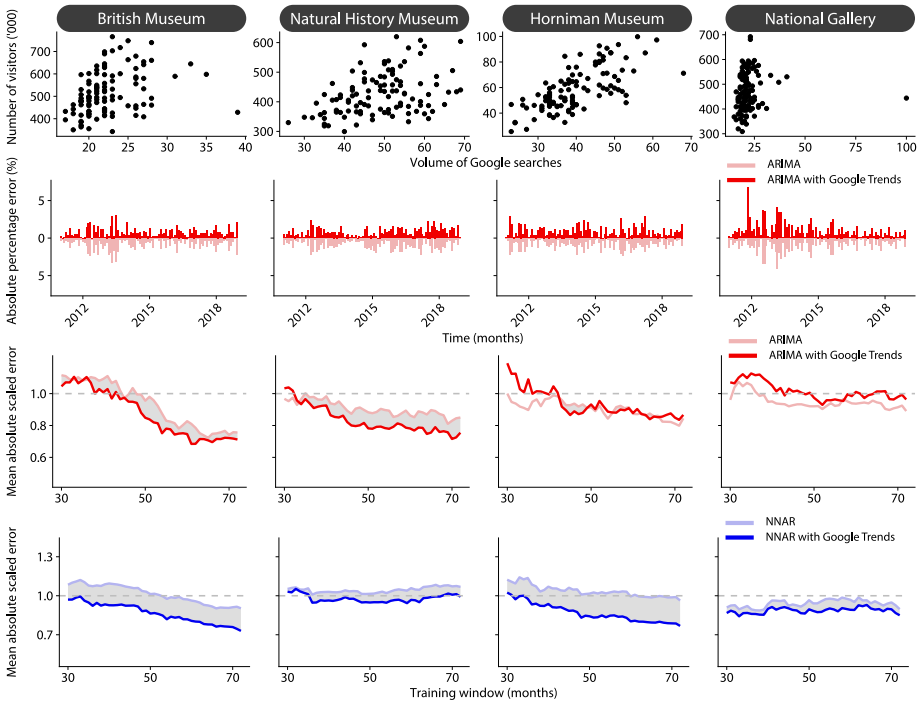

Figure S3 Rapid estimates of museum and gallery visitor numbers using Google data | Refer to the caption of Fig. 1 in the main text for further details.

