Detecting and Gauging Impact on Wikipedia Page Views

Xiaoxi Chelsy Xie
cxie@wikimedia.org
Wikimedia Foundation

Isaac Johnson
isaac@wikimedia.org
Wikimedia Foundation

Anne Gomez
agomez@wikimedia.org
Wikimedia Foundation

ABSTRACT

Understanding how various external campaigns or events affect readership on Wikipedia is important to efforts aimed at improving awareness and access to its content. In this paper, we consider how to build time-series models aimed at predicting page views on Wikipedia with the goal of detecting whether there are significant changes to the existing trends. We test these models on two different events: a video campaign aimed at increasing awareness of Hindi Wikipedia in India and the page preview feature roll-out—a means of accessing Wikipedia content without actually visiting the pages—on English and German Wikipedia. Our models effectively estimate the impact of page preview roll-out, but do not detect a significant change following the video campaign in India. We also discuss the utility of other geographies or language editions for predicting page views from a given area on a given language edition.

CCS CONCEPTS

• **Information systems** → **Web mining**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*.

KEYWORDS

wikipedia; bayesian structural time series; page views; causal inference

ACM Reference Format:
Xiaoxi Chelsy Xie, Isaac Johnson, and Anne Gomez. 2019. Detecting and Gauging Impact on Wikipedia Page Views. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316751

1 INTRODUCTION

Wikipedia is the fifth-most-visited website worldwide [1] at 190 billion page views in 2018 alone [7] and is turned to by readers for all sorts of reasons ranging from simple curiosity to fact-checking to making a personal decision [20]. Despite its success, there are still many regions in the world where it is relatively unknown [9] or access is blocked [6]. In an attempt to improve access and awareness worldwide, the Wikimedia Foundation (WMF) has conducted various campaigns and efforts aimed at improving access to Wikipedia in various regions.[1]

Many researchers have sought to estimate the impact of external events on Wikipedia page views. This has included the effect of posting Wikipedia articles on external websites [14, 23], privacy concerns on viewing of sensitive Wikipedia articles [15], and censorship [24]. Conversely, much research has also sought to use Wikipedia page views as a predictor—i.e. *nowcasting* or *forecasting* [16]—of external entities such as the stock market [13], box office returns for movies [12], and disease incidence [16].

Evaluating the impact, however, of a given campaign or external event can be difficult. Daily page views to Wikipedia projects can be quite noisy, being affected by weekly, seasonal, holiday-related trends [21]. A change in the volume of page views following an awareness campaign could also easily be the result of an unrelated event—e.g., a celebrity marriage or World Cup game [7]. To account for these challenges, many researchers rely on some form of regression discontinuity design that focuses on changes between a short time period (e.g., two weeks) before and after an event (e.g., [14, 23, 24]). While powerful, this approach is limited to studying short-term effects and does not naturally lend itself to the task of nowcasting or forecasting.

In this paper, we explore the utility of Bayesian structural time series (BSTS) models for evaluating the impact of external events. BSTS models are designed to predict a given time series based on historical data, seasonality components, and additional control time series. They naturally incorporate uncertainty and the resulting forecast can then serve as a counterfactual—i.e. be compared against the actual time series following a given intervention to determine whether there is evidence that the intervention increased or decreased the magnitude of the time series. We test the BSTS model in two scenarios: the page preview roll-out in German and English Wikipedia as well as a video campaign in India designed to increase awareness about Hindi Wikipedia.

Our contributions are as follows:

- **Page Previews**: using the roll-out of page previews on the German and English Wikipedia, we demonstrate that our BSTS model can effectively detect changes in page views given predictive control series.
- **Impact of video campaigns in India**: applying our BSTS model to online and TV awareness campaigns in India, we do not find evidence of increased page views as a result of the online or TV campaigns.
- **Correlations across languages and regions**: we evaluate the predictive power of page-view time-series between pairs of Wikipedia language editions and regions. We find evidence that page view trends are unique to a given country and language edition and that control series ideally originate from the same language edition followed by same country to be a useful predictor.

[1] https://meta.wikimedia.org/wiki/New_Readers

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316751

2 RELATED WORK

In this work, we draw methods from the time series prediction literature and motivation from the literature that has sought to understand the impact of external events on Wikipedia activity.

2.1 Time Series Modeling

The goal of the time series modeling that we employ in this paper is to understand whether a specific intervention significantly impacts a given metric for which we have temporal data—e.g., whether the roll-out of a new feature causes a change in daily traffic. There are many approaches to time series modeling that span from quite simple to much more complex in accordance with how many assumptions they make. Common to these models, however, is that their validity depends on the model being able to make direct comparisons between the time series prior to an intervention and the time series following the intervention [2]. Threats to this validity come from a variety of sources that may affect time series independent of the intervention being studied: seasonality effects such as natural variation by day of week or month of year, unaccounted external events such as holidays or changes in the size of the underlying population. The likelihood that these events affect the time series increases as the time period being studied increases. A strong model, then, incorporates covariates that can control for seasonality, holidays, and other external factors that might affect the time series. A strong model also effectively represents its own uncertainty about predictions when there are insufficiently strong controls in place.

The core distinguishing features of approaches are 1) whether they include a control time series, and, 2) whether they directly compare metrics before and after the intervention or predict the time series after the intervention and compare this counterfactual prediction with the actual data. A control time series is a time series that is highly correlated with the "treated" times series, but, importantly, is known to not be affected by the intervention. The value of a control time series is that it helps to ensure that if seasonality or another event affects the treated time series, this effect is not conflated with the impact of the intervention because the effect should also be present in the control time series. The value of comparing the post-intervention time series with a counterfactual, as opposed to just the values from prior to the intervention, is greater flexibility to changing conditions. A model that produces the counterfactual can take into account more data about the time series prior to the intervention and, therefore, better account for shifts in covariates that might occur following the intervention. This is especially important when considering the long-term effects of an intervention. For these models, if the actual time series falls outside of the bounds of the counterfactual time series, this provides evidence that the intervention had a significant impact.

There are many considerations for how to build a robust time series model in order to produce the counterfactual predictions. Primarily, some models that use static regression to produce the counterfactual predictions falsely assume the data to be independent and identically distributed (i.i.d), which would result in an underestimation of the uncertainty [3]. Secondarily, to help avoid over-fitting, we need to choose appropriate control time series. Castle et al. [5] reviews and compares 21 methods for variable selection, including significance testing (e.g., forward and backward step-wise regression) and information criteria (e.g., AIC, BIC). Other popular model selection algorithms in time-series forecasting includes principal component and factor models, and penalized regression models (e.g., Lasso, ridge regression). However, these techniques force us to use a fixed set of selected variables, or do not account for the uncertainty in variable selection. Lastly, in order to gauge the uncertainty of the impact, we need to account for various sources of uncertainty in the model. Besides the uncertainty in variable selection and auto-correlation mentioned above, we also want to account for uncertainties in the historical relationships between treated and control time series, as well as uncertainties in seasonality and other components in the model.

2.2 Impact of External Events on Wikipedia

A number of papers have considered the challenge of establishing how an external event has affected dynamics within Wikipedia. Vincent et al. [23] and Moyer et al. [14] take what is known as a *interrupted time series* (ITS) approach to examine how posts on Reddit that include links to Wikipedia articles affect page view traffic on Wikipedia. Both model a Reddit post with a Wikipedia link as a "shock" to that Wikipedia article and compare the mean number of page views in a short period of time before and after the post to determine whether there is a significant difference. Zhang and Zhu [24] take a similar approach for the rate of contributions to Chinese Wikipedia from outside editors before and after a block on mainland China. Zhang and Zhu also seek to control for seasonal trends by examining the same time period in prior years. These approaches build on the assumption that there should be no difference in the expected page views between the pre- and post-intervention periods, and therefore any difference in page views can be causally tied to the Reddit post or block. As discussed above, these are assumptions that may hold true for short time-spans like one week, given that they do not happen to coincide with major holidays or events. This assumption is increasingly tenuous, however, as more long-term trends are considered.

Penney [15] also starts with an ITS model to understand the impact of the Edward Snowden revelations on page views to "terrorism-related" Wikipedia articles. Notably, because Penney examines a much longer time-period comprising 32 months, their analysis also includes a "control time series" that are security-related articles that are similar in content but less likely to be affected by the Snowden revelations. This approach is often referred to as difference-in-differences (DD) and is similar to how we construct the BSTS models considered in this work, but the BSTS models directly incorporate the concept of a control series as a core component of the models and captures the uncertainty of the relationship between the treated series and the control series. This makes for a much more explicit and robust means of controlling for additional external effects that may otherwise be conflated with the treatment under study.

3 METHODS

We use a single time series model architecture, described below, and apply it to two different events. Each event involves an external event that led to a potential shift in page views. We describe each event alongside its results.

3.1 Bayesian Structural Time Series Model

In this work, we use Bayesian structural time series (BSTS) model [19]. Per the components discussed in §2.1, these models can incorporate control covariates and time series, generate counterfactual predictions of page views for the post-intervention period assuming that the intervention did not take place, and naturally model their own uncertainty about these counterfactuals. We can then compare the counterfactual predictions and actual page views to quantify the causal impact of the intervention.

BSTS models combine three statistical methods into an integrated architecture [18]:

- A structural time series model for trend and seasonality, estimated using Kalman filters;
- Spike and slab regression for variable selection;
- Bayesian model averaging for the final prediction.

Structural Time Series Model: Under different assumptions, a very large class of models can be expressed in the form of structural time series models, including all ARIMA models [10]. This flexibility allows BSTS models to accommodate multiple sources of variations, including trends, seasonality, and latent evolutions of the treated series that cannot be explained by known trends or events. Specifically, a structural time series model (e.g. Eq. 1) decomposes the time series into four components: a level (μ_t), a local trend or slope (δ_t), seasonal effects (τ_t) and error terms. The model described here adds a regression component ($\beta^T \mathbf{x}_t$) to incorporate the control time series and other covariates. It is a stochastic generalization of the constant-trend regression model (e.g. $y_t = \mu + \delta t + \beta^T \mathbf{x}_t + \epsilon_t$), where the level μ_t and slope δ_t parameters each follow a random walk model instead of a constant. This allows for greater flexibility in the trends expressed within the model.

$$\begin{aligned} y_t &= \mu_t + \tau_t + \beta^T \mathbf{x}_t + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + u_t, u_t \sim N(0, \sigma_u^2) \\ \delta_t &= \delta_{t-1} + v_t, v_t \sim N(0, \sigma_v^2) \\ \tau_t &= -\sum_{s=1}^{S-1} \tau_{t-s} + w_t, w_t \sim N(0, \sigma_w^2) \end{aligned} \tag{1}$$

Spike and Slab: There are often many potential control series but including them all would likely lead to over-fitting and very complex models. A spike-and-slab prior over coefficients [8, 11] is designed to solve this challenge. The spike part controls the probability of whether a given variable would be chosen for the model—i.e. having a non-zero coefficient. The slab part shrinks the non-zero coefficients toward prior expectations which is often zero. Upon observing data, Bayes' theorem updates the inclusion probability of each coefficient. Then when sampling from the posterior distribution of a regression model, many of the simulated regression coefficients will be exactly zero [17].

Bayesian Model Averaging: To generate counterfactual predictions, the procedure uses the Markov chain Monte Carlo (MCMC) algorithm to draw samples from the parameter's posterior distribution and then combine that with the available data to yield a distribution of the counterfactual predictions. The model can then compute the difference between the actual values of a treated series in the post-intervention period and the distribution of counterfactual samples to yield an estimate of the distribution of the impact [4]. Because the structural time series model, spike-and-slab regression and model averaging all have natural Bayesian interpretations, BSTS is able to account for various sources of uncertainties using MCMC. This allows us to gauge confidence in the magnitude of causal impact and estimate the posterior probability that the causal impact is non-existent.

We use the *BSTS*[2] and *CausalImpact*[3] R packages to fit the BSTS models. The following parameters comprise each BSTS model. Figure 1 illustrates the parameters of the model for Hindi online video campaign in §4.2.

- **Treated Time Series**: this is the time series under study—e.g., daily internally-referred page views to English Wikipedia over the period of several months.
- **Intervention**: an event that occurred on a specific date during the study period that is believed to have affected the treated time series—e.g., the introduction of a new feature onto Wikipedia that might change the daily number of page views. Model validation is conducted entirely on data prior to the intervention.
- **Pre-Intervention Period**: time period from the first data point of the treated series to the day before the intervention. For each model in this work, we explore four different pre-intervention period length using grid-search: 12 weeks, 18 weeks, 183 days and 400 days.
- **Post-Intervention Period**: time period following the intervention for which the impact is being estimated—e.g., daily page views for the six weeks following the roll-out of a new feature. For each model in this work, we set this to 6 weeks.
- **Covariates**: additional variables that help explain the treated time series—e.g., total population online in a country. This also includes the control series described below.
- **Control Series**: a time series that is predictive of the treated time series prior to the intervention, but that is not be impacted by the intervention—e.g., daily page views for a similar Wikipedia edition for which the feature was not rolled out. The authors of the *CausalImpact* library we use for estimating the models suggest using 3-50 covariates.[4] Thus, for some models in this work where we have hundreds of control series—e.g., many combinations of different regions and language editions—we use correlation and dynamic time warping (DTW) [22] algorithms with pre-intervention data to prescreen and trim the list of control series before feeding them into the BSTS model.
- **Seasonality**: weekly and seasonal trends, or holiday effects that did not get captured by the control series. For each model in this work, we include features for day-of-week and month. We also include major holidays for the regions under study as described in the Results section.

[2] https://cran.r-project.org/package=bsts

[3] https://cran.r-project.org/package=CausalImpact

[4] https://stats.stackexchange.com/questions/162930/causalimpact-should-i-use-more-than-one-control/163554#163554

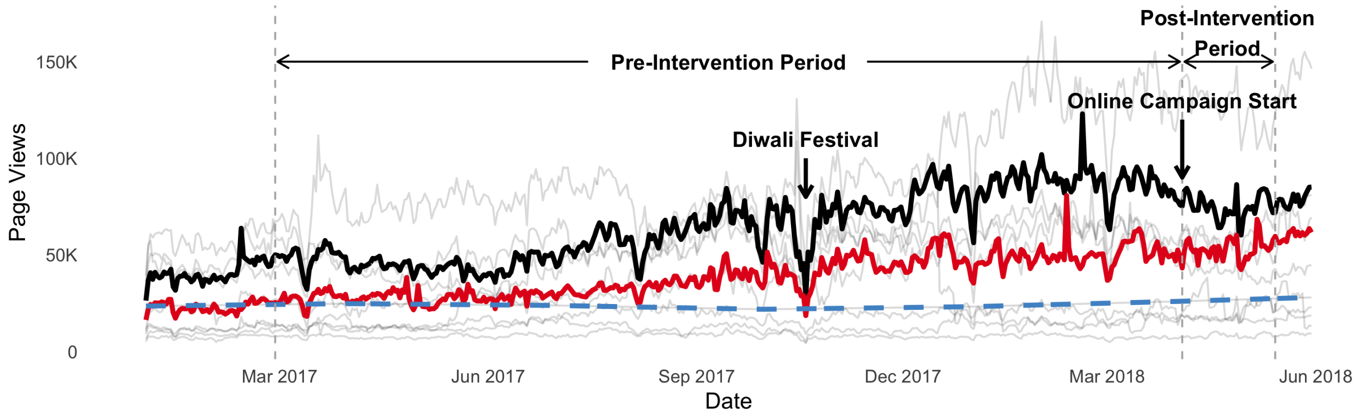

Figure 1: Hindi Wikipedia daily externally-referred page views from the top 10 states with the most page views in India. The black (top highlighted) series is the treated series: page views from Madhya Pradesh. The red (bottom highlighted) series is one of the most predictive control series according to the model in §4.2: page views from the neighboring state of Rajasthan. The blue dashed line is a covariate: interpolated number of internet subscribers in thousands from Madhya Pradesh. The intervention under study (online video campaign) started on 3 April 2018.

- **Trend Model**: the architecture for the model that predicts how the treated time series evolves. For each model in this work, we explore the following types of trend components using grid-search: local level, local linear, semi-local linear, and static intercept term.

4 RESULTS

Below we describe the context for two events on which we tested our BSTS model and their results.

4.1 Wikipedia Page Previews

Beginning in 2014, Wikipedia began exploring a new feature that would allow for page previews for the desktop version of the site. When a user moused over a link, a card would appear with an image and part of the first paragraph from the article that the link pointed to (see the Wikimedia Blog post[5] for more information and an example of a page preview on Wikimedia Commons[6]). This allows users to preview the article content without clicking on it (and thereby recording a page view). As a result of this ability to explore Wikipedia content without actually visiting the pages, it was expected that page views would actually drop with the roll-out of this feature.

In late 2017, the feature was finally rolled out to a proportion of anonymous users on German (de-wiki) and English (en-wiki) Wikipedia in a series of A/B tests. These tests were analyzed and it was determined that page preview feature led to a drop of approximately 4% in page views across these Wikipedia communities.[7] The full deployment of page previews to all anonymous users of these communities occurred on April 11 and 17 respectively. These A/B tests present us with an opportunity to explore the power of our BSTS models because they experimentally determined the expected effect on overall page views from the full deployment of page previews. Specifically, from the A/B tests starting on December 21 2017, in which 1.5% and 4% of anonymous users on English and German Wikipedias by default had access to page preview functionality respectively, we expect our BSTS model to detect:

- **de-wiki**: a 3.0% decrease in page views after April 11 2018.[8]
- **en-wiki**: a 4.7% decrease in page views after April 17 2018.[9]

4.1.1 Model Parameters. For each model, we set the pre-intervention period to be 400 days and the post-intervention period to be 6 weeks. For English Wikipedia, this means that the time series starts on 13 March 2017 and includes daily page views data through 28 May 2018, with the intervention occurring on 17 April 2018. For German Wikipedia, this means that the time series starts on 7 March 2017 and includes daily page views data through 22 May 2018, with the intervention occurring on 11 April 2018. Alongside day-of-week and monthly seasonality, we also include the following holidays: Christmas and New Year's. For trend modeling, we choose a static intercept term for German Wikipedia—i.e. we expect the trend of the time series to be soaked up by the regression component, and a local level model for English Wikipedia—i.e. the trend will be predicted around the weighted average values of recent observations.

For the control series, we rely on the assumption that while page previews should impact the internally-referred page views—i.e. page views as a result of navigating from one Wikipedia page to another—there is no reason that the previews would impact externally-referred page views—i.e. page views that result from someone navigating from a search engine or other, non-Wikimedia website—or direct page views without referrer. Specifically, we select the daily internally-referred page views from en-wiki or de-wiki as the treated time series. For our control time series, we include the daily externally-referred page views and direct page views from the same Wikipedia language edition as the treated time series under

[5] https://blog.wikimedia.org/2018/05/09/page-previews-documentation/
[6] https://commons.wikimedia.org/w/index.php?curid=47213242
[7] https://www.mediawiki.org/wiki/Page_Previews/2017-18_A/B_Tests
[8] https://phabricator.wikimedia.org/T191966
[9] https://phabricator.wikimedia.org/T191101

study (i.e. en-wiki or de-wiki), as well as daily externally-referred page views from the other top-20 largest Wikipedia editions (e.g. Russian and Spanish Wikipedia).

4.1.2 Results. We find that our time series models for both English and German Wikipedia are quite accurate. The validation statistics associated with the model provide an indication of how effective the model was at predicting the pre-intervention time series. With 10-fold cross validation and prediction evaluated on 6 weeks of daily page views (from the end of the pre-intervention period), the holdout mean absolute percentage error (MAPE) of the English Wikipedia model is 2.54%, and the holdout MAPE of the German Wikipedia model is 3.92%.

Turning to the estimate of the causal impact of the page preview roll-out, the treated time series along with counterfactual estimates from model for en-wiki are shown in Figure 2. Recall that the early A/B tests indicated that there would be a 4.7% decrease in page views for en-wiki. Our BSTS model, using the externally-referred and direct page views as control series, estimates a 3.0% decrease and correctly determines that no impact—i.e. 0% change—falls outside of the 95% credible interval [1.9%, 3.9%], indicating that the roll-out resulted in a significant change in page views. The most predictive control series in this model is the search engine referred page views on English Wikipedia with an average standardized coefficient of 0.65—i.e. when search engine referred page views change 1 standard deviation, we expect to see internally-referred page views change 0.65 standard deviation, and the posterior inclusion probability—i.e. the probability of this coefficient being different from zero—is 100%.

We see analogous results for de-wiki: the BSTS model estimated a 2.6% decrease in page views with a 95% credible interval of [1.9%, 3.4%], in line with the 3.0% decrease that had been determined via A/B testing. Similarly, the most predictive control series in this model is the search engine referred page views on German Wikipedia with an average standardized coefficient of 0.95 and a posterior inclusion probability of 100%.

4.2 Hindi Video Campaign

In India, only 33% of Hindi internet users have heard of Wikipedia [9] and, while there are 120,000 Wikipedia articles in Hindi, many people do not know that Hindi content exists. Meanwhile, internet access is growing 20%+ per year across India[10] and Hindi online content consumption is growing 94% per year[11]. In July 2017, the Wikimedia Foundation and the Hindi Wikimedians User Group began collaborating to reach "New Readers" in India through production and promotion of an online video.[12] The goal is to increase awareness and drive new usage of Wikipedia among Hindi speaking internet users.

To explain and promote Hindi Wikipedia (hi-wiki), the Wikimedia Foundation launched the video campaign on 3 April 2018. The Ektara[13] video was promoted on YouTube and Facebook targeting Hindi internet users in Madhya Pradesh, many of whom who had not heard of Wikipedia. The online promotion ran for three weeks and the video gathered 2.61 million views. This was followed by a second push over TV during a major Cricket event (on DD Sports during the Indian Premier League finals) on 27 May 2018 to the whole country, which reached 1.37 million viewers.[14]

4.2.1 Model Parameters. The pre-intervention period is 400 days and the post-intervention period is 6 weeks. Alongside day-of-week and monthly seasonality, we also include the following major Hindu holidays: Diwali, Raksha Bandhan, Holi, Dussehra, and New Year.[15] Local level model and a static intercept term are chosen as the trend for online campaign model and TV campaign model respectively.

For the evaluation of the impact of the online campaign, we set the treated time series to be daily externally-referred page views to hi-wiki from the Indian state of Madhya Pradesh (as determined by IP geolocation) because the promotion of the online video campaign was targeted at Madhya Pradesh. We select the externally-referred page views because it is a good indicator of the general brand awareness. The time series starts on 27 February 2017 and includes daily page view data through 14 May 2018, where the intervention occurred on 3 April 2018. For the control series, we use daily hi-wiki page views, as well as page views to other popular Wikipedia language editions and Wikimedia projects,[16] from the rest of India by states. We also included the daily number of internet subscribers in Madhya Pradesh as a covariate, which is linearly interpolated from a quarterly series reported by Telecom Regulatory Authority of India.[17]

For the evaluation of the impact of the TV campaign, we set the treated time series to be daily externally-referred hi-wiki page views from the entire country of India because there was no state-specific targeting of the campaign. The time series starts on 22 April 2017 and includes daily page view data through 7 July 2018, where the intervention occurred on 27 May 2018. For the control series, we use daily hi-wiki page views and page views to other popular Wikipedia language editions and Wikimedia projects from other countries[18]. Additionally, we included the daily number of internet subscribers in India as a covariate, which is linearly interpolated from a quarterly series reported by Telecom Regulatory Authority of India.

4.2.2 Results. For both the online and TV campaigns for Hindi Wikipedia, we do not detect a significant change in page views. As we discuss below and in §5.1, this is likely a combination of factors: low impact and imprecise control series. The results from the BSTS models for the online campaign are in Figure 3 and TV campaign in Figure 4.

First we examine the results for the online video campaign that was targeted at the state of Madhya Pradesh. As before, we performed a 10-fold cross validation with the pre-intervention time

[10] http://www.internetlivestats.com/internet-users/india/

[11] https://economictimes.indiatimes.com/tech/internet/hindi-content-consumption-on-internet-growing-at-94-google/articleshow/48528347.cms

[12] https://meta.wikimedia.org/wiki/New_Readers/Raising_Awareness_in_India

[13] https://commons.wikimedia.org/wiki/File:Wikipedia_-_Ektara_(English_subtitles).webm

[14] TV data was collected by Eurodata TV via BARC in India.

[15] These Hindu festivals are picked because of their relative big impact on the treated time series. Their dates of each year are obtained from https://www.officeholidays.com/countries/india/index.php

[16] We selected the top 10 Wikimedia projects in India with the most page views, and Wikipedia of major Indian languages spoken by more than 4% of the population, according to 2011 census of India.

[17] https://www.trai.gov.in/release-publication/reports/performance-indicators-reports

[18] Countries that contribute more than 5% of Hindi Wikipedia page views, countries whose official language is Hindi, and other nearby countries.

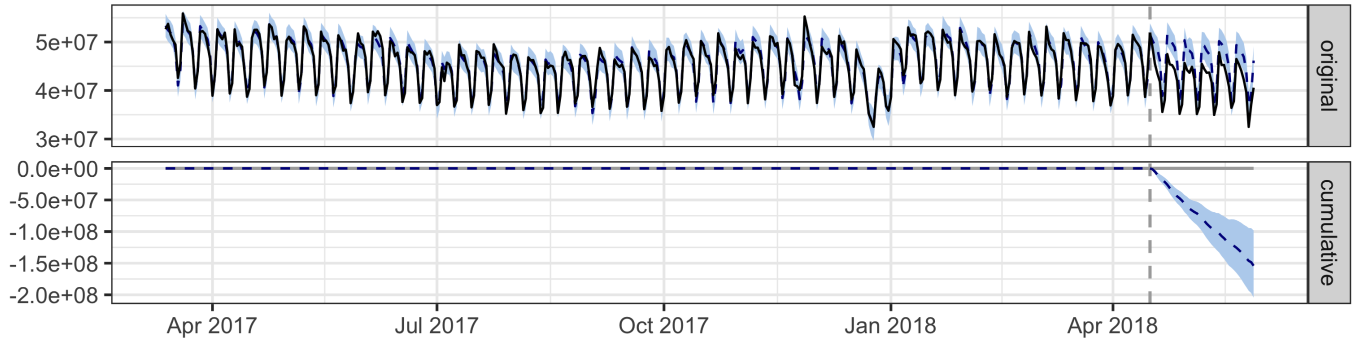

Figure 2: Impact of the page preview feature on 6 weeks of English Wikipedia internally-referred page views. Vertical dashed line represents the date of the roll-out (17 April 2018). Shaded areas indicate 95% credible intervals. The first panel shows the data (black solid line) and counterfactual prediction (blue dashed line) for the post-intervention period. The second panel sums the difference between observed data and counterfactual predictions—i.e. point-wise causal effect as estimated by the model, resulting in a plot of the cumulative effect of the intervention. The figures of point-wise causal effect are removed in this paper for space consideration.

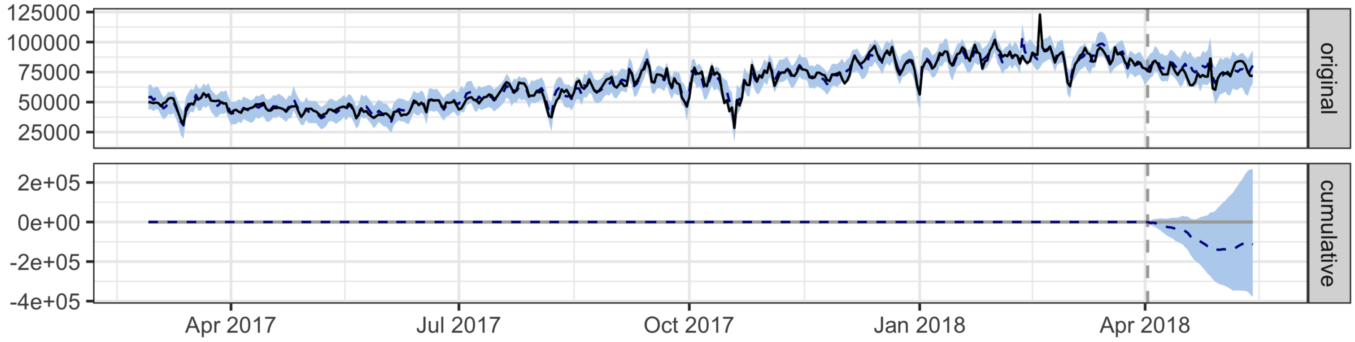

Figure 3: Impact of the online campaign in 6 weeks on Hindi Wikipedia externally-referred page views from Madhya Pradesh. Vertical dashed line represents the start date of the campaign 3 April 2018.

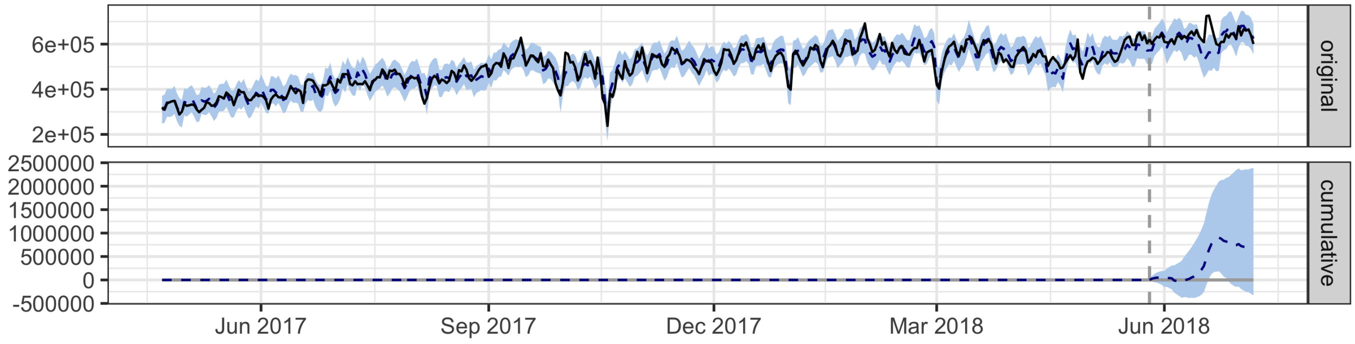

Figure 4: Impact of the TV campaign in 6 weeks on Hindi Wikipedia externally-referred page views from all of India. Vertical dashed line represents the date of the campaign 27 May 2018.

series and predict 6-week's daily page views in each fold. The average holdout MAPE is 7.6%.

As Figure 3 indicates, no significant impact on page views was detected following the intervention. While there does appear to be a downward trend in page views, zero change is still within the 95% credible interval. Page views to hi-wiki from the states of Rajasthan and Chhattisgarh, both of which border Madhya Pradesh, are the most predictive control series in the model, with average standardized coefficients of 0.23 and 0.13 respectively. The posterior probabilities that their coefficients are different from zero are greater than 95%.

Next we turn to the results for the country-wide TV campaign. The average holdout MAPE from the cross validation is 10.2%. Figure 4 shows the results from the BSTS model. As the graphs indicate, no significant impact on page views was detected in the first 3 weeks following the intervention. There was a bump in the 4th week after the campaign, but it is most likely to be the result of an unknown event. Overall, we did not detect significant impact in 6 weeks. The number of internet subscribers in India, hi-wiki page views from the United States, Bengali Wikipedia page views from Bangladesh and English Wiktionary pageviews from Nepal are the most predictive control series in the model, with an average standardized coefficient of 0.84, 0.35, 0.14 and 0.16 respectively. The posterior probabilities that their coefficients are different from zero are greater than 95%.

We did not include the page views to other Wikipedia language editions that also were geolocated to India—e.g., page views to en-wiki from India—in the set of control series because it may violate the independence assumption. Most people in India are multilingual, so if our brand awareness was affected by the campaign, the impact would likely be revealed on page views of other Wikipedia language editions from the target region as well. After seeing the relatively high MAPE (10.2%) of the TV campaign model, we tried to include the page views to other Wikipedia language editions and other Wikimedia projects from India into the model to see if they help. The average holdout MAPE decreased to 8.5%, but that model also does not detect significant impact.

5 DISCUSSION

5.1 Choosing a Control Time Series

A predictive control series is one of the most important aspects of a BSTS model and also the part of the model that is often most difficult to choose. From the page preview roll-out analysis (§4.1), we see that the BSTS model, with a well-chosen control time series, can effectively estimate the impact of a given external event. In that case, externally-referred and internally-referred page views for the same project are highly correlated, but only internally-referred page views were believed to be impacted by the page preview roll-out. For both German and English Wikipedia, the estimated causal impact was slightly conservative—i.e. lower in magnitude than expected based on the A/B tests—but still quite close to the expected impact. We had less success selecting an effective control series for the Hindi video campaign. Furthermore, the fact that the online campaign model for Madhya Pradesh had less error than the TV campaign model for all of India raises questions about how factors like geography or language edition affect the predictive power of a control series for Wikipedia.

To better understand the power of different types of control series, we tested four additional control time series models for the Hindi analysis. All are trained on 400 days of daily externally-referred page views and evaluated via 10-fold cross validation on 6 weeks of page views prior to the intervention (3 April 2018). All models include day-of-week and monthly seasonality, holiday effects, local level trends, and a set of control time series but no further covariates. The control series for each model and its respective validation error are shown in Table 1.

We see that the control time series that are from the same language edition as the treated time series (Models 1 and 3) have a consistently lower error than the models that are from different language edition but the same geographic region (Models 2 and 4). This indicates that language edition plays a more important role than geographic region in the page view trends on Wikipedia. Comparing Model 1 and 3 where the control and treated series are from the same language edition, page views between states within the same country are more predictive of each other (adding these controls into the model decrease the MAPE from 11.54% to 7.54%) than page views between different countries (adding them into the model only decrease the MAPE from 7.92% to 7.22%), which indicates that while language appears to be most important, country borders are still a highly salient aspect of page view trends on Wikipedia. Further research would be needed to understand how these effects play out in other language communities and the inter-relatedness of different countries and language pairs.

6 FUTURE WORK AND LIMITATIONS

While this work has a number of limitations, as we lay out below, we believe it lays the groundwork for exploring more standardized methods of predicting trends such as page views on Wikipedia with the goal of understanding the effect of external events. Limitations for this work largely relate to temporal evolution of impact, data pre-processing, prior distributions of parameters in BSTS, and the need for additional experiments.

In this work, we focus only on the cumulative effect by the end of the post-intervention period—its existence and magnitude—without discussing the temporal evolution of an impact. In practice, how an effect evolves over time, especially its onset and decay structure, is often a valuable question as well. The point-wise effect from BSTS reflects the temporal evolution and future implementation should consider analyzing this result.

Small volume Wikipedia editions such as hi-wiki are more sensitive to undetected bot behavior, which can cause anomalies in page-view data. Anomalies in the prediction or post-intervention period would increase the error rate of validation, or the model might detect an impact that is unrelated to the known intervention. When the number of control series is very large, removing outliers manually is not feasible and thus requires a robust algorithm to detect and adjust outliers while preserving those known "outliers" such as holiday effects. It is possible that further pre-processing would also provide benefits—e.g., including more holidays, removing seasonal patterns in predictors before fitting the model, more extensive grid search for parameters like the length of pre-intervention period.

We were expecting that spike-and-slab prior in the BSTS would prevent over-fitting by forcing the coefficients of poor predictors to zero, so we would at least have predictions not worse than that of a model which only contains the historical information of the treated series itself. Contrary to this expectation, Model 4 from Table 1 (predicting hi-wiki page views in India) shows that including control series from other Wikipedia language editions and other Wikimedia projects within India actually added noise to the prediction. To solve

Table 1: Comparison of the predictive power of four sets of control time series for the Hindi Wikipedia campaign, of which the treated series and the control series either share the same geographic region or the same Wikipedia language edition.

