Estimating tourism statistics with Wikipedia page views

Christian M Alis
University College London
Gower Street, London
WC1E 6BT, UK
c.alis@ucl.ac.uk

Adrian Letchford
Data Science Lab
Behavioural Science
Warwick Business School
University of Warwick
Coventry, CV4 7AL, UK
Adrian.Letchford@wbs.ac.uk

Helen Susannah Moat
Data Science Lab
Behavioural Science
Warwick Business School
University of Warwick
Coventry, CV4 7AL, UK

Tobias Preis
Data Science Lab
Behavioural Science
Warwick Business School
University of Warwick
Coventry, CV4 7AL, UK

ABSTRACT

Decision makers depend on socio-economic indicators to shape the world we inhabit. Reports of these indicators are often delayed due to the effort involved in gathering and aggregating the underlying data. Our increasing interactions with large scale technological systems are generating vast datasets on global human behaviour which are immediately accessible. Here we analyse whether data on how often people view *Wikipedia* articles might help us to improve estimates of the current number of tourists leaving the UK. Our analyses suggest that in the absence of sufficient history, *Wikipedia* page views provide an advantage. We conclude that when using adaptive models, *Wikipedia* usage opens up the possibility to improve estimates of tourism demand.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Time series analysis

Keywords

Computational Social Science, Complexity Science, Data Science, Forecasting, Nowcasting, Wikipedia

1. INTRODUCTION

Society's increasing interaction with large technological systems is generating vast amounts of data on human behaviour. These new humanly generated data sets capture human behaviour on large scales thereby allowing, for example, forecasting of future behaviour with Wikipedia page view data [4]. In addition, these data sets are available almost immediately after their generation.

Recent research has focused on using online data on human behaviour to improve estimates of the current levels of socio-economic indicators. Instead of forecasting the future, these approaches aim to forecast present values—commonly referred to as *nowcasting*. An example in the area of tourism is the use of *Google* search data to improve the current estimate of tourists visiting Hong Kong [1].

Here, we built a prototype of an online service analysing *Wikipedia* page views. The platform allows users to upload a reference time series in order to retrieve the names of the 100 *Wikipedia* articles whose article views time series are most correlated with this reference time series. We use this system to identify *Wikipedia* articles that may provide extra information concerning tourism levels in the UK. Using the platform outputs, we build nowcasting model to quantify the extent to which correlated *Wikipedia* page view time series might be used to improve current estimates of monthly UK tourism levels—which is reported by the *UK Office for National Statistics* with a time lag of one to two months.

2. METHODS

We developed a prototype of an online platform called TREE[1] for identifying Wikipedia articles whose page views are most correlated with a reference time series. TREE outputs the 100 most correlated articles ranked by Pearson's correlation coefficient. The platform uses page views data extracted from the hourly *Wikimedia* dumps[2] covering the period from 9$^{\text{th}}$ December 2007 to 25$^{\text{th}}$ March 2014.

The statistics of overseas travel and tourism are released monthly, with a one to two month delay in the release. In this study, we analysed the number of UK residents travelling abroad for tourism (outbound tourists). We restrict our analysis to the period January 2008 until February 2014 for which *Wikipedia* data is available.

We use a standard approach for creating nowcasting models. Specifically, we apply standard automatic model selec-

[1] `http://tree.mu`
[2] `http://dumps.wikimedia.org/other/pagecounts-raw/`

WebSci '15, June 28 - July 01, 2015, Oxford, United Kingdom
 ISBN 978-1-4503-3672-7/15/06...$15.00
DOI: http://dx.doi.org/10.1145/2786451.2786925

