Measuring Ad Effectiveness Using Geo Experiments

Jon Vaver, Jim Koehler

Google Inc.

Abstract

Advertisers have a fundamental need to quantify the effectiveness of their advertising. For search ad spend, this information provides a basis for formulating strategies related to bidding, budgeting, and campaign design. One approach that Google has successfully employed to measure advertising effectiveness is geo experiments. In these experiments, non-overlapping geographic regions are randomly assigned to a control or treatment condition, and each region realizes its assigned condition through the use of geo-targeted advertising. This paper describes the application of geo experiments and demonstrates that they are conceptually simple, have a systematic and effective design process, and provide results that are easy to interpret.

1 Introduction

Every year, advertisers spend billions of dollars on online advertising to influence consumer behavior. One of the benefits of online advertising is access to a variety of metrics that quantify related consumer behavior, such as paid clicks, website visits, and various forms of conversions. However, these metrics do not indicate the incremental impact of the advertising. That is, they do not indicate how the consumer would have behaved in the absence of the advertising. In order to understand the effectiveness of advertising, it is necessary to measure the behavioral changes that are directly attributable to the ads.

A variety of experimental and observation methods have been developed to quantify advertising's incremental impact (see [1], [5], [3], [2]). Each method has its own set of advantages and disadvantages.

Observational methods of measurement impose the least amount of disruption on an advertiser's ongoing campaigns. In an *observational study* ad effectiveness is assessed by observing consumer behavior in the presence of the advertising over a period of time. The analyses associated with these studies tend to be complex, and their results may be viewed with more skepticism, because there is no control group. That is, a statistical model is used to infer the behavior of a comparable set of consumers without ad exposure, as opposed to directly observing their behavior via an unexposed control group. At Google, observational methods have been used to measure the ad effectiveness of display advertising in the Google Content Network [1] and Google Search [2].

The most rigorous method of measurement is a randomized experiment. One application of randomized experiments that is used to analyze search ad effectiveness is a *traffic experiment*. At Google, these are performed using the AdWords Campaign Experiments (ACE) tool [3]. In these experiments, each incoming search is assigned to a control or treatment condition and the subsequent user behavior associated with each condition is compared to determine the incremental impact of the advertising. These experiments are very effective at providing an understanding of consumer behavior at the query level. However, they do not account for changes in user behavior that occur further downstream from the search.

For example, conversion level behavior may involve multiple searches and multiple opportunities for ad exposures, and a traffic experiment does not follow individual users to track their initial control/treatment assignment or observe their longer-term behavior.

An alternative approach is to vary the control/treatment condition at the cookie level. In a *cookie experiment*, each cookie belongs to the same control/treatment group across time. However, ad serving consistency is still a concern with cookie experiments because some users may have multiple cookies due to cookie churn and their use of multiple devices to perform online research. Cookie experiments have been used at Google to measure display ad effectiveness [5].

This paper describes one additional method for measuring ad effectiveness; the *geo experiment*. In these experiments, a region (e.g. country) is partitioned into a set of geographic areas, which we call "geos". These geos are randomly assigned to either a treatment or control condition and geo-targeting is used to serve ads accordingly. A linear model is used to estimate the return on ad spend.

2 Geo Experiment Description

Online advertising can impact a variety of consumer behaviors. In this paper, we refer to the behavior of interest as the response metric. The response metric might be, for example, clicks (paid as well as organic), online or offline sales, website visits, newsletter sign-ups, or software downloads. The results of an experiment come in the form of return on ad spend (ROAS), which is the incremental impact that the ad spend had on the response metric. For example, the ROAS for sales indicates the incremental revenue generated per dollar of ad spend. This metric indicates the revenue that would not have been realized without the ad spend.

A geo experiment begins with the identification of a set of geos, or geographic areas, that partition a region of interest. For a national advertiser, this region may be an entire country. There are two primary requirements for these geos. First, it must be possible to serve ads according to a geographically based control/treatment prescription with reasonable accuracy. Second, it must be possible to track the ad spend and the response metric at the geo level. Ad serving inconsistency is a concern due to finite ad serving accuracy, as well as the possibility that consumers will travel across geo boundaries. The location and size of the geos can be used to mitigate these issues. It is not generally feasible to use geos as small as, for example, postal codes. The generation of geos for geo experiments is beyond the scope of this paper. In the United States, one possible set of geos is the 210 DMAs (Designated Market Areas) defined by Nielson Media, which is broadly used as a geo-targeting unit by many advertising platforms.