	Treated Series (lang; region)	Control Series (lang; region)	Avg MAPE	Avg MAPE (No Control Series)
Model 1	hi-wiki page views; Madhya Pradesh	hi-wiki page views; other states of India	7.54%	11.54%
Model 2	hi-wiki page views; Madhya Pradesh	other wikis page views; Madhya Pradesh	9.31%	11.54%
Model 3	hi-wiki page views; all of India	hi-wiki page views; other countries	7.22%	7.92%
Model 4	hi-wiki page views; all of India	other wikis page views; all of India	9.14%	7.92%

this problem, we can further tune the hyper-parameter that controls the expected model size—the expected number of coefficients that are different from zero—so that when most of the predictors do not have enough predictive power, we can lower the expected model size and force more coefficients to be zero (we set the expected model size to be 10% of total number of controls in the model but not greater than 5 in this work).

Finally, future work should continue to explore these models in more contexts. This would hopefully provide guidance for how to select control time series—e.g., which pairs of regions and language editions (or even other Wikimedia projects) are predictive, what is the best way to split Wikipedias into control and treatment articles, how to take advantage of more granular information as with the internal/external referrer information. This would also hopefully provide guidance for how to set priors in BSTS models—e.g., a prior likelihood of relationships between treatment and control time series or prior standard deviation of the Gaussian random walk of the trend models (conservatively, we use a non-informative prior for the former and 0.01 for the latter). In future analyses, we can increase the prior inclusion probabilities for control series that are likely to be correlated with the treated series, and increase the prior standard deviation for the trend models if we believe the volatility of residuals would be large after regressing out known predictors.

ACKNOWLEDGMENTS

We would like to thank Tilman Bayer, Mikhail Popov, Leila Zia, Kate Zimmerman and Jon Katz for guidance and support as well as our anonymous reviewers for their feedback.

REFERENCES

[1] Alexa. 2018. *wikipedia.org Traffic Statistics.* Technical Report. Alexa. https://www.alexa.com/siteinfo/wikipedia.org

[2] Joshua D. Angrist and Pischke Jörn-Steffen. 2009. *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton University Press.

[3] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. 2002. How Much Should We Trust Differences-in-Differences Estimates? (2002). https://doi.org/10.3386/w8841

[4] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9, 1 (2015), 247–274. https://doi.org/10.1214/14-aoas788

[5] Jennifer L. Castle, Xiaochuan Qin, and W. Robert Reed. 2009. How To Pick The Best Regression Equation: A Review And Comparison Of Model Selection Algorithms, by Jennifer L. Castle; Xiaochuan Qin; W. Robert Reed. https://ideas.repec.org/p/cbt/econwp/09-13.html

[6] Justin Clark, Robert Faris, and Rebekah Heacock Jones. 2017. *Analyzing Accessibility of Wikipedia Projects Around the World.* Berkman Klein Center for Internet & Society Research Publication.

[7] Ed Erhart. 2019. Wikipedia's most-popular articles of 2018 show that pop culture rules over us all. https://wikimediafoundation.org/2019/01/02/wikipedias-most-popular-articles-of-2018-show-that-pop-culture-rules-over-us-all/

[8] E. I. George and R. E. McCulloch. 1997. Approaches for Bayesian variable selection. *Statistica Sinica* 7 (1997), 339–374.

[9] Satdeep Gill and Zack McCune. 2018. How we're building awareness of Wikipedia in India. https://blog.wikimedia.org/2018/04/03/building-awareness-wikipedia-india/

[10] Andrew C. Harvey. 1991. *Forecasting, structural time series models and the Kalman filter.* Cambridge Univ. Press.

[11] D. Madigan and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* 89 (1994), 1535–1546.

[12] Márton Mestyán, Taha Yasseri, and János Kertész. 2013. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one* 8, 8 (2013), e71226.

[13] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H Eugene Stanley, and Tobias Preis. 2013. Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports* 3 (2013), 1801.

[14] Daniel Cheng Moyer, Samuel L Carson, Thayne Keegan Dye, Richard T Carson, and David Goldbaum. 2015. Determining the Influence of Reddit Posts on Wikipedia Pageviews. In *Ninth International AAAI Conference on Web and Social Media.*

[15] Jonathon W Penney. 2016. Chilling effects: Online surveillance and Wikipedia use. *Berkeley Tech. LJ* 31 (2016), 117.

[16] Reid Priedhorsky, Dave Osthus, Ashlynn R Daughton, Kelly R Moran, Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, and Sara Y Del Valle. 2017. Measuring global disease with Wikipedia: Success, failure, and a research agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* ACM, 1812–1834.

[17] Steven L Scott. 2017. Fitting Bayesian structural time series with the bsts R package. http://www.unofficialgoogledatascience.com/2017/07/fitting-bayesian-structural-time-series.html

[18] Steven L Scott and Hal R Varian. 2013. *Bayesian Variable Selection for Nowcasting Economic Time Series.* Working Paper 19567. National Bureau of Economic Research. https://doi.org/10.3386/w19567

[19] Steven L Scott and Hal R Varian. 2013. Predicting the present with bayesian structural time series. *Available at SSRN 2304426* (2013).

[20] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 1591–1600.

[21] Marijn ten Thij, Yana Volkovich, David Laniado, and Andreas Kaltenbrunner. 2012. Modeling and predicting page-view dynamics on Wikipedia. *CoRR abs/1212.5943* (2012).

[22] Giorgino Toni. 2009. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software* 31 (2009). https://doi.org/10.18637/jss.v031.i07

[23] Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 566.

[24] Xiaoquan Michael Zhang and Feng Zhu. 2011. Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101, 4 (2011), 1601–15.