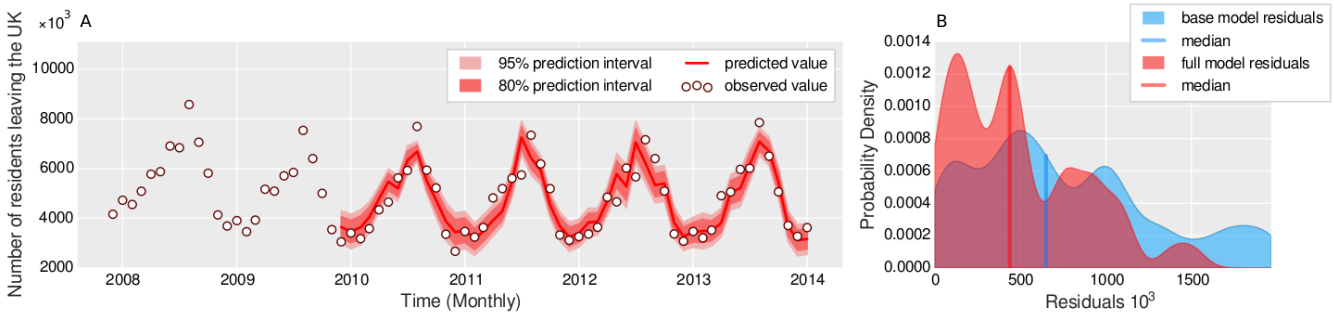

Figure 1: **Out-of-sample nowcasting of UK tourism.** (**A**) We use the first 24 months to train both base and full models and nowcast one step ahead. We repeat this step for all other values in our time series using a sliding window approach. Here, we show the 80% and 95% prediction interval and the predicted value in different shades of red. (**B**). The median of the full model's absolute residuals is significantly lower than the base model's median absolute residuals. (median of the base model's absolute residuals = 650.8, median of the full model's absolute residuals = 438.3; $V = 323$, $p < 0.01$, two sample paired Wilcoxon signed rank test).

tion procedures [3] for an autoregressive integrated moving average (ARIMA) time series model as described in more detail by Stock & Watson [5]. In this paper, we utilized the `auto.arima()` function of the forecast [2] R package for estimating parameters of the ARIMA model. It selects the best model based on the corrected Akaike information criterion (AICc). We then apply this model to compute out-of-sample values in our analyses.

We use the total number of page views of a selected set of *Wikipedia* articles as an external regressor in our analyses. Specifically, we investigate whether the number of page views at time t can be used to improve estimates of tourism statistics at time t where the latter is not known at time t.

3. RESULTS

We investigate whether or not *Wikipedia* usage can significantly improve out-of-sample estimation of tourist statistics. We train an ARIMA model on a sliding window of 24 months and nowcast the following month. The base model only uses the historical tourist levels. The full model's external regressor is the total views count of the 50 *Wikipedia* articles whose page views are most correlated with the monthly number of outbound tourists in each window.

The full model's nowcast results are shown in Figure 1A. The probability density of both the base model's and full model's residuals is depicted in Figure 1B. Again, we find that the full model is significantly better than the base model (median of the base model's absolute residuals = 650.8, median of the full model's absolute residuals = 438.3; $V = 323$, $p < 0.01$, two sample paired Wilcoxon signed rank test).

The out-of-sample window size of 24 gives a significantly improved estimate when including *Wikipedia* page view counts. We repeat the same out-of-sample analysis using a sliding window of 12 months and 36 months. When the window size is reduced to 12 months, the base model performs no better than when the window is 24 months (window size = 12 median of the absolute residuals = 665.5, window size = 24 median of the absolute residuals = 650.8; $V = 783$, $p = 0.1616$, two sample paired Wilcoxon signed rank test). Also when the window size is 12 months, the full model does not perform significantly different to the base model (median of the base model's absolute residuals = 665.5, median of the full model's absolute residuals = 438.3; $V = 791$, $p = 0.1946$, two sample paired Wilcoxon signed rank test). When the window size is increased to 36 months, using *Wikipedia* page views does not improve the model's performance (median of the base model's absolute residuals = 182.7, median of the full model's absolute residuals = 245.8; $V = 471$, $p = 0.1485$, two sample paired Wilcoxon signed rank test).

4. DISCUSSION

Using our prototype platform Tree for identifying *Wikipedia* articles whose page views are most correlated with a reference time series, we investigate in a case study whether estimates of current UK tourism statistics can be improved by incorporating online data. We find out-of-sample improvements of the full model incorporating *Wikipedia* article views when the sliding training window size is 24 months. However, the model performs worse when the sliding window size increases to 36 months. This shows, that given enough data, the ARIMA based approach is able to sufficiently model and nowcast the underlying time series. In the absence of sufficient data, *Wikipedia* page views provide an advantage.

5. ACKNOWLEDGEMENTS

The authors acknowledge the support of Research Councils UK Digital Economy via grant EP/K039830/1.

6. REFERENCES

[1] H. Choi and H. Varian. Predicting the Present with Google Trends. *Economic Record*, 88:2–9, 2012.

[2] R. J. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 7 2008.

[3] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman. *Forecasting methods and applications*. John Wiley & Sons, London, UK, 3rd edition, 1998.

[4] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3:1801, May 2013.

[5] J. H. Stock and M. Watson. *Introduction to econometrics*. Pearson Education, Harlow, UK, 2011.