The next step is to randomly assign each geo to a control or treatment condition. Randomization is an important component of a successful experiment as it guards against potential hidden biases. That is, there could be fundamental, yet unknown, differences between the geos and how they respond to the treatment. Randomization ensures that these potential differences are equally distributed - statistically speaking - across the treatment and control groups. It also may be helpful to constrain this random assignment in order to better balance the control and treatment geos across one or more characteristics or demographic variables. For example, we have found that grouping the geos by size prior to assignment can reduce the confidence interval of the ROAS measurement by 10%, or more.

Each experiment contains two distinct time periods: pretest and test (see Figure 1). During the pretest period there are no differences in campaign structure across geos (e.g. bidding strategy, keyword set, ad creatives, etc.). In this time period, all geos operate at the same baseline level and the incremental differences between the treatment and control geos in the ad spend and response metric are zero.

During the test period the campaigns for the

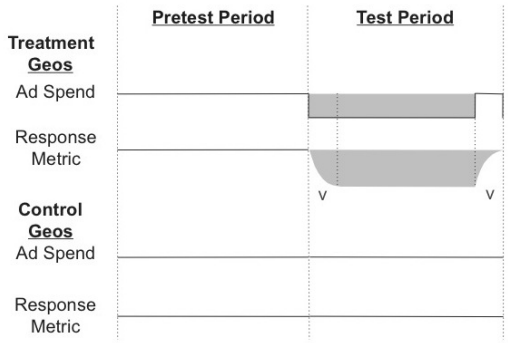

Figure 1: Diagram of a geo experiment. Ad spend is modified in one set of geos during the test period, while it remains unchanged in another. There may be some delay before the corresponding change in a response metric is fully realized

treatment geos are modified. This modification generates a nonzero differential in the ad spend in the treatment geos relative to the control geos. That is, the ad spend differs from what it would have been if the campaign had not been modified. This differential will be negative if the campaign change causes the ad spend to decrease in the treatment geos (e.g. campaigns turned off), and positive if the change causes an increase in ad spend (e.g. bids increased or keywords added). This ad spend differential will generate a corresponding differential in the response metric, perhaps with some time delay, ν. Offline sales is an example of a response metric that is likely to have a positive value of ν. It takes time for consumers to complete their research, make a decision, and then visit a store to make their purchase. The test period extends beyond the end of the ad spend change by ν to fully capture these incremental sales.

3 Linear Model

After an experiment is executed, the results are analyzed using the following linear model:

$$y_{i,1} = \beta_0 + \beta_1 y_{i,0} + \beta_2 \delta_i + \epsilon_i \qquad (1)$$

where $y_{i,1}$ is the aggregate of the response metric during the test period for geo i, $y_{i,0}$ is the aggregate of the response metric during the pretest period for geo i, δ_i is the difference between the actual ad spend in geo i and the ad spend that would have occurred without the experiment, and ϵ_i is the error term. This model is fit using weights $w_i = 1/y_{i,0}$ in order to control for heteroscedasticity caused by the differences in geo size.

The first two parameters in the model, β_0 and β_1, are used to account for seasonal differences in the response metric across the pretest and test periods. The parameter of primary interest is β_2, which is the return on ad spend (ROAS) of the response metric.

The values of $y_{i,1}$ and $y_{i,0}$ (e.g. offline sales) are generated by the advertiser's reporting system. The geo level ad spend is available through AdWords. If there is no ad spend during the pretest period then the ad spend differential, δ_i, required by Equation 1 is simply the ad spend during the test period. However, if the ad spend is positive during the pretest period and is either increased or decreased, as depicted in Figure 1 , then the ad spend differential is found by fitting a second linear model:

$$s_{i,1} = \gamma_0 + \gamma_1 s_{i,0} + \mu_i \qquad (2)$$

Here, $s_{i,1}$ is the ad spend in geo i during the test period, $s_{i,0}$ is the ad spend in geo i during the pretest period, and μ_i is the error term. This model is fit with weights $w_i = 1/s_{i,0}$ using only the control geos (C).

