A Time-Based Regression Matched Markets Approach for Designing Geo Experiments

Timothy C. Au
Google LLC
timau@google.com

Abstract

Although randomized controlled trials are regarded as the "gold standard" for causal inference, advertisers have been hesitant to embrace them as their primary method of experimental design and analysis due to numerous technical difficulties that can arise when implementing them in the online advertising context. To help mitigate some of these challenges while still providing the rigor of a randomized controlled trial, Vaver and Koehler (2011) introduced the concept of a "geo experiment." However, it may not always be possible to rely on randomization when designing a geo experiment. For example, it may not be realistic to expect randomization to create balanced experimental groups when some of the geos are markedly different from all of the others, or when there are only a few geos available for experimentation. In addition, randomization may not always be feasible given some of the specific requirements that advertisers often must impose on their experiments in practice—such as the need to run a smaller scale geo experiment within a given budget, or the need to include certain geos in specific experimental groups. Consequently, advertisers may sometimes prefer to forgo some of the benefits of randomization, and in this paper we introduce a more systematic "matched markets" approach that, subject to the advertiser's constraints, greedily searches for experimental group assignments that appear to satisfy some of the critical assumptions of the "Time-Based Regression" (TBR) model for analyzing geo experiments that was introduced in Kerman et al. (2017). If the modeling assumptions of TBR do indeed hold, then the experimental designs that are recommended by our matched markets approach lead to straightforward causal estimates of the geo experiments that are run.

1 Introduction

To evaluate and optimize their marketing strategies, advertisers need to be able to accurately measure the effectiveness of their online marketing campaigns. Unfortunately, such measurement efforts have proven to be very challenging.

Although they are widely employed in the online advertising industry, observational studies have been particularly difficult to analyze despite many recent and significant improvements in observational methods—see, for instance, Imbens and Rubin (2015). In particular, Lewis and Rao (2015) showed that observational studies in the online advertising context are particularly susceptible to selection bias issues due to its inherently targeted nature. Furthermore, because they lack a proper control group, Gordon et al. (2017) empirically demonstrated that observational studies are likely to yield biased estimates and, as a result, and are oftentimes insufficient for measuring the causal effect of online advertising when compared to randomized controlled trials.

Indeed, randomized controlled trials are generally regarded as the "gold standard" for causal inference since they minimize selection bias by randomly assigning the experimental units to different experimental groups. Advertisers, however, have been hesitant to embrace randomized controlled trials as the primary method of designing and analyzing their online marketing experiments. Perhaps the biggest reason for this slow adoption are the technical difficulties that arise when implementing randomized controlled trials in the online advertising context—for example, issues such as cookie churn and multiple device usage can oftentimes make it challenging to design a randomized controlled trial that is capable of maintaining the integrity of the randomization since a non-trivial number of individuals in the control group may inadvertently be exposed to the treatment condition (Gordon et al., 2017).

To help mitigate some of these difficulties while still providing the rigor of a randomized controlled trial in the online advertising context, Vaver and Koehler (2011) introduced the concept of a "geo experiment." In a geo experiment, a geographic region of interest (e.g., a country) is first partitioned into a set of smaller non-overlapping "geos" subject to several constraints:

1. The advertiser must be able to serve their ads to each individual geo with some reasonable amount of accuracy.

2. It must be possible to track the metric(s) of interest (generally the online advertising spend and some other response metric) at the geo level.

3. The geos must be relatively self-contained so as to minimize any contamination effects that may incidentally occur (e.g., when consumers travel across geo boundaries).

After these geos have been defined, a subset of them will then be selected for experimentation by assigning them to the different experimental groups. This assignment is typically done at random, where it can oftentimes be helpful to constrain the randomization (e.g., by first stratifying the geos on some characteristic) since doing so may increase the precision of the causal estimates (Vaver and Koehler, 2011; Kerman et al., 2017).

However, it may not always be possible to rely on randomization when designing a particular geo experiment. For example, it may not be realistic to expect randomization to create balanced experimental groups when some of the

geos are markedly different from all of the others (e.g., those containing large cities), or when there are only a few geos available for experimentation (e.g., when designing a geo experiment in smaller countries). In addition, randomization may not always be feasible given some of the specific requirements that advertisers often must impose on their experiments in practice—such as the need to run a smaller scale experiment within a given budget, or the need to include certain geos in specific experimental groups. Consequently, advertisers may sometimes prefer to forgo some of the benefits of randomization in favor of a more systematic way of assigning their geos to experimental groups.

In this paper, we introduce a "matched markets" approach for designing geo experiments that allows advertisers to constrain the experimental group assignments of their geos. In particular, we propose a hill climbing algorithm that, subject to the advertiser's assignment constraints, greedily searches for experimental designs which appear to satisfy some of the critical assumptions of the "Time-Based Regression" (TBR) model for analyzing geo experiments that was introduced in Kerman et al. (2017). If the assumptions of the TBR model do indeed hold, then the experimental designs that are recommended by our matched markets approach lead to straightforward estimates of the causal effects of the geo experiments that are run.

The rest of this paper is organized as follows. In Section 2, we provide the necessary background by first reviewing how a geo experiment is designed, and then by discussing how a geo experiment can be subsequently analyzed using TBR. We then introduce our proposed matched markets approach for designing geo experiments in Section 3, which we further motivate in Section 4 through the use of two simulated examples and one real data example. Finally, Section 5 concludes.

2 Background

In this section, we first briefly review how a geo experiment is designed in the online advertising context. Afterwards, we provide a short overview of the TBR framework for analyzing geo experiments that was proposed by Kerman et al. (2017).

2.1 Designing a Geo Experiment

Suppose that an advertiser has geos $i = 1, \ldots, N$ available for experimentation, where the advertiser's goal is to measure the causal effect that some modification to their online advertising campaign (e.g., adding new targeted keywords) has on some metric of interest (e.g., their online sales revenue). In addition to deciding on how large of a change to make to their online advertising campaign, the advertiser must also consider several other design parameters when setting up their geo experiment—some of which we discuss in this section. For a more complete description of the design and structure of a geo experiment, we refer the reader to Vaver and Koehler (2011) and Kerman et al. (2017).

When designing a geo experiment, the advertiser must specify the experimental group assignment for each geo i, which we denote as

$$a_i \in \{control, treatment, unassigned\}.$$

These assignments subsequently induce the geo experiment's treatment, control, and unassigned groups, which we respectively denote as

$$\begin{aligned} \mathcal{G}_{trt} &= \{i \mid a_i = treatment\}, \\ \mathcal{G}_{ctl} &= \{i \mid a_i = control\}, \\ \mathcal{G}_{uad} &= \{i \mid a_i = unassigned\}, \end{aligned} \tag{1}$$

where $\mathcal{G}_{trt}$ and $\mathcal{G}_{ctl}$ must both be nonempty for the experimental design to be valid, and where the possibly empty set $\mathcal{G}_{uad}$ contains all of the geos that will be excluded from the experiment. As discussed in Section 1, this assignment is typically done using some form of randomization, although other assignment strategies are possible. However, it is important to note that different assignment mechanisms offer varying levels of protection against potential confounding variables and, as a result, may also either facilitate or preclude certain types geo experiment analysis or conclusions.

In addition to the experimental group assignments, the advertiser must also define the geo experiment's pretest, intervention, and cooldown periods. During the pretest period, the treatment and control groups are both kept in some common baseline state. Afterwards, in the intervention period, the campaign modification of interest is applied to all of the geos in the treatment group. Finally, during the cooldown period, the geos in the treatment group are returned to their original baseline state (e.g., by removing any targeted keywords that were added during the intervention period). In terms of notation, we let $\mathcal{T}_0 = \{1, \ldots, T_0\}$ denote the set of T_0 dates belonging to the pretest period, we let $\mathcal{T}_1 = \{T_0 + 1, \ldots, T\}$ denote the set of $T_1 = T - T_0$ "test period" dates belonging to either the intervention or cooldown periods, and we let $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1$ denote the set of all T dates that are under consideration for the geo experiment. Note that the test period $\mathcal{T}_1$ includes both the intervention and cooldown periods in order to help account for any delayed advertising effects that the intervention may have caused.

When designing their geo experiment, it is also imperative for the advertiser to do a power analysis ahead of time to determine whether their planned experiment offers a suitable amount of statistical power. An experiment that has too little power may result in causal estimates that have too much uncertainty to be practically useful, while an experiment that has too much power may lead to overspending and causal estimates that are more precise than necessary for the advertiser to make an informed decision. *A priori* power calculations that can be used to help guide the choice of acceptable geo experiment design parameters are discussed in more detail in Vaver and Koehler (2011) and Kerman et al. (2017).

Finally, at the conclusion of the geo experiment, various methods can be used for estimating its causal effect—such as the "Geo-Based Regression" linear

model discussed by Vaver and Koehler (2011, 2012) or the Bayesian structural time series "CausalImpact" model described by Brodersen et al. (2015). However, because one of the primary use cases of our proposed matched markets approach is to facilitate the design of smaller scale geo experiments, our assumes that the Time-Based Regression (TBR) model will be used for analysis since it is capable of being applied to experiments with a limited number of geos (Kerman et al., 2017).

2.2 Time-Based Regression (TBR)

Introduced in Kerman et al. (2017), TBR estimates the causal effect of a geo experiment by predicting the counterfactual time series of the treatment group.

For some metric of interest m (e.g., online advertising spend or online sales), let $m_{i,t}$ denote the observed value of m for geo i at time t. TBR begins by separately aggregating this metric within the treatment and control groups on each date during the geo experiment $t \in \mathcal{T}$:

$$\begin{aligned} y_t &= \sum_{i \in \mathcal{G}_{trt}} m_{i,t}, \\ x_t &= \sum_{i \in \mathcal{G}_{ctl}} m_{i,t}, \end{aligned} \tag{2}$$

where y_t and x_t denote the observed time series for metric m at time t for the treatment and control groups defined by equation (1), respectively.

Following the Rubin causal model framework (Holland, 1986), for each date $t \in \mathcal{T}$, denote the treatment group's potential outcomes that would have occurred in the absence or presence of an intervention as $y_t^{(0)}$ and $y_t^{(1)}$, respectively. Similarly, for each date $t \in \mathcal{T}$, let $x_t^{(0)}$ and $x_t^{(1)}$ denote the control group's potential outcomes that would have been observed in the absence or presence of an intervention, respectively. One of the goals of TBR is to estimate the cumulative causal effect of the intervention on the treatment group's metric m during the geo experiment's test period $\mathcal{T}_1$. In the Rubin causal model, this quantity can be expressed as

$$\Delta_m = \sum_{t \in \mathcal{T}_1} \left(y_t^{(1)} - y_t^{(0)} \right). \tag{3}$$

However, for any time t, it is impossible to simultaneously observe all of the potential outcomes. Instead, the observed data for the control group's time series is $x_t = x_t^{(0)}$ for all $t \in \mathcal{T}$ since the intervention is never applied to any of its geos, while the observed data for the treatment group's time series is

$$y_t = \begin{cases} y_t^{(0)} & \text{if } t \in \mathcal{T}_0 \\ y_t^{(1)} & \text{if } t \in \mathcal{T}_1 \end{cases}.$$

Therefore, to measure the cumulative causal effect of the intervention as defined in equation (3), it is necessary to estimate the unobserved potential outcomes $y_t^{(0)}$ for each test period date $t \in \mathcal{T}_1$.

TBR accomplishes this by positing the following linear relationship between the treatment and control group's time series of potential outcomes that would have occurred in the absence of an intervention for the entire duration of the geo experiment:

$$y_t^{(0)} = \alpha + \beta x_t^{(0)} + \epsilon_t \quad (t \in \mathcal{T}), \tag{4}$$

where $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. But because the potential outcomes $y_t^{(0)}$ are not actually observed for any of the test period dates $t \in \mathcal{T}_1$, TBR assumes that the model parameters (α, β, σ) from equation (4) coincide with the parameters of the analogous linear model that is defined using just the geo experiment's pretest period

$$y_t^{(0)} = \alpha + \beta x_t^{(0)} + \epsilon_t \quad (t \in \mathcal{T}_0), \tag{5}$$

which can be estimated from the observed data since $y_t = y_t^{(0)}$ and $x_t = x_t^{(0)}$ for all pretest period dates $t \in \mathcal{T}_0$ as the intervention had not yet taken place. TBR then uses Bayesian inference to estimate the joint posterior distribution of (α, β, σ) under a standard noninformative prior distribution that is uniform on $(\alpha, \beta, \log(\sigma))$, which results in the posterior mean and covariance of the conditional normal distribution of (α, β) given σ coinciding with the classical regression point estimates and covariance matrix (Gelman et al., 2013).

TBR will then use this estimated joint posterior distributions for (α, β, σ), the observed control time series x_t, and the hypothesized linear relationship in equation (4) to derive posterior predictive distributions for the unobserved potential outcomes $y_t^{(0)}$ for each test period date $t \in \mathcal{T}_1$. Finally, as can be seen in equation (3), this subsequently induces a posterior distribution for Δ_m which can be used to estimate the cumulative causal effect of the intervention on the treatment group's metric m during the geo experiment's test period $\mathcal{T}_1$.

Although the process that was just described allows advertisers to infer their marketing change's causal effect on a single metric m, advertisers are oftentimes more interested in understanding the efficiency of their marketing change on some response metric r relative to some cost metric c. Within the TBR framework, this type of analysis is accommodated by following the TBR procedure for the two different metrics of interest $m \in \{r, c\}$. After the posterior distributions for the geo experiment's cumulative causal effects on the response and cost metrics have been inferred—which we denote as Δ_r and Δ_c, respectively—the posterior distribution for the intervention's cumulative incremental return on ad spend (iROAS) during the geo experiment's test period $\mathcal{T}_1$ can then be measured through the ratio

$$\text{iROAS} = \frac{\Delta_r}{\Delta_c},$$

which Kerman et al. (2017) estimate by sampling from the two posterior distributions.

3 A TBR Matched Markets Approach for Designing Geo Experiments

As with any type of study, advertisers must determine the extent to which they trust the validity of the experimental design and modeling assumptions that are being made. In particular, for randomized controlled trials, experimenters rely on randomization because it will, *on average*, yield experimental groups that are balanced on all potential confounding factors. However, for any *particular* experiment, it is still possible for randomization to lead to groups that are noticeably unbalanced and which may, in turn, result in confounded conclusions. Indeed, there is a large literature which warns of the dangers of depending on randomization to balance covariates (Urbach, 1985; Krause and Howard, 2003; Rubin, 2008), as well as research which discusses the appropriate steps that should be taken if an imbalance in the covariates is observed (Urbach, 1985; Rubin, 2008; Bruhn and McKenzie, 2009; Morgan and Rubin, 2012).

Randomization can be especially challenging in the case of geo experiments. For example, it may not always be realistic to expect randomization to generate well balanced treatment and control groups in practice when there are some geos which are markedly different from all of the others, or when there are only a few geos available for experimentation. Moreover, randomization may not always be feasible in a geo experiment given some of specific requirements that advertