Estimating Ad Effectiveness using Geo Experiments in a Time-Based Regression Framework

Jouni Kerman, Peng Wang, and Jon Vaver

Google, Inc.

March 2017

Abstract

Two previously published papers (Vaver and Koehler, 2011, 2012) describe a model for analyzing geo experiments. This model was designed to measure advertising effectiveness using the rigor of a randomized experiment with replication across geographic units providing confidence interval estimates. While effective, this geo-based regression (GBR) approach is less applicable, or not applicable at all, for situations in which few geographic units are available for testing (e.g. smaller countries, or subregions of larger countries) These situations also include the so-called matched market tests, which may compare the behavior of users in a single control region with the behavior of users in a single test region. To fill this gap, we have developed an analogous time-based regression (TBR) approach for analyzing geo experiments. This methodology predicts the time series of the counterfactual market response, allowing for direct estimation of the cumulative causal effect at the end of the experiment. In this paper we describe this model and evaluate its performance using simulation.

1 Introduction

The ability to measure the impact and the return on investment of online ad campaigns is essential for advertisers. Proper measurement facilitates informed strategic and tactical decision-making regarding campaign management. Unfortunately, generating truly causal measurements is nontrivial; suitable causal estimation methodologies have been discussed widely (see e.g., Chan et al., 2010; Ye et al., 2016; Johnson et al., 2015; Chan et al., 2011). Thoughtful analysis of historical observational data may provide hints and hypotheses regarding ad effectiveness. However, such observational studies are not usually capable of assessing true causal effects, and are susceptible to biased estimates of return on investment (Lewis and Rao, 2014; Gordon et al., 2016). Access to reliable causal information requires the rigor of a controlled experiment, such as the randomized controlled trial, which is considered the gold standard for inferring causal assessments.

Measurement needs are varied and come with a variety of limitations, e.g., data availability, ad targeting restrictions, wide-ranging measurement objectives, budget availability, time constraints, etc. This diversity requires flexible analysis methodologies or, in some situations, the development of alternative analysis methodologies to provide advertisers

with options for rigorous measurement. Geo experiments (Vaver and Koehler, 2011, 2012) meet a large range of measurement needs. They use non-overlapping geographic regions, or simply "geos," that are randomly, or systematically, assigned to a control or treatment condition. Each region realizes its assigned treatment condition through the use of geo-targeted advertising. These experiments can be used to analyze the impact of advertising on any metric that can be collected at the geo level. Geo experiments are also privacy-friendly since the outcomes of interest are aggregated across each geographic region in the experiment. No individual user-level information is required for the "pure" geo experiments, although hybrid geo + user experiments have been developed as well (Ye et al., 2016).

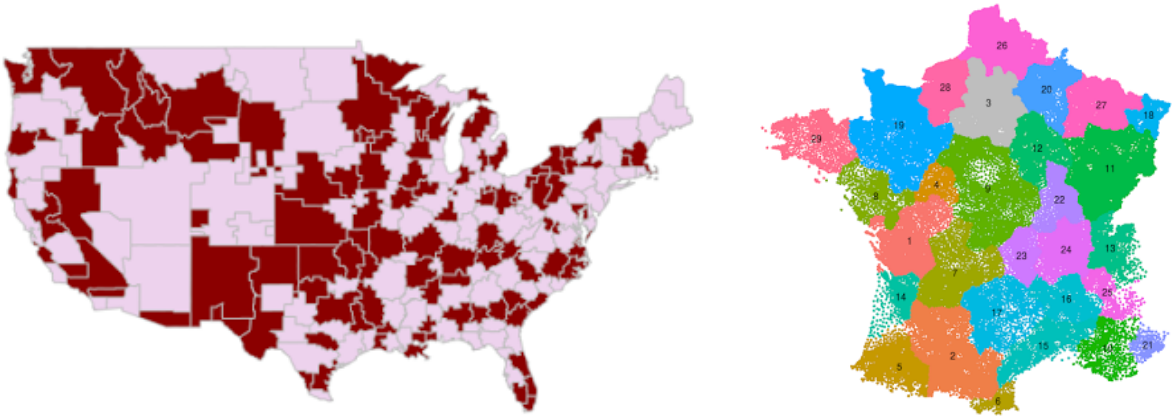
Matched market tests (see e.g., Gordon et al., 2016) are another specific form of geo experiments. They are widely used by marketing service providers to measure the impact of online advertising on offline sales. In these tests, geos are carefully selected and paired. This matching process is used instead of a randomized assignment of geos to treatment and control. Although these tests do not offer the protection of a randomization experiment against hidden biases, they are convenient and relatively inexpensive, since the testing typically uses a small subset of available geos. These tests often use time series data at the store level. Another matching step at the store level is used to generate a lift estimate and confidence interval.

An analysis methodology for geo experiments was introduced in Vaver and Koehler (2011). In this paper, we refer to this methodology as Geo-Based Regression (GBR). The GBR approach estimates the incremental Return On Ad Spend (iROAS) directly using a linear regression model. However, GBR draws its statistical power from the number of available geos in the analysis, and therefore is not applicable for situations with only a few geos available (such as matched market tests).

Brodersen et al. (2015) introduced a time-series based analysis methodology, referred to as *Causal Impact*. This diffusion-regression state-space model was developed to analyze observational data by selecting and applying synthetic controls. While it can be applied to experimental data, including experimental data with a small number of geos, it includes more flexibility and complexity than is required for these analyses.

Time-Based Regression (TBR) offers an alternative analysis approach. It can be applied in any situation in which GBR is applicable without any change to experiment execution. Like Causal Impact, TBR uses the observed time series in the pretest period as training to predict the counterfactual time series during the marketing intervention period and beyond. TBR can be applied to an experiment with a limited number of geos, including the case in which there is only one test geo and one control geo (e.g., a matched market test).

Section 2 briefly reviews the key concepts of a geo experiment: the geographical areas (geos), and the description of time periods. Section 3 introduces the TBR model and the estimation procedure. Section 4 discusses the experiment design process along with the design parameters and their impact on the estimation precision. Section 5 contains a performance assessment, which includes evaluation of the bias and coverage of the estimation procedure across various scenarios. Finally, we summarize the paper in Section 7. Appendix (Section 9) presents some technical calculations useful for implementing the method, and compares the TBR model to an alternative model form that applies to a slightly different analysis scenario.



(a) A randomized assignment of the 210 Designated Marketing Areas (DMAs) of the United States.

(b) A partition of France into 29 geos.

Figure 1: Examples of sets of geos. On the left, the U.S. market is divided into 210 standard geos, the Nielsen DMAs, that can be directly targeted in Google AdWords. On the right, geos constructed for France using Google AdWords Cities.

2 Structure of a Geo Experiment

All geo experiments have a common structure. In this section we describe this structure to facilitate the subsequent description of TBR.

2.1 Geographic areas

Setting up a geo experiment begins by segmenting a geographic area (such as a country) into a set of sub-regions, or "geos." It must be possible for the ad system under test to individually target each geo. For example, Nielsen Media has partitioned the United States into 210 geos. These regions, called Designated Marketing Areas (DMAs, The Nielsen Company, 2017), were originally formed based on television viewing behavior of their residents (see Figure 1a). The DMAs are one example of a set of geos that can be targeted individually in Google AdWords. Other countries, such as France, do not have a set of geos analogous to DMAs. Figure 1b shows a set of geos that were constructed from the set of Google AdWords Cities, which are smaller geo-targetable units in AdWords (Google, Inc., 2017a).

The set of geos used in an experiment may be a subset of the available geos. Once the set of experiment geos is specified, they are assigned to treatment and control groups. The assignment can be randomized (possibly using stratification), or "matched," or anything in between, depending on the requirements and the analysis method that will be used. The randomized assignment of a large number of geos provides the most protection against known and unknown differences across geos, while a matched market test deliberately assigns as few as one geo (e.g., a DMA) to the treatment group and one to control.

2.2 Pretest, intervention, and cooldown periods

A geo experiment consists of several, distinct time periods: the pretest, intervention, and cooldown periods. The latter two periods combined make up the 'test' period.

During the pretest period, ad campaigns in the treatment and control geos are in their unmodified base state. All geos operate with the same baseline level campaign settings (e.g., common bidding, keyword lists, ad targeting, etc.).

Ad campaigns are modified in the treatment geos during the intervention period. We expect that this modification, by design, will cause a change effect on ad spend. What we do not know is whether it also will cause a significant effect on the response metric, e.g., sales. It is worth noting that modified campaigns may cause the ad spend either to increase (e.g., add keywords or increasing bids in the ad campaign) or decrease (e.g., turn campaigns off). Either way, we typically expect the response metric to be affected in the same direction as the spend.

Finally, ad campaigns are reset to their original state during the cooldown period. This does not always mean their effect will cease instantly. Incremental offline sales, for example, may continue to accrue across subsequent days or even weeks. Including data from the cooldown period in the analysis makes it possible to capture these lagged effects from the advertising change.

3 Time-Based Regression

3.1 Data Aggregation

Both Geo-Based Regression (GBR) and Time-Based Regression (TBR) analyses use data with the same canonical structure. A data set consists of several concurrent time series (of the response metric of interest), one for each geo; see Figure 2a.

In GBR, the data are are aggregated over time to provide an observation of the response metric volume across the entire pretest period and the entire test period (that is, the intervention and cooldown periods) for each geo, resulting in a data set with two data points per geo: see Figure 2b.

In contrast, for TBR, data are aggregated across geos to provide one observation of response metric volume at every time interval for the control and treatment groups: see Figure 2c. The time series are most commonly daily or weekly.

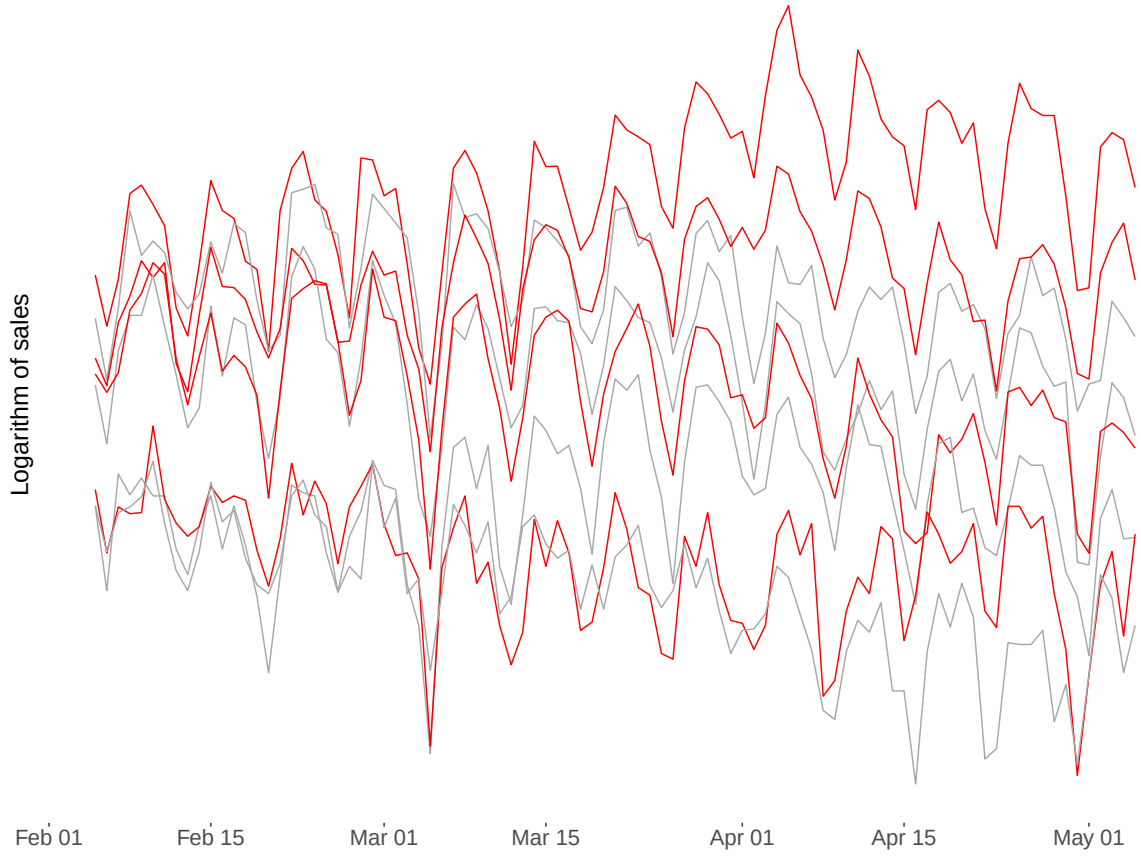
3.2 TBR Causal Effect Analysis

The primary objective of TBR is to estimate the incremental, causal effect that an advertising intervention had on a response metric, such as sales revenue, but also on the incremental cost of the marketing intervention. The incremental return on ad spend (iROAS) can be computed using these two quantities, which is a more useful measure for evaluating the effectiveness of an intervention (See Section 3.4).

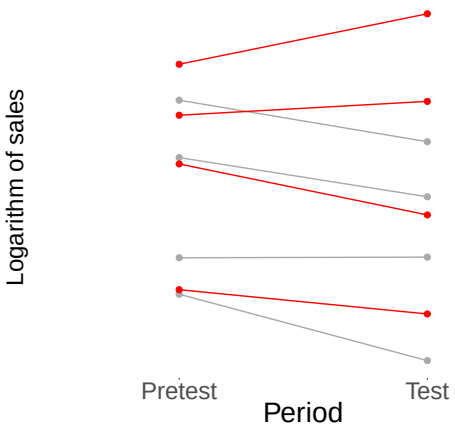
The TBR Causal Effect Analysis uses the treatment and control group time series, y_t and x_t, respectively, from the pretest period to estimate the relationship between the treatment and control groups for the outcome metric prior to the marketing intervention. This relationship can take different forms depending on the assumptions we are willing to make. The simplest possible relationship is given by the regression model,

$$y_t = \alpha + \beta x_t + \epsilon_t, \quad t \text{ in pretest period}, \tag{1}$$

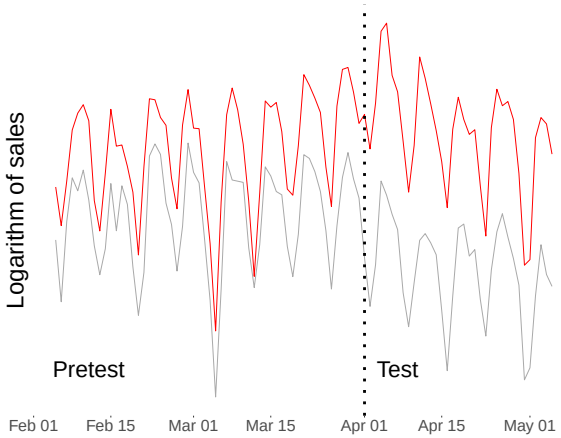
where ϵ_t are independent Normal errors with standard deviation σ. This model assumes that the control time series acts as a sufficient predictor for the behavior of the treatment



(a) Logarithm of sales plotted for 8 geos from a geo experiment.



(b) Same data set, aggregated for the GBR method.



(c) Same data set, aggregated for the TBR method.

Figure 2: Example of a geo experiment data set, aggregated for Geo-Based Regression (GBR) and Time-Based Regression (TBR) methods. Red color: treatment group, gray color: control group.

time series across the pretest, intervention, and cooldown periods. The relationship is therefore assumed to be stable, (i.e., the residuals are assumed to be stationary) in the absence of the experiment-related marketing intervention.

First, this model is trained using data from the pretest period to estimate the unknown parameters (α, β, σ)—along with their uncertainty. Then, this fitted model is used to predict the counterfactual behavior of the treatment group during the test period. For each t in the test period, the main quantities of interest are the counterfactual observations y_t^* (the potential outcomes):

$$y_t^* = \alpha + \beta x_t + \epsilon_t^*, \quad t \text{ in the test period}$$

The estimation of y_t^* includes both the estimated uncertainty in (α, β, σ) *and* the prediction uncertainty in the new, unobserved errors ϵ_t^*. Bayesian inference is used to estimate the unknowns and to derive the posterior predictive distribution of y_t^* for each t. Assuming a standard noninformative distribution for the parameters, the joint posterior distribution of y_t^* follows a shifted and scaled t-distribution, making the inferences straightforward. See Gelman et al. (2013) and Section 9.1 below for details.

Having estimated the counterfactual time series y_t^*, the causal estimates of the effect are the differences between the observed response metric volumes in the treatment group, y_t, and the potential outcomes y_t^*,

$$\phi_t = y_t - y_t^*, \quad t \text{ in the test period.}$$

Since y_t^* have posterior (uncertainty) distributions and y_t are fixed, ϕ_t also have a posterior distribution.

Finally, the cumulative causal effect at time t during the test period is the sum of these differences up to time t starting from the first day of the intervention period, $t' = 1$.

$$\Delta(t) = \sum_{t'=1}^{t} \phi_{t'}.$$

The posterior distribution of $\Delta(t)$ is again a shifted and scaled t-distribution. See Section 9.1 for a description of the process for recovering the uncertainty associated with the partial sums without resorting to simulation.

3.3 TBR Causal Effect Analysis: Example

A geo experiment was executed in the U.S. using all 210 DMAs, which were randomly assigned to a treatment and a control group. On the first day of the intervention period, April 1, a set of Paid Search ad campaigns were modified in all treatment geos, while the campaigns were unchanged in the control geos. The response metric was daily sales. The marketing intervention caused the incremental number of impressions and clicks to increase, and consequently the incremental cost of the campaigns increased. On April 29, four weeks after the start of intervention, the campaigns in the treatment geos were reset to their original configuration.

The following analysis uses a pretest period of eight weeks and a cooldown period of one week. Figure 3 shows the results of the analysis for the revenue. The first panel shows the actual, observed treatment group time series, y_t, along with the median and the 90% middle posterior intervals of the predicted counterfactual time series, y_t^*.

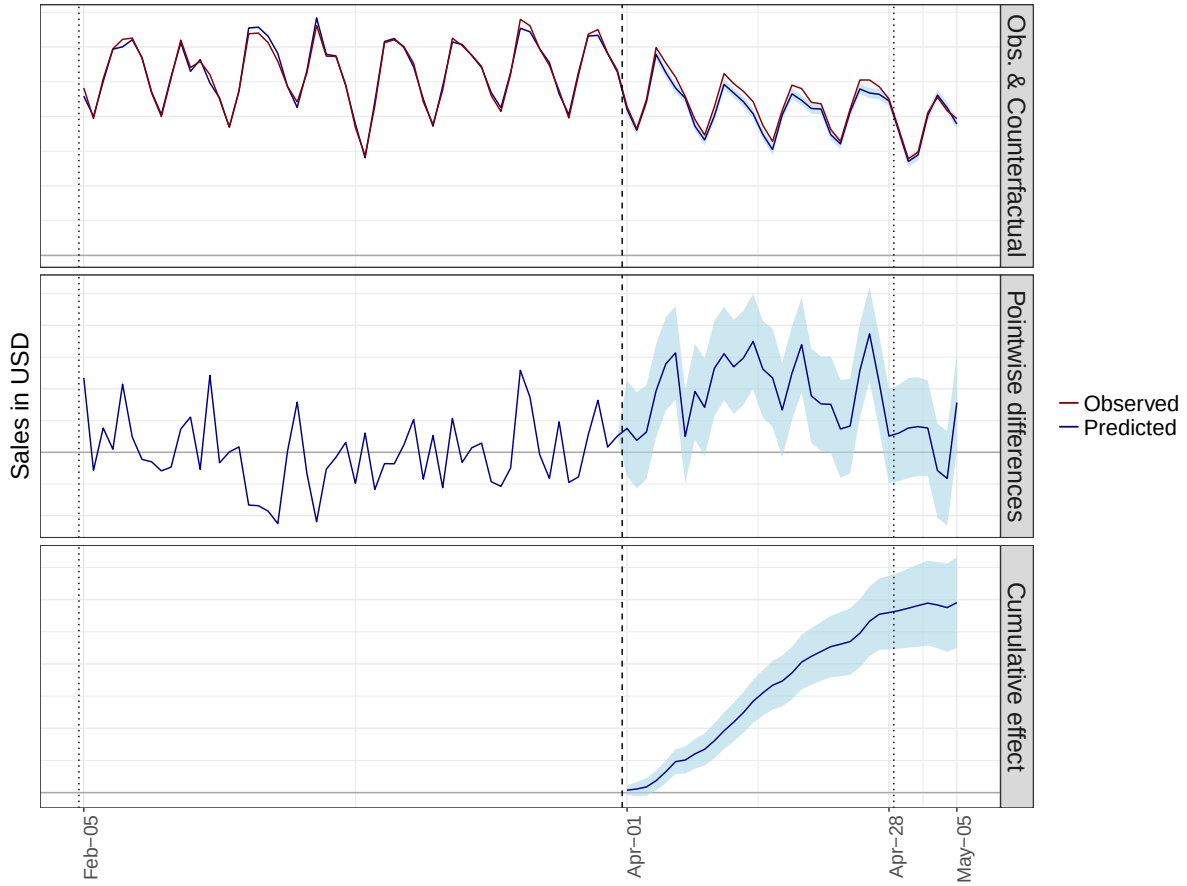


Figure 3: Example: TBR Causal Effect Analysis of Incremental Revenue. The panel is divided into pretest, intervention, and cooldown periods, starting on February 5, April 1, and April 29, respectively. Note that the scales on the vertical axes are different in each panel. The three panels show (a) the observed treatment time series y_t along with its estimated counterfactual y_t^*; (b) the pointwise (daily) causal effects ϕ_t, and (c) the cumulative causal effect $\Delta(t)$.

The median estimate and the corresponding 90% posterior intervals for the "pointwise" differences (i.e., daily differences in this particular data set) ϕ_t are shown in the middle panel. Starting from the intervention period (April 1), the differences correspond to the daily incremental revenue ϕ_t, showing positive incremental effects. The pointwise differences for the pretest period (February 5 – April 1) are given for the purpose of a visual diagnostic only, as they correspond to the residuals of the model.

The bottom panel shows, similarly, the estimates of the cumulative incremental revenue $\Delta_{revenue}(t)$. These results show that the cumulative incremental revenue increases up to the end of the marketing intervention on April 28, and flattens out afterward. Observing the extent of the lagged impact on a response metric is a useful diagnostic. Nothing is plotted for the pretest period, since a plot of cumulative residuals does not serve as a useful regression diagnostic.

Figure 4 shows a similar TBR Causal Effect analysis for incremental ad spend (cost of clicks) $\Delta_{cost}(t)$. Due to the marketing intervention, incremental cost of clicks increased during the test period. We also see that the incremental cost did not continue to accumulate beyond the end of the test period, as expected, since the marketing intervention should not have a lagged impact on spend.

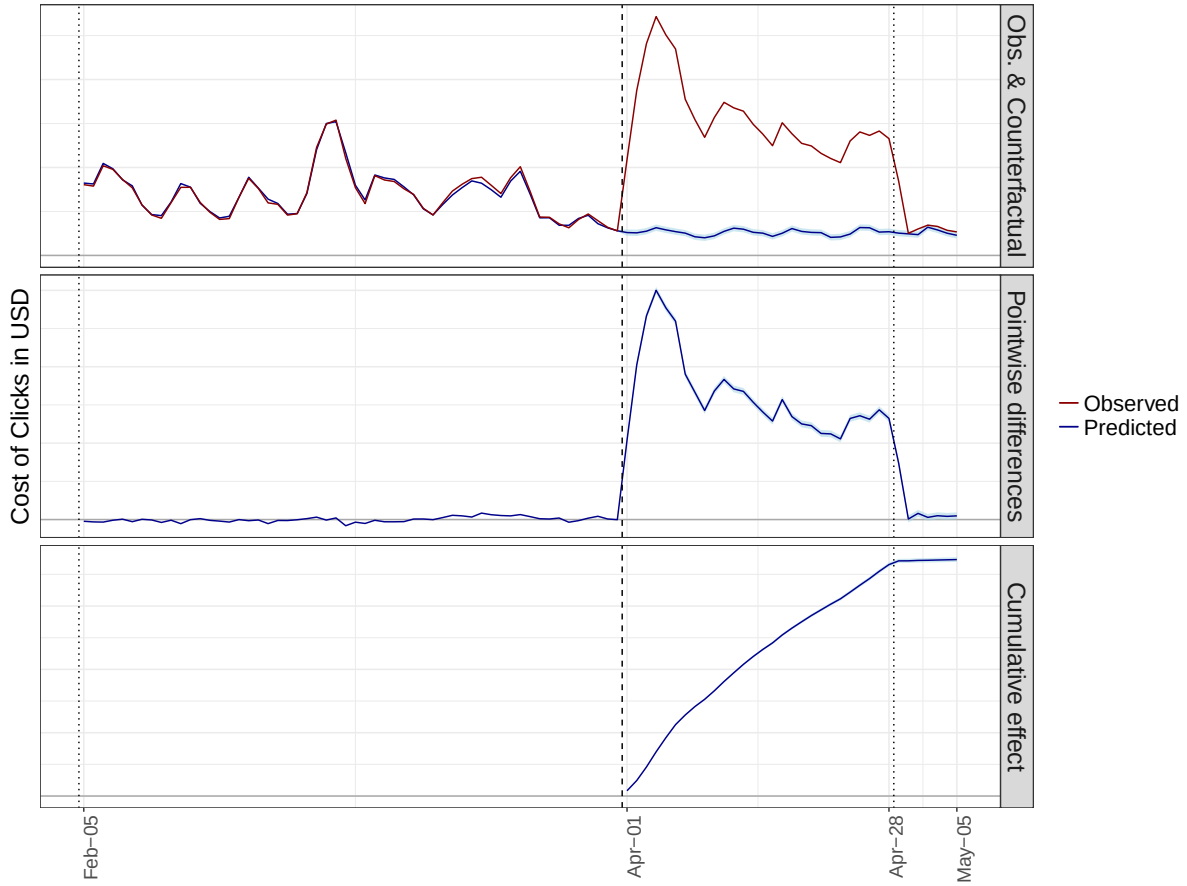


Figure 4: Example: TBR Causal Effect Analysis of Incremental Cost of Clicks. The estimated effect is very large compared to the estimation uncertainty, which makes the posterior intervals very narrow.

3.4 TBR Incremental ROAS Analysis

Knowing the causal effect of a marketing change on a single response metric, such as sales, is informative. However, understanding the efficiency of a marketing change also requires knowledge of the cost associated with generating this change. iROAS is the ratio of the cumulative causal effect on a response metric relative to the cumulative causal effect on marketing cost, $\text{iROAS}(t) = \Delta_{resp}(t)/\Delta_{cost}(t)$, for a time t in the test period. The main quantity of interest is the total cumulative iROAS, $\text{iROAS}(T)$, calculated for the last day T of the cooldown period.

Once the posterior distributions of the cumulative causal effects for both the response $\Delta_{resp}(t)$ and the cost $\Delta_{cost}(t)$ are recovered using TBR Causal Effect Analysis described in Section 3.2, the ratio is estimated using simulated draws from the two posterior distributions:

$$\text{iROAS}^{(s)}(t) = \Delta^{(s)}_{resp}(t)/\Delta^{(s)}_{cost}(t),$$

where $s = 1, \ldots, N$, for e.g., $N = 10000$. This can be done very rapidly since both the numerator and denominator can be simulated from their respective t distributions. As point estimates of $\Delta_{resp}(t), \Delta_{cost}(t)$, and $\text{iROAS}(t)$, we chose the medians of the corresponding posterior distributions.

As a special case, if the cost metric is zero during the pretest period—which would happen if the media in question is used for the first time as the marketing intervention—the

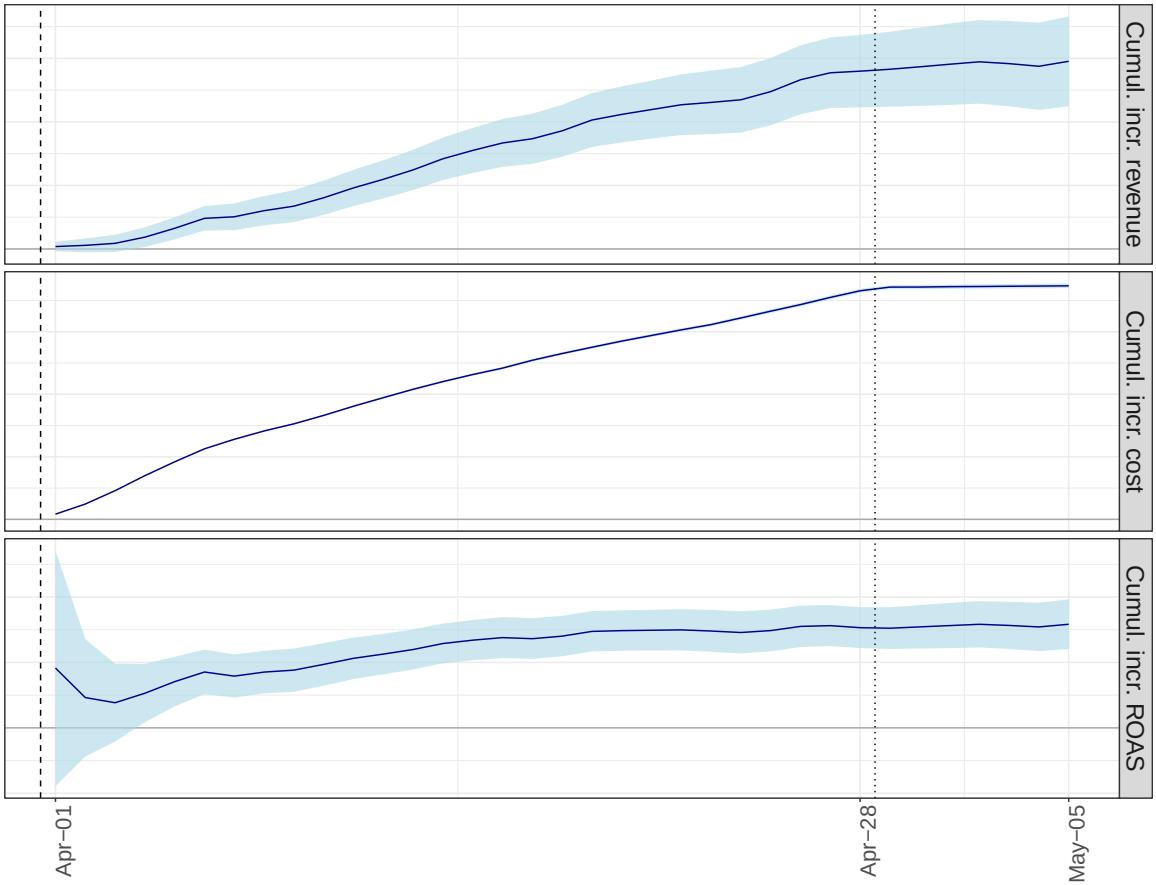


Figure 5: Example: The iROAS estimate from the TBR analysis. This plot shows the cumulative incremental revenue, cumulative incremental cost, and the iROAS, the ratio of the two.

counterfactual predictions are all zero, with complete certainty. This implies that $\Delta_{cost}(t)$ is simply the total sum of cost generated by the test. Hence, since $\Delta_{cost}(t)$ is a constant in this case, the posterior distribution of iROAS(t) is again a shifted and scaled t-distribution, making it unnecessary to resort to simulations as the quantiles of the t-distribution are readily available from statistical software packages.

3.5 TBR Incremental ROAS Analysis: Example

Figure 5 summarizes the iROAS estimation of the example of Section 3.3, showing the cumulative incremental revenue, cost, and their ratio, and the iROAS. Only the test period is shown, as iROAS is not defined in the pretest period. The estimate of iROAS in the beginning of the intervention period is unstable, but stabilizes soon to its final estimate.

The plot of cumulative iROAS varies more significantly near the beginning of the test period and then stabilizes across time. If it is obvious from this plot that the incremental effects have ceased to increase beyond the end of the intervention period, there is no need to add a cooldown period, since doing so only adds more uncertainty to the estimate of iROAS. More generally, end of the cooldown period can be chosen to correspond to the time at which the cumulative effects cease and iROAS stabilizes. Hence, it is important to look for the cumulative effects beyond the intervention period. However, if the cumulative

effects do *not* cease within a reasonable time period after the marketing intervention stops, it may indicate a problem in the configuration of the campaigns, an unexpected event that interfered with the experiment. It could also indicate that the cooldown period needs to be longer.

4 Design Process for TBR

A design process featuring a statistical power analysis is a crucial step in planning of an effective experiment. Not having one will lead to over spending on some experiments (results with more precision than is necessary to make a decision) and underspending on others (results with so much uncertainty that they cannot be used to make a decision).

As the primary interest is to measure the iROAS, it is necessary to predict the precision of an iROAS estimate that a proposed experimental design will deliver. This is directly related to the concept of statistical power, that is, the probability of detecting an effect of a given magnitude or greater under a specified design scenario. Estimates that have higher precision correspond to higher statistical power. We define the precision as the half-width of a 2-sided credible interval ("CI half-width") of the iROAS estimate; this is a more concrete measure of the efficiency of the experiment than that of power. The CI half-width also specifies the smallest effect size that can be detected with a given statistical significance level.

A change in marketing strategy that is small may produce a small change in ad spend and ad delivery, which may produce an effect that is too small to detect. Sometimes the change is limited by budget constraints. In other situations it may be limited by ad inventory.

4.1 Technical implementation

We predict the uncertainty of the estimate in a future experiment using data from historical time series data using a process similar to the one already described in Vaver and Koehler (2011). It applies equally well to GBR and TBR.

1. Assign treatment geos. Each geo is randomly assigned to either the treatment or control group in the manner specified by the experimental design. For example, geos may be stratified by their sizes (in terms of volume of response variable) to ensure a more balanced set of groups. This is accomplished by sorting the geos by their response metric volume across a set of pretest period data. For each adjacent and non-overlapping pair of geos, one is randomly assigned to test, and the other to control. This assignment process could be generalized to balance across additional geo-level characteristics related to sales behavior and ad effectiveness.

2. Fix design parameters. Fix the total experiment length, say N_{exp} days, including the pretest, intervention, and cooldown periods that correspond to the candidate test design. Fix also the total ad spend change, which is assumed to be a known number (i.e., estimated without uncertainty).

3. Generate a distribution of CI half-widths. Generate a sequence of simulated "pseudo-geo experiment data sets," that is, full data sets of length N_{exp} days, constructed from intervals extracted from the available historical time series data. Each of these simulated data sets represents a possible outcome in the future experiment. They are analyzed using the process described in Section 4.2, which among other things, generates a CI half-width for the the iROAS.

To generate the sequence of such data sets, we use the following procedure. Let $t = 1, \ldots, T$ index each available date in the historical time series. We extract a data set D_i, indexed similarly by $i = 1, \ldots, T$, starting from a date $t = i$ onward consisting of N_{exp} days. When the end date of the pseudo-geo experiment extends beyond date T, complement the time series by adding dates starting from $t = 1$, thus "recycling" data points as necessary. For example, assuming daily data, a data set D_{T-1} contains data from days $\{T-1, T, 1, 2, 3, \ldots, N_{exp}-2\}$. Each of the data sets i is assigned a new, consecutive range of dates. Thus we obtain T unique data sets in total.

4. Obtain an estimate. Take the median of these T generated CI half-widths as the estimate of the CI half-width that we expect to obtain in the future experiment.

4.2 Design Considerations

We consider a two-group design, but the ideas described below are directly applicable to more complex designs.

Intensity of ad spend change. By far the most influential parameter is the magnitude of the campaign change in the test group, which corresponds to the magnitude of the ad spend change (Δ_{cost}). Increasing the intensity of ad spend change (i.e., ad spend change per time unit, such as week) by a factor of f, while keeping other design parameters constant, reduces the CI half-width by a factor of $1/f$. Thus doubling the spend will halve the uncertainty. This is straightforward to see from the definition of iROAS: consider its cumulative posterior distribution, $\Pr(\Delta_{resp}/\Delta_{cost} < u|y) = \alpha$ for some fixed quantile u and probability α (y represents all fixed and observed quantities). Then multiplying the cost by a factor of f yields a scaled quantile u/f, since $\Pr(\Delta_{resp}/(f\Delta_{cost}) < u/f|y) = \alpha$. The effect of the ad spend intensity on the precision of iROAS can also be seen in Section 9.2. [1]

Pretest period length. Increasing the length of the pretest period increases the precision of the iROAS, via increasing the precision of the parameters (α, β, and σ). As shown in Section 9.3, the CI half-width of the iROAS estimate decreases at a rate of $1/\sqrt{n}$, where n is the number of data points included in the pretest period. However, it is not possible to reduce the CI half-width without bound.

Test period length. Increasing the length of the test period will improve the precision as long as the intensity of ad spend change does not decrease during the test period. This strategy has more impact for shorter periods than for longer periods. However, Section 9.4 shows that there is a limit to how much the size of the credible interval can be reduced.

Fraction of control geos. Although the number of geos does not affect the TBR estimates of iROAS directly, changing the number of geos in control or treatment group may affect the precision of the iROAS estimate via the change in the volume (magnitude) of x_t and y_t. As discussed in Section 9.6, only the change in the volume of treatment geos will have a direct impact on the precision of iROAS. For example, reducing the treatment group volume by 50% by reducing the number of geos in the treatment group (and keeping the ad spend intensity fixed), we can expect that the uncertainty in the estimate is halved. This means, consequently, that the ad spend requirement to obtain the same CI half-width will be correspondingly lower. However, if the number of geos is

[1] Increasing the ad spend intensity improves precision. However, it also increases the chances of hitting the point of diminishing returns. Therefore the uncertainty will be lower, but with a potentially lower iROAS. Also, the ad spend cannot be increased without bound: at some point the ad spend increase will not be feasible due to limitations in ad inventory.

very low, the model fit will likely be worse.

Number of geos. As already mentioned above, the number of geos does not have a direct impact on the estimation process. However, with only a few geos in either group, there is less randomization in the assignment process (and less protection from observed/unobserved biases). Hence, we can expect a worse fit and a higher residual variance, which affects the precision of the iROAS estimate directly, as can be seen in Section 9.2. A high correlation between the two groups will improve the fit.

4.3 Design Process: Example

We ran a preanalysis using the historical data from the experiment described in Section 3.3 to obtain an estimate for the required ad spend for a specified precision of the iROAS estimate. The lengths of the pretest, intervention and cooldown periods were fixed to 8, 4, and 1 week, respectively. As in the actual experiment, we randomized all 210 geos into two groups of 105 geos each.

The geos were assigned to groups and the power analysis was run to obtain the estimate for the required spend, which was $\$22,000$.[2] This estimate was very precise due to the fact that the variance of the pseudo-experiment geo time series was essentially constant over the simulations. Since this figure was over the budget of $\$20,000$, it was decided that a lower precision was acceptable if the ad spend could be kept under the limit. The expected CI half-width was calculated directly without further simulations by rescaling, yielding $0.5 \cdot \$22,000/\$20,000 = 0.55$.

In the end, the four-week experiment was executed with the total spend of $\$18,273$ that yielded a precision of 0.62, slightly over the target of 0.55.

We can evaluate the accuracy of the predicted estimate as follows. The predicted cost for a unit CI half-width was $\$20,000 \cdot 0.55 = \$11,000$; however the actual cost for a unit CI half-width was $\$18,273 \cdot 0.62 = 11,329$, Hence the difference of the predicted cost and that of the actual precision was only 3%.

5 Performance Evaluation

To understand the performance and the characteristics of the TBR algorithm, we applied TBR to simulated sales time series with varying numbers of geos, noise, and correlational features.

5.1 Simulation Set-up

We generated multiple sets of complete geo time series with different levels of noise and correlations that are similar to data from actual experiments.

In sales data from actual experiments, the geo time series are aggregates of even smaller geographic units. So, it is reasonable to assume that the sum of such units is proportional to their variance, but this only holds when these smaller units are independent. Analysis of dozens of sets of experimental data indicates that response metrics across geos are highly correlated, and the standard deviation of a response metric in a geo tends to be proportional to its mean. A similar observation was noted by Owen and Launay (2016).

[2] All numbers have been scaled to anonymize the actual experiment.

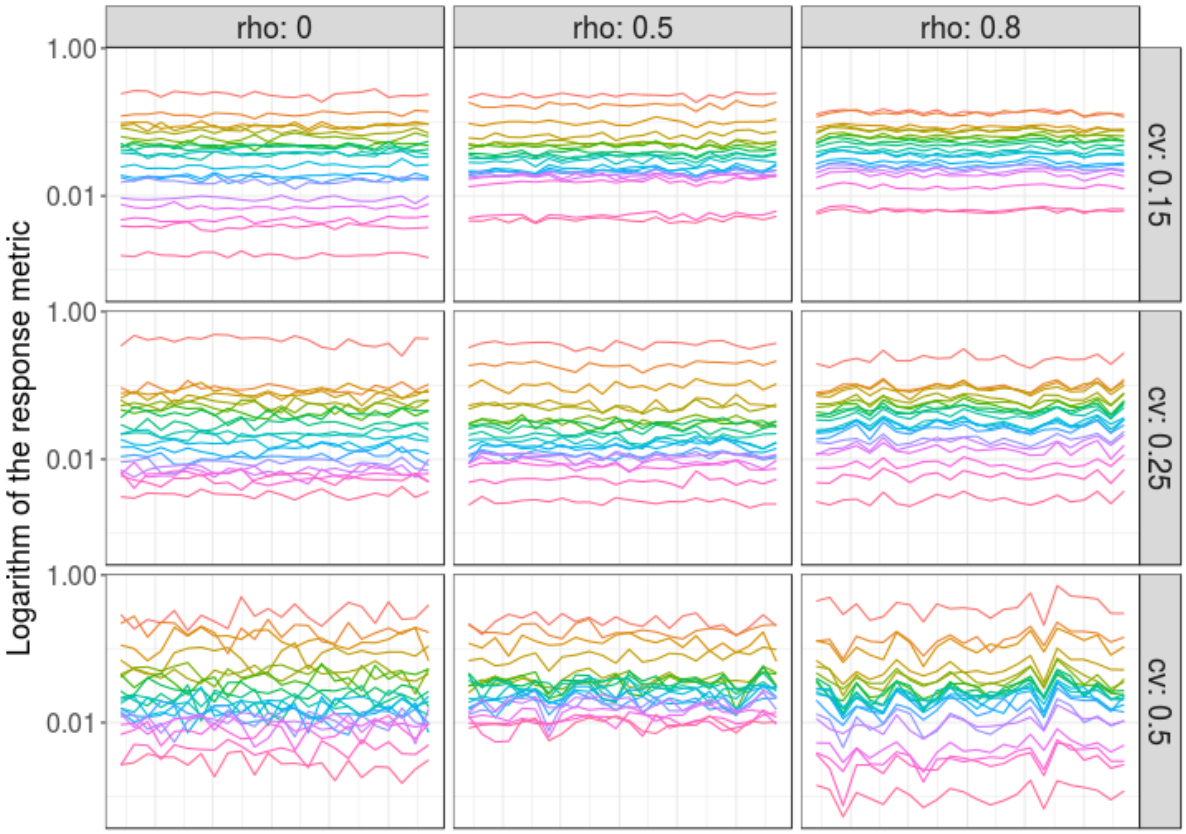


Figure 6: Each plot in this grid shows simulated time series for 20 geos with a different combination of c and ρ. Each time series shows the weekly sales share for each geo. The vertical axis is on a log scale to better indicate the variation in the smallest geos.

It is also not uncommon to observe very high ratios of standard deviation to mean. These ratios may be as high as 50%, indicating high volatility in the response metric.

We generate weekly[3] aggregate geo-level sales time series as follows. For a specified number of geos, we sample from a lognormal distribution to find a baseline volume of the response metric, m_i, for each geo. These baselines are normalized so that $\sum_i m_i = 1$.

The time series for simulated sales has two components: one that contains the seasonal pattern, W_t, that is common to all geos, and one that contains the noise, Z_{it}, for geo i at time t. These variables have a mean of one and standard deviations of c_w and c_z, respectively. They are combined to form the actual time series:

$$y_{it} = m_i(0.5 \cdot W_t + 0.5 \cdot Z_{it}),$$

such that at any time point t, the expected value of y_{it} is m_i.

The correlation between geos i and j is $\rho = \text{Cor}(y_{it}, y_{jt}) = c_w^2/(c_w^2 + c_z^2)$. We define the total coefficient of variation, c to be $c^2 = c_w^2 + c_z^2$. Thus, we have $c_w^2 = \rho c^2$ and $c_z^2 = (1-\rho)c^2$. We generate sets of simulated data using one of three different values for $\rho \in [0, 1)$ and one of three different values for $c \in [0.15, 0.5]$.

Figure 6 shows a sample of simulated data sets of 20 geos each. 2000 such simulated data sets were used for each scenario, i.e., combination of (ρ, c).

[3]Since TBR does not distinguish between daily and weekly data, without loss of generality we assume that the simulated data is generated at a weekly aggregated level.

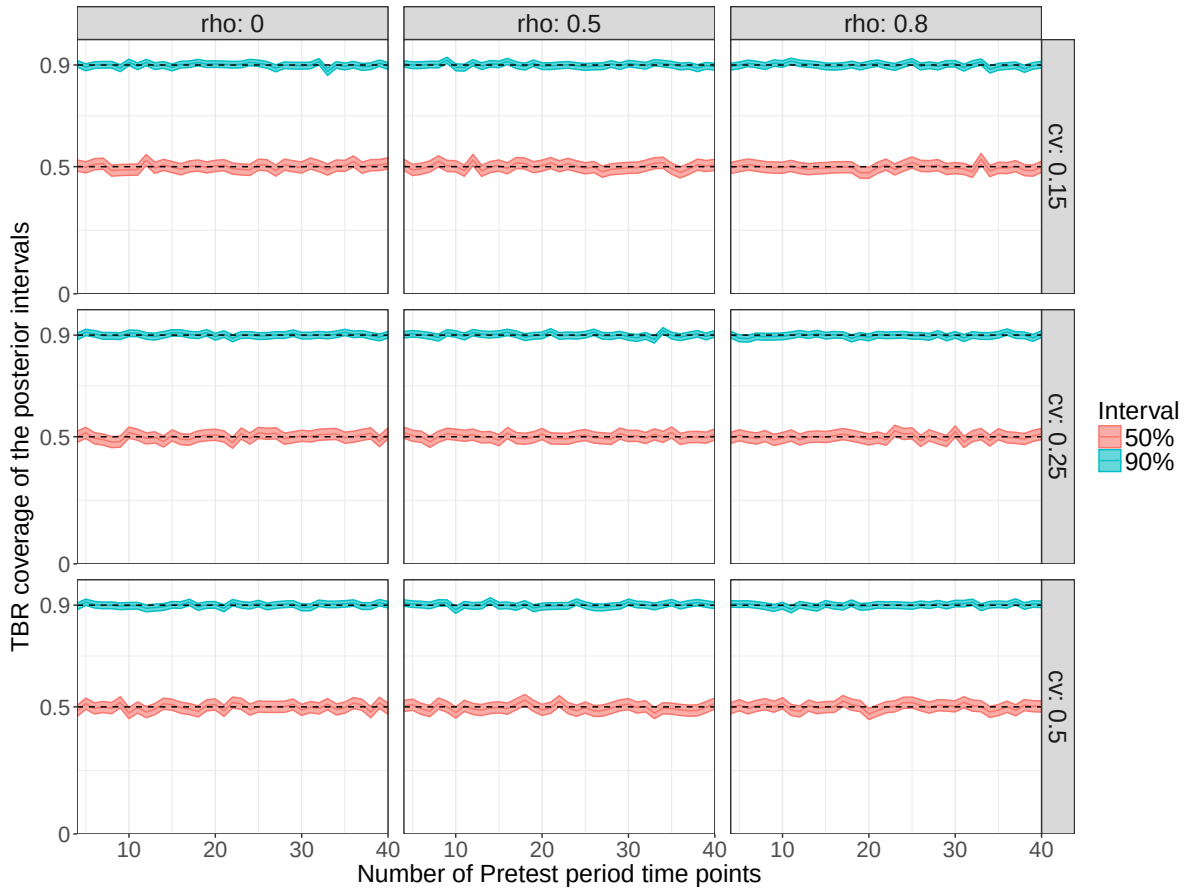


Figure 7: Coverage rates of the 90% and 50% posterior intervals. TBR achieves nominal coverage regardless of the level of noise and correlation, and the length of the pretest period. The bands are 95% posterior intervals derived from a binomial-Beta conjugate model.

For situations in which the advertiser has existing online advertising in the pretest period, the incremental cost is estimated from the cost time series and there is uncertainty associated with this estimate. However, more commonly, the incremental cost of online advertising is estimated with much higher precision than changes in the response metric. Therefore, for simplicity, in these performance evaluations we assume that the incremental cost is a known constant.

5.2 Coverage and bias

Figure 7 shows the estimated coverage rates for the two-sided 90% and 50% posterior intervals based on 2000 simulated data sets of 20 geos each. There were four time points in the test period and no cooldown period. To account for the effect of the sample size, instead of plotting the empirical means, we show the 2.5% and 97.5% posterior quantiles of a Bayesian estimate of the coverage rate. These plots show that TBR achieves nominal coverage regardless of the level of noise, correlation, or of the number of time points in the pretest period (i.e., length of the training period). The mean estimates of the probabilities are shown with posterior intervals obtained from the binomial-Beta conjugate model $\text{Beta}(1/3+y, 1/3+n-y)$, where y is the number of successes out of $n = 2000$ trials. This estimate is based on the neutral prior distribution $\text{Beta}(1/3, 1/3)$ introduced in Kerman

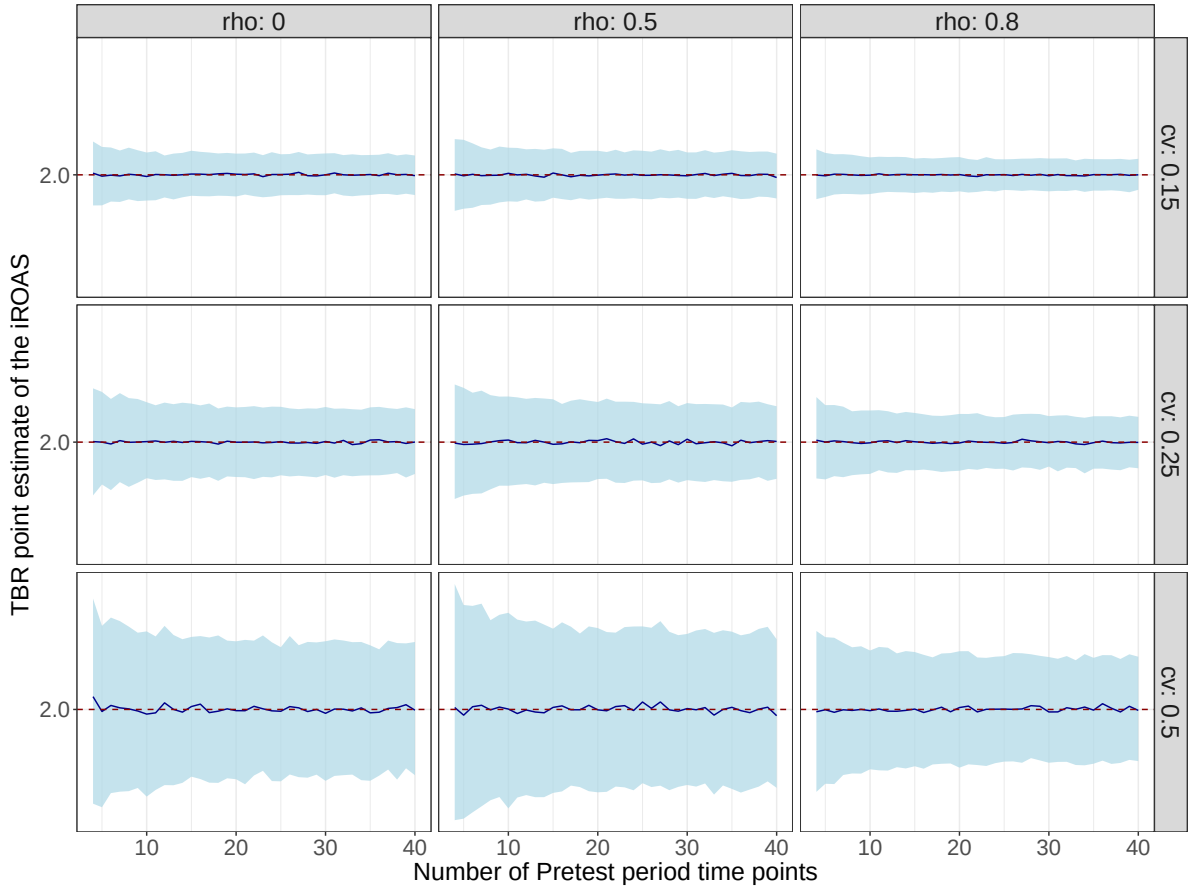


Figure 8: Bias of the point estimate for iROAS. Median point estimate of iROAS generated via TBR for 2000 sets of simulated data (dark blue line). The true value is 2.0. The blue bands show the empirical 50% middle interval of the distribution of iROAS estimates. The median point estimate is practically unbiased in these scenarios.

(2011), which yields the median approximately equal to the empirical estimate.

Figure 8 shows the empirical estimates of bias (mean differences between the estimate and the true iROAS), across different combinations of c and ρ and pretest periods with different lengths. The dark blue line is the median point estimate of iROAS generated via TBR for 2000 simulated data sets in which the true value of iROAS is 2.0. The light blue band indicates the empirical 50% middle interval. The plots demonstrate that the median point estimate, $(\hat{\theta})$, is practically unbiased in these scenarios. For another point of view, we calculated the proportion of (squared) bias of the mean squared error (MSE), that is, $(\mathrm{E}[\hat{\theta} - \theta_{\mathrm{True}}])^2 / \mathrm{E}[(\hat{\theta} - \theta_{\mathrm{True}})^2]$. Over all scenarios, this proportion was 0.04% with standard deviation 0.06%. These observations support that the TBR posterior median performs well as an estimator of the incremental ROAS.

The TBR analyses were performed using the open-source software by Google, Inc. (2017b).

6 Software

We have released an open-source software package (Google, Inc., 2017b) that provides necessary functions and object classes to perform analyses and preanalyses using both

TBR and GBR methods.

7 Summary

We have presented a new methodology, TBR, for estimating causal effects in geo experiments. The underlying model, a simple linear regression, is used to predict the counterfactual time series from the fully observed time series that is represented by the aggregate of all control group geos. The total incremental cost and revenue are estimated separately, and combined to form an estimate of iROAS, the incremental return on ad spend.

The underlying model is straightforward to fit using standard statistical software. We have shown it to perform well under various levels of noise and correlation. TBR is flexible as it is applicable for analyzing experiments with a very low number of geos (two or more), such as those arising from experiments in smaller national markets, and matched market studies.

8 Acknowledgements

We thank Tim Au, Anne-Claire Haury, Nicolas Remy, Fan Zhang, and Jim Koehler for their participation in the research.

9 Appendix

9.1 Posterior distribution of the cumulative causal effect

Joint posterior distribution of (α, β)

Let X be the $n \times 2$ matrix with the first column constant and second column as the time series x_t from the pretest period (see Equation 1). y denotes the corresponding vector with the treatment time series y_t. Following Gelman et al. (2013), we assume a standard noninformative prior distribution, uniform on the joint distribution of $(\alpha, \beta, \log \sigma)$. The conditional posterior distribution of (α, β), given σ, is then Normal with the mean vector $VX'y$ and covariance matrix $\sigma^2 V$, where $V = (X'X)^{-1}$ is the unscaled covariance matrix. With this chosen prior, these quantities are conveniently equivalent to the classical point estimates of (α, β) and their covariance matrix.

To obtain the marginal posterior distribution of (α, β), we integrate over σ to obtain a scaled and shifted bivariate t-distribution for (α, β) with $n - 2$ degrees of freedom, where n denotes the number of time points in the pretest period. This distribution has a location equal to the mean of the conditional posterior $(VX'y)$ and a squared scale matrix $s^2 V$, where s^2 is the classical point estimate of the residual variance. Hence, in the following calculations, we can derive the conditional distribution given σ and replace σ by s.

When characterizing the uncertainty represented by this distribution, we refer to the 'scale' of the distribution rather than the standard deviation, as the scale of the distribution is directly proportional to the width of the posterior intervals. Standard deviation is also undefined for the case of $n \leq 4$. Similarly, we avoid speaking of the 'mean' estimate and instead speak of the median estimate, which exists for all t-distributions.

Median estimate of $\Delta(T)$

Assuming for the moment that σ is known, we derive the form of the cumulative effect, $\Delta(T)$ in terms of the unknowns. Let T be an integer corresponding the Tth time point within the test period (that is, day T or week T). Then,

$$\Delta(T) \equiv \sum_t^T \phi_t = \sum_t^T (y_t - y_t^*) = \sum_t^T (y_t - (\alpha + \beta x_t) + \epsilon_t^*) \,. \tag{2}$$

Taking the expectation,

$$\begin{aligned} \mathrm{E}(\Delta(T)) &= \sum_t y_t - \left(T\mathrm{E}(\alpha) + \left(\sum_t^T x_t \right) \mathrm{E}(\beta) \right) && (3) \\ &= T\left(\bar{y}_T - \mathrm{E}(\alpha) - \bar{x}_T \mathrm{E}(\beta)\right) && (4) \end{aligned}$$

where $(\mathrm{E}(\alpha), \mathrm{E}(\beta))$ are the components of the mean vector $VX'y$, and $\bar{x}_T := \frac{1}{T}\sum_{t=1}^T x_t$, and similarly $\bar{y}_T$ are averages over the test period. Averaging over σ does not change the median estimate of $\Delta(T)$.

Scale of $\Delta(T)$

Still working with the conditional joint distribution of (α, β) given σ, we apply the variance operator to (2),

$$\mathrm{Var}(\Delta(T)|\sigma) = T^2\mathrm{Var}(\alpha) + \mathrm{Var}(\beta)\left(\sum_t^T x_t\right)^2 + 2T\mathrm{Cov}(\alpha,\beta)\left(\sum_t^T x_t\right) + T\sigma^2, \tag{5}$$

where $\mathrm{Var}(\alpha)$, $\mathrm{Var}(\beta)$ and $\mathrm{Cov}(\alpha, \beta)$ are the equal to the corresponding entries in the covariance matrix $\sigma^2 V$.

Integrating over σ yields a scaled and shifted t-distribution with $n-2$ degrees of freedom. The scale of the t-distribution equals the square root of (5), with σ replaced by s, the classical point estimate for the residual standard deviation. Rearranging terms,

$$Ts\left(v_\alpha + 2\bar{x}_T v_{\alpha,\beta} + v_\beta \bar{x}_T^2 + 1/T\right)^{1/2} \tag{6}$$

where v_α, v_β and $v_{\alpha,\beta}$ are the entries in the unscaled covariance matrix $V = \begin{pmatrix} v_\alpha & v_{\alpha,\beta} \\ v_{\alpha,\beta} & v_\beta \end{pmatrix}$.

The half-width of the $100p\%$ two-sided middle posterior interval is therefore (6) multiplied by the $0.5 \cdot (1+p) \cdot 100\%$ percentile of the t-distribution with $n-2$ degrees of freedom.

9.2 Posterior distribution of iROAS

For simplicity, we assume that the total cost of the intervention is estimated without uncertainty, that is, it is a known constant. The cumulative cost is $\Delta_{cost}(T) = \sum_{t=1}^T c_t = \bar{c}T$, where c_t are the daily or weekly increments and $\bar{c}$ is the average cost increment, that is, the ad spend intensity.

The mean estimator of iROAS(T) is then equal to the mean estimator of $\Delta(T)$ divided by $\bar{c}T$:

$$\mathrm{E}(\mathrm{iROAS}(T)) = \frac{\mathrm{E}(\Delta(T))}{\bar{c}T} = \frac{1}{\bar{c}}\left(\bar{y}_T - \mathrm{E}(\alpha) - \bar{x}_T \mathrm{E}(\beta)\right). \tag{7}$$

The scale of the posterior of iROAS is then equal to (6) divided by $\bar{c}T$, yielding,

$$\frac{s}{\bar{c}}\left(v_\alpha + 2\bar{x}_T v_{\alpha,\beta} + v_\beta \bar{x}_T^2 + 1/T\right)^{1/2}. \tag{8}$$

The resulting posterior distribution is again a scaled and shifted t-distribution. In the case of nonconstant cost, the posterior distribution of iROAS(T) is the ratio of two t-distributions, which is a nonstandard distribution; in practice this is estimated by simulation.

9.3 Effect of changing the pretest period length

The unscaled covariance matrix V can be written as,

$$V = \frac{1}{n} \cdot \frac{1}{\mathrm{Var}(x)} \begin{pmatrix} \frac{1}{n}\sum^n x_i^2 & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix} \tag{9}$$

where $\mathrm{Var}(x) := \frac{1}{n}\sum_i^n x_i^2 - \bar{x}_n^2$ and $\bar{x}_n := \frac{1}{n}\sum_i x_i$ (x_i in the pretest period), showing that all entries decrease to 0 as n increases, assuming that $\mathrm{Var}(x)$ approaches a nonzero constant.

Increasing the pretest period length n will increase the precision of the parameters (α, β, σ), making the uncertainty of the parameters go to zero at a rate of $1/\sqrt{n}$. However, the uncertainty of iROAS(T), for a fixed T and $\bar{c}$ will not reach zero:

$$\frac{s}{\bar{c}\sqrt{T}}\left(1 + O(1/n)\right)^{1/2}. \tag{10}$$

Hence the width of the posterior interval decreases at a rate of $1/\sqrt{n}$ but approaches $\sigma_0/(\bar{c}\sqrt{T})$, where σ_0 denotes the true residual standard deviation.

9.4 Effect of changing the intervention period length

Equation 8 shows that if T increases while the average ad spend intensity $\bar{c}$ is kept constant (and s and $\bar{x}_T$ do not change), the only change in the scale of the posterior of iROAS(T) is due to the term $1/T$ inside the square root. Eventually, this scale approaches,

$$\frac{s}{\bar{c}}\left(v_\alpha + 2\bar{x} v_{\alpha,\beta} + v_\beta \bar{x}^2\right)^{1/2}, \tag{11}$$

where $\bar{x}$ is the limit of $\frac{1}{T}\sum_t^T x_t$. This shows that increasing the length of the test period has a limited ability to reduce the uncertainty of the iROAS estimate.

9.5 Effect of changing the cooldown period length

Increasing the length of the cooldown period decreases the ad spend intensity $\bar{c}$, since the overall spend change stays the same while the length of the analysis period increases. As a result, the width of the credible interval grows across the cooldown period. This can be

seen by replacing $\bar{c}$ in Equation (8) with C/T, where C is the fixed, total amount of ad spend change:

$$\frac{sT}{C}\left(v_\alpha + 2\bar{x}_T v_{\alpha,\beta} + v_\beta \bar{x}_T^2 + 1/T\right)^{1/2}. \tag{12}$$

Therefore, for large T, the scale of the posterior of iROAS grows proportional to $\sqrt{T}$.

9.6 Effect of changing control or treatment group volume

Change in control group volume. Suppose that x_t is multiplied by a factor of κ_x, a fixed positive quantity, due to omitting or adding geos to the control group. From Equation (9) we see that v_α stays unchanged, v_β is scaled by $1/\kappa_x^2$, and $v_{\alpha,\beta}$ is scaled by $1/\kappa_x$. However, $\bar{x}_T$ is scaled by a factor of κ_x, hence none of the terms inside the square root of Equation 8 changes. Residuals do not change, therefore neither does residual standard deviation, s. So, the scale of the uncertainty of the iROAS estimate (8) does not change in this situation.

Change in treatment group volume. Multiplying y_t by a factor of, say κ_y, will change the residual standard deviation s by a factor of κ_y. The terms under the square root in Equation (8) stay unchanged as they are all functions of x_t only. Hence the scale of the uncertainty of the iROAS estimate is changed by a factor of κ_y. Although increasing the size of the treatment group helps to lower uncertainty, it also (proportionally) increases the cost of the experiment.

9.7 TBR-OR: an alternative model using orthogonal regression

We have shown that the model performs well even under high noise and correlation across geos. However, if there is strong seasonality across the analysis period, the model structure changes. One approach for dealing with this change is to use orthogonal regression, also referred to as Deming regression (see e.g., Fuller, 1987). The error-in-variables model, which we call TBR-OR, attempts to find the best fit to $y_t = \alpha + \beta x_t$, but the covariate x_t is modeled as $x_t = x_t' + \eta_t$, where η_t are independent error terms with an unknown variance σ_x^2.

For the purposes of model fitting, the ratio of the variance of y_t (say, σ_y^2) and that of x_t (σ_x^2) is assumed to be known. Estimating both variance parameters simultaneously is not possible, since this would cause the model to be over-parameterized. We estimate this ratio using the empirical point estimate of the ratio $\mathrm{Var}(y)/\mathrm{Var}(x)$, where y_t and x_t are in the pretest period.

This model attempts to produce the best fit whether we regress y on x *or* vice versa. At first, this sounds satisfying due to the symmetry in x and y imposed by the geo assignment process. However, if there is little or no seasonality in the data, TBR-OR will produce extremely unstable estimates, resulting in unreliable predictions. This can be seen in the estimator of β (Fuller, 1987, Equation 1.3.7), which has the point estimate of the correlation between y and x in the denominator. It follows that TBR-OR should not be used at all when the correlation between y and x is near zero. This is much more likely to be true when there is no strong seasonality. Having low correlation with strong seasonality requires very large variability in the response metric; so large that the variability dominates the seasonality. In this case, the natural estimate for a counterfactual when x_t is observed to have near constant value would also simply be a flat time series,