This ad spend model characterizes the impact of seasonality on ad spend from the pretest period to the test period, and it is used as a counterfactual [1] to calculate the ad spend differential. The ad spend differential in the control and treatment geos (T) is found using the following prescription:

$$\delta_i = \begin{cases} s_{i,1} - (\gamma_0 + \gamma_1 s_{i,0}) & \text{for i} \in T \\ 0 & \text{for i} \in C \end{cases} \qquad (3)$$

The zero ad spend differential in the control geos reflects the fact that these geos continue to operate at the baseline level during the test period.

[1] The *counterfactual* is the ad spend that would have occurred in the absence of the treatment.

4 Example Results

One issue that is of primary concern to advertisers is the potential cannibalization of cost-free organic clicks by paid search clicks (i.e. users will click on a paid search link when they would have clicked on an organic search link). Although perhaps unlikely, it is also possible that the co-occurrence of a paid link and an organic link will make an organic click more likely. *Cost per click* (CPC) does not provide the advertiser with a complete picture of advertising impact because of competing effects such as these. A more useful metric is the *cost per incremental click* (CPIC), which can be measured with a geo experiment.

One of Google's advertisers ran an experiment to measure the effectiveness of their search advertising campaign. During this experiment, which lasted several weeks, the advertiser's search ads were shown in half of the geos. Figure 2 shows the result of fitting the linear model in Equation 1 with successively longer sets of test period data to find the ROAS for clicks. At first, the confidence interval of this metric is large, but it decreases quickly as more test period data are accumulated. Each dollar of ad spend generates 1/3 of an incremental click or, equivalently, the CPIC is \$3. In this case, the reported CPC in AdWords is \$2.40, which underestimates CPIC by 20% [2]. So, the paid clicks do displace some organic clicks, but certainly not the bulk of them.

To further illustrate the ability of paid search advertising to generate incremental clicks, Figure 3 shows the cumulative incremental ad spend across the test period along with the cumulative incremental clicks. The number of incremental clicks is zero at the beginning of the test period and increases steadily with time along with the incremental ad spend. However, once the ad spend in the test geos returns to a pretest level, the accumulation of incremental ad spend stops. At the same time, the accumulation of incremental clicks stops as well. This behavior indicates

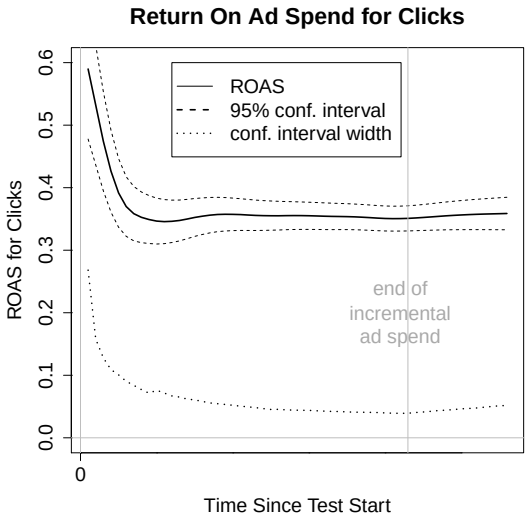

Figure 2: Measurement of return on ad spend for clicks as a function of test period length. The uncertainty in this estimate decreases until the ad spend returns to normal levels in all of the geos.

that, in this case, search advertising does not increase the number of clicks beyond the day in which the ad spend occurred.

As mentioned in Section 2, the impact of ad spend is not as time-limited for all response metrics. Figure 4 is analogous to Figure 3, except the response metric is offline sales. Even after the ad spend differential returns to normal, the impact of the ad spend continues to generate incremental sales for some period of time before fading.

5 Design

Design is a crucial aspect of running an effective geo experiment. Before beginning a test, it is helpful to understand how characteristics such as experiment length, test fraction, and magnitude of ad spend differential will impact the uncertainty of the ROAS measurement. This understanding allows for the design of an effective and efficient experiment. Fortunately, it is possible to make such assessments for the linear model in Equation 1.

[2] In [2] the authors define IAC (incremental ad clicks) as the fraction of paid clicks that are incremental. IAC = CPC / CPIC, so IAC = 80% in this example.

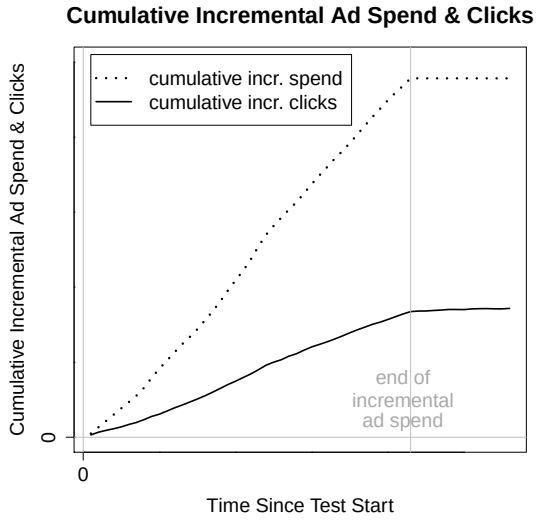

Figure 3: Cumulative incremental ad spend and clicks across the test period. The accumulation of incremental clicks stops as soon as the ad spend returns to the pretest level in all geos.

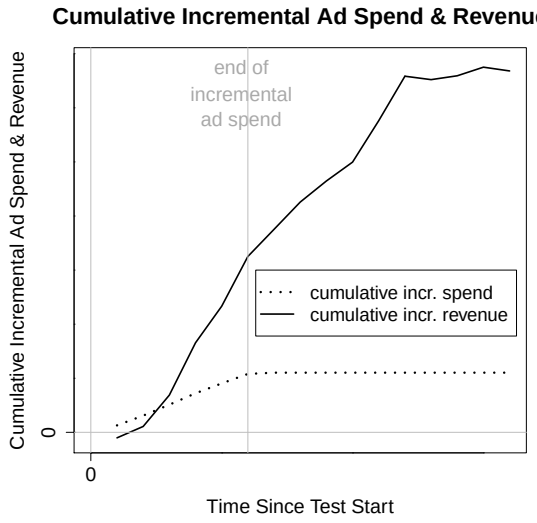

Figure 4: Cumulative incremental sales across the length of the test period. Incremental sales continue to be generated even after the ad spend returns to pretest levels in all geos.

For an experiment with N geos, let $\bar{y}_0 = (1/N)\sum_{i=1}^{N} y_{i,0}$ and $\bar{\delta} = (1/N)\sum_{i=1}^{N} \delta_i$. Linear theory indicates that the variance of β_2 from Equation 1 is

$$\text{var}(\beta_2) = \frac{\sigma_\epsilon^2}{\left(1-\rho_{y\delta}^2\right)\left[\sum_{i=1}^{N} w_i(\delta_i-\bar{\delta})^2\right]} \quad (4)$$

where σ_ϵ is the residual variance, and

$$\rho_{y\delta}^2 = \frac{\left[\sum_{i=1}^{N} w_i(y_{i,0}-\bar{y}_0)(\delta_i-\bar{\delta})\right]^2}{\sum_{i=1}^{N} w_i(y_{i,0}-\bar{y}_0)^2 \sum_{i=1}^{N} w_i(\delta_i-\bar{\delta})^2} \quad (5)$$

(see Appendix) . Using a set of geo-level pretest data in the response variable, it is possible to use this expression to estimate the width of the ROAS confidence interval for a specified design scenario.

The first step in the process is to select a consecutive set of days from the pretest data to create pseudo pretest and test periods. The lengths of the pseudo pretest and test periods should match the lengths of the corresponding periods in the hypothesized experiment. For example, an experiment with a 14 day pretest period and a 14 day test period should have pseudo pretest and test periods that are each 14 days long. The data from the pseudo pretest period are used to estimate $y_{i,0}$ and w_i in Equation 4.

The next step is to randomly assign each geo to the treatment or control group. We have found that confidence interval estimates are lower by about 10% when this random assignment is constrained in the following manner. The geos are ranked according to $y_{i,0}$. Then, this ranked list of geos is partitioned into groups of size M, where the test fraction is $1/M$. One geo from each group is randomly selected for assignment to the treatment group.

It may be possible to directly estimate the value of δ_i at the geo level. For example, if the ad spend will be turned off in the treatment geos, then δ_i is just the average daily ad spend for treatment geo i times the number of days in the experiment. Otherwise, an aggregate ad spend differential Δ can be hypothesized and the geo-

level ad spend differential can be estimated using

$$\delta_i = \begin{cases} \Delta(y_{i,0}/\sum_i y_{i,0}) & \text{for i} \in T \\ 0 & \text{for i} \in C \end{cases} \quad (6)$$

The last value to estimate in Equation 4 is σ_ϵ. This estimate is generated by considering the reduced linear model;

$$y_{i,1} = \hat{\beta}_0 + \hat{\beta}_1 y_{i,0} + \hat{\epsilon} \quad (7)$$

This model has the same form as Equation 1 except the ad spend differential term has been dropped. Fitting this model using the pseudo pretest and test period data results in a residual variance of $\sigma_{\hat{\epsilon}}$, which is used to approximate σ_ϵ.

To avoid any peculiarities associated with a particular random assignment, Equation 4 is evaluated for many random control/treatment assignments. In addition, different partitions of the pretest data are used to create the pseudo pretest and test periods by circularly shifting the data in time by a randomly selected offset. The half width estimate for the ROAS confidence interval is $2\sqrt{\overline{\text{var}(\beta_2)}}$, where $\overline{\text{var}(\beta_2)}$ is the average variance of β_2 across all of the random assignments. This process can be repeated across a number of different scenarios to evaluate and compare designs. Note that if a limited set of pretest data is available, circular shifting of the data makes it possible to analyze scenarios with extended test periods. However, doing so requires data points to be used multiple times in generating each estimate of $\text{var}(\beta_2)$, and the example below demonstrates that this reuse of the data leads to estimates that are overly optimistic.

Figure 5 shows the confidence interval prediction as a function of experiment length for the click example from Section 2. The dashed line corresponds to the predicted confidence interval half width and the solid line corresponds to results from the experiment. For this comparison, the ad spend differential from the experiment was used as input to the prediction. The predictions are quite accurate beyond the very beginning of the test period. Additionally, they maintain this accuracy until the combined

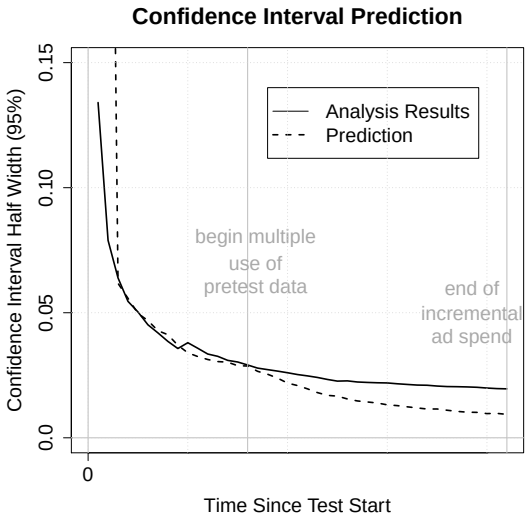

Figure 5: ROAS confidence interval prediction across the length of the test period. The prediction is quite good until the test period becomes long enough that some of the pretest data must be used multiple times to generate each estimate of $\text{var}(\beta_2)$.

length of the hypothesized pretest and test periods becomes longer than the (deliberately) limited set of pretest data used to generate the estimates. The good match between these two curves demonstrates that the absolute size of the confidence interval can be predicted quite well, at least as long as the ad spend differential can be accurately predicted.

6 Concluding Remarks

Measuring ad effectiveness is a challenging problem. Currently, there is no single methodology that works well in all situations. However, geo experiments are worthy of consideration in many situations because they provide the rigor of a randomized experiment, they are easy to understand, they provide results that are easy to interpret, and they have a systematic and effective design process. Geo experiments can be applied to measure a variety of user behavior and can be used with any advertising medium that allows for geo-targeted advertising, Furthermore, these experiments do not require the tracking of

individual user behavior over time and therefore avoid privacy concerns that may be associated with alternative approaches.

Acknowledgments

We thank those who reviewed this paper (with special thanks to Tony Fagan and Lizzy Van Alstine for their many helpful suggestions), others at Google who made this work possible, and the forward looking advertisers who shared their data with us.

References

[1] D. Chan, et al. "Evaluating Online Ad Campaigns in a Pipeline: Causal Models at Scale." *Proceedings of ACM SIGKDD 2010*, pp. 7-15.

[2] D. Chan et al. "Incremental Clicks Impact Of Search Advertising." *research.google.com/pubs/archive/37161.pdf*, 2011.

[3] Google Ads Team. "AdWords Campain Experiments." Sept. 1, 2011. Ad Innovations. http://www.google.com/ads/innovations/ace.html

[4] M. H. Kutner, et al. Applied Linear Statistical Models. New York: McGraw-Hill/Irwin, 2005.

[5] T. Yildiz, et al. "Measuring and Optimizing Display Advertising Impact Through Experiments" In preparation (research.google.com)

7 Appendix

To derive Equation 4, consider the centered versions of the variables $y_{i,1}$, $y_{i,0}$, and δ_i from Equation 1; $y'_{i,1} = y_{i,1} - \bar{y}_1$, $y'_{i,0} = y_{i,1} - \bar{y}_0$, and $\delta'_i = \delta_i - \bar{\delta}$ for $i \in 1...N$ and $\bar{y}_j = (1/N)\sum_i y_{i,j}$. With these translations, the relevant linear model becomes

$$y'_{i,1} = \beta_1 y'_{i,0} + \beta_2 \delta'_i + \epsilon_i \tag{8}$$

Or,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{9}$$

where

$$\mathbf{Y} = \begin{bmatrix} y'_{1,1} \\ . \\ . \\ . \\ y'_{N,1} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} y'_{1,0} & \delta'_1 \\ . & . \\ . & . \\ . & . \\ y'_{N,0} & \delta'_N \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_N \end{bmatrix}$$

With the model in this form, the variance-covariance matrix of the weighted least squares estimated regression coefficients is:

$$var(\boldsymbol{\beta}) = \sigma^2_\epsilon (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \tag{10}$$

(see [4]), where $\mathbf{W}$ is a diagonal matrix containing the weights w_i,

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & . & . & 0 \\ 0 & w_2 & & & . \\ . & & . & & . \\ . & & & . & 0 \\ 0 & . & . & . & w_N \end{bmatrix}. \tag{11}$$

Now,

$$var(\boldsymbol{\beta}) = \sigma^2_\epsilon \begin{bmatrix} \sum_i w_i {y'_{i,0}}^2 & \sum_i w_i y'_{i,0} \delta'_1 \\ \sum_i w_i y'_{i,0} \delta'_i & \sum_i w_i {\delta'_i}^2 \end{bmatrix}^{-1} \tag{12}$$

and the last component of this matrix is the variance of β_2,

$$var(\beta_2) = \frac{\sigma^2_\epsilon \sum_i w_i {y'_{i,0}}^2}{\left(\sum_i w_i {y'_{i,0}}^2\right)\left(\sum_i w_i {\delta'_i}^2\right) - \left(\sum_i w_i y'_{i,0} \delta'_i\right)^2}. \tag{13}$$

Using Equation 5,

$$\left(\sum_i w_i {y'_{i,0}}^2\right)\left(\sum_i w_i {\delta'_i}^2\right)(1 - \rho_{y\delta})^2 =$$

$$\left(\sum_i w_i {y'_{i,0}}^2\right)\left(\sum_i w_i \delta_i'^2\right) - \left(\sum_i w_i y'_{i,0}\delta'_i\right)^2 \tag{14}$$

which, after substituting into Equation 13, leads to Equation 4.