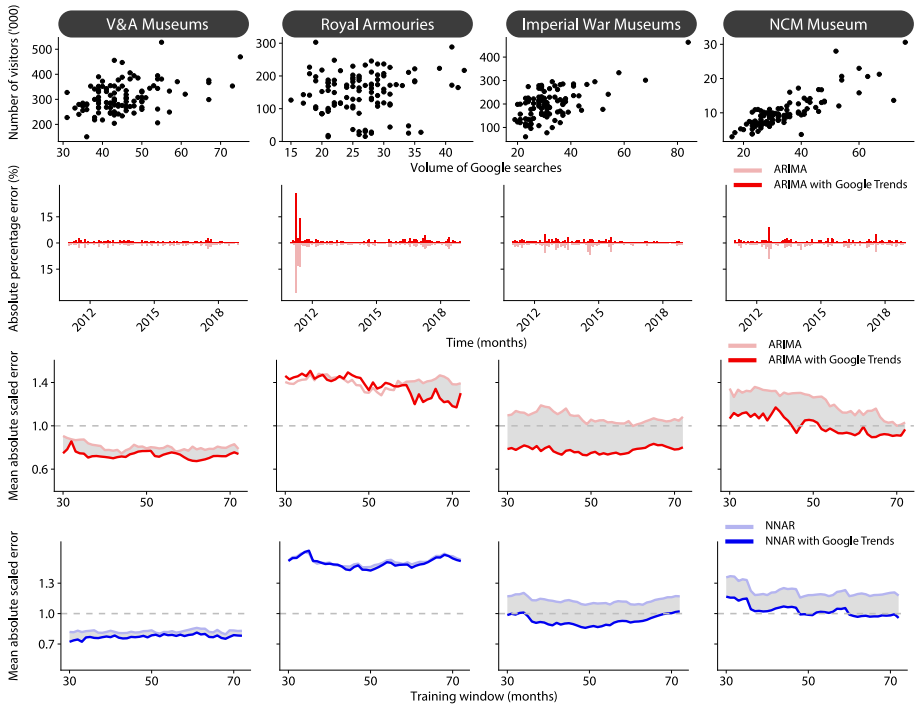

Figure S4 Rapid estimates of museum and gallery visitor numbers using Google data | Refer to the caption of Fig. 1 in the main text for further details. In this figure, *NCM Museum* stands for *National Coal Mining Museum of England*.

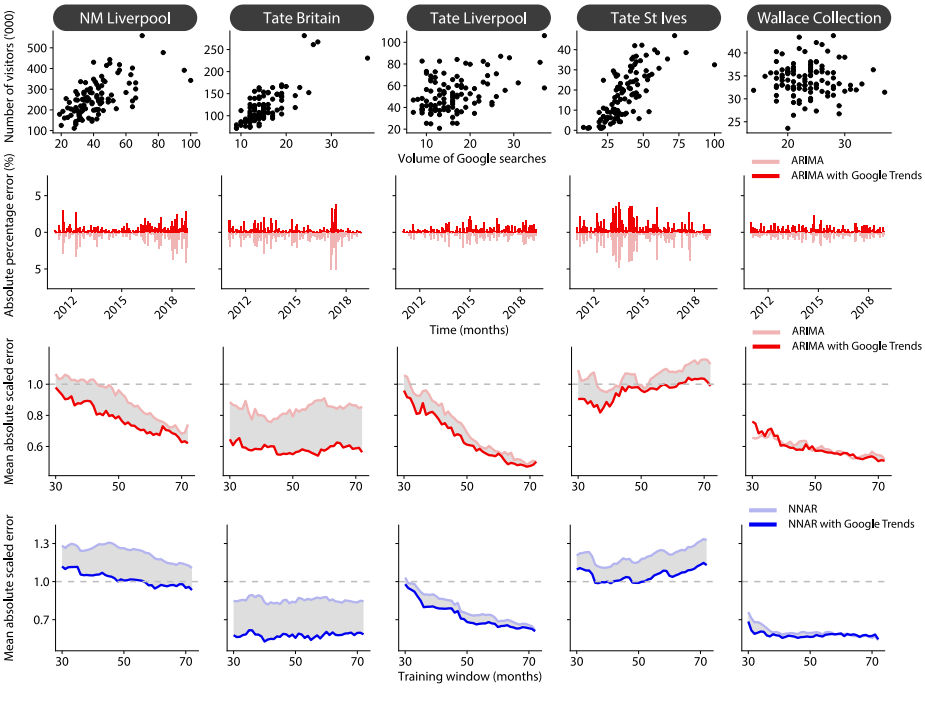

Figure S5 Rapid estimates of museum and gallery visitor numbers using Google data | Refer to the caption of Fig. 1 in the main text for further details. In this figure, *NM Liverpool* stands for *National Museums Liverpool*.

Author details
[1]Data Science Lab, Behavioural Science, Warwick Business School, University of Warwick, Scarman Road, Coventry, CV4 7AL, UK. [2]Department of Computer Science, University of Exeter, North Park Road, Exeter, EX4 4QF, UK. [3]The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK.

References

1. Crone SF, Hibon M, Nikolopoulos K. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. International Journal of Forecasting. 2011;27(3):635–660.
2. Hornik K, Leisch F. Neural network models. In: Peña D, Tiao GC, Tsay RS, editors. A Course in Time Series Analysis. Wiley Online Library; 2001. p. 348–362.
3. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2018.
4. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, et al.. forecast: Forecasting functions for time series and linear models; 2018. R package version 8.4. Available from: http://pkg.robjhyndman.com/forecast.
5. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software. 2008;26(3):1–22. Available from: http://www.jstatsoft.org/article/view/v027i03.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57:289–300.