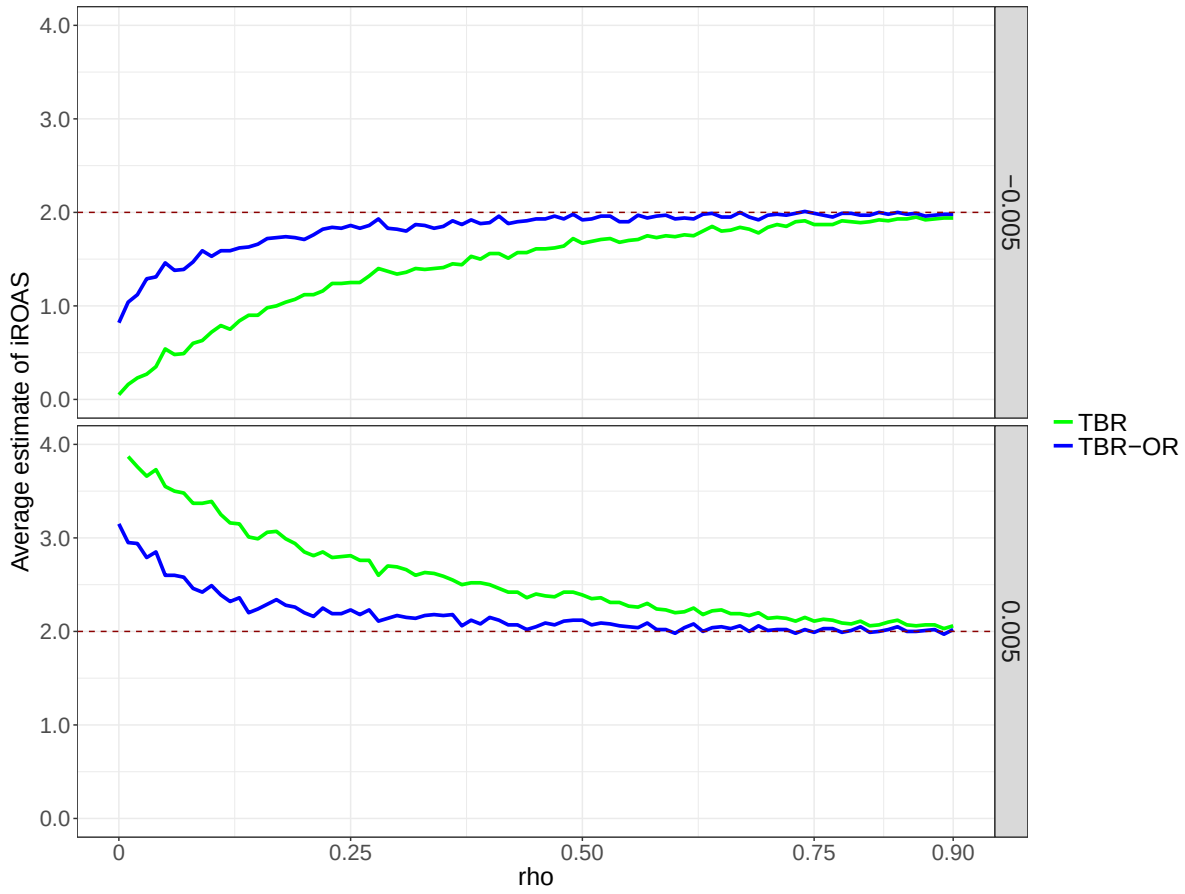


Figure 9: Average iROAS point estimates under exponential baseline growth for both TBR and TBR-OR. The true iROAS is 2.0.

as would be predicted by TBR, but TBR-OR is unable to produce a stable estimate. In the case when correlation is high, TBR may not be the most appropriate model and can introduce some bias, although the amount of bias is smaller with higher correlation.

Comparing TBR with TBR-OR

We implemented TBR-OR as a frequentist model, and used bootstrap to generate confidence intervals. To demonstrate the performance of TBR-OR, we simulated geo time series with an exponential baseline (i.e., non-ad related) growth in the response metric. 20 geos were randomly assigned to treatment and control groups, and we used a pretest period with 20 time points and an intervention period with four time points.

Figure 9 shows the iROAS estimate under the scenario in which the baseline response metric is changing at a rate of +0.5% per week. High, and sustained, growth is one way the baseline could change very significantly. The more likely scenario of strong change would be holiday-related, which operates on an "experiment" type time scale. The true iROAS is 2.0. TBR-OR performs better than TBR under this scenario. Although when the true underlying correlation ρ is low (say, less than 0.5), there is also bias in the TBR-OR estimates of iROAS.

Figure 10 shows the interval coverage rates for the same set of scenarios as in Figure 6. At 90%, TBR provides relatively consistent coverage, although somewhat low at higher correlations. Whereas, the corresponding TBR coverage is low for lower correlations

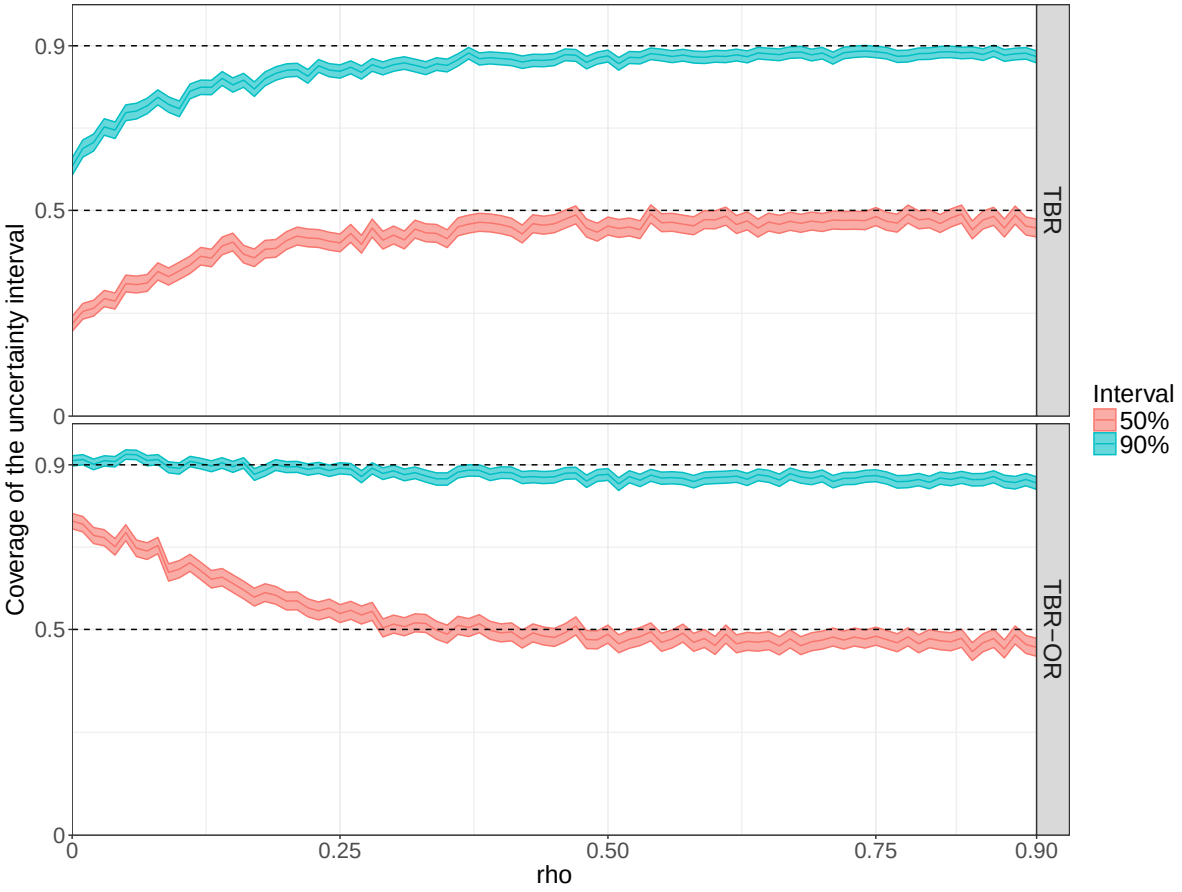


Figure 10: Interval coverage rates under exponential baseline growth for both TBR and TBR-OR.

(< 0.50). At 50%, TBR and TBR-OR coverage are somewhat low at higher correlations, and too low and high at lower correlations, respectively.

However, even when correlation increases, the TBR-OR bootstrap intervals seem to be too short to produce nominal coverage. TBR, on the other hand, performs surprisingly similarly although the intervals are too short throughout, albeit perform at least as well as those of TBR-OR for $\rho \geq 0.5$. Although not shown here, the plots are similar when growth rates are negative.

In conclusion, when the experiment spans a period of time that includes strong sustained seasonal trends (e.g. the holiday season or a sales event), TBR-OR may be a better choice for the analysis. Still, TBR is our default choice for analysis over TBR-OR.

References

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274. `http://research.google.com/pubs/pub41854.html`.

Chan, D., Ge, R., Gershony, O., Hesterberg, T., and Lambert, D. (2010). Evaluating online ad campaigns in a pipeline: Causal models at scale. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,

KDD '10, pages 7–16, New York, NY, USA. ACM. https://research.google.com/pubs/archive/36552.pdf.

Chan, D. X., Yuan, Y., Koehler, J., and Kumar, D. (2011). Incremental clicks: The impact of search advertising. *Journal of Advertising Research*, 51, no. 4:643–647. https://research.google.com/pubs/archive/37161.pdf.

Fuller, W. A. (1987). *Measurement error models*. Wiley, 1st edition.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CR, London, 3 edition.

Google, Inc. (2017a). AdWords API: Geographical targeting. https://developers.google.com/adwords/api/docs/appendix/geotargeting.

Google, Inc. (2017b). R package GeoexperimentsResearch. https://github.com/google/GeoexperimentsResearch/.

Gordon, B., Zettelmeyer, F., Bargava, N., and Chapsky, D. (2016). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Technical report, Facebook, Inc.

Johnson, G. A., Lewis, R. A., and Nubbemeyer, E. I. (2015). Ghost ads: A revolution in measuring ad effectiveness. *SSRN*. http://dx.doi.org/10.2139/ssrn.2620078.

Kerman, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electron. J. Statist*, 5:1450–1470. http://dx.doi.org/10.1214/11-EJS648.

Lewis, R. A. and Rao, J. M. (2014). The unfavorable economics of measuring the returns to advertising. *Available at SSRN*. http://dx.doi.org/10.2139/ssrn.2367103.

Owen, A. and Launay, T. (2016). Multibrand geographic experiments. Technical report, Google Inc. https://arxiv.org/abs/1612.00503.

The Nielsen Company (2017). DMA regions. http://www.nielsen.com/intl-campaigns/us/dma-maps.html.

Vaver, J. and Koehler, J. (2011). Measuring ad effectiveness using geo experiments. Technical report, Google Inc. http://research.google.com/pubs/pub38355.html.

Vaver, J. and Koehler, J. (2012). Periodic measurement of advertising effectiveness using multiple-test-period geo experiments. Technical report, Google Inc. http://research.google.com/pubs/pub38356.html.

Ye, Q., Malik, S., Chen, J., and Zhu, H. (2016). The seasonality of paid search effectiveness from a long running field test. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, pages 515–530, New York, NY, USA. ACM. http://doi.acm.org/10.1145/2940716.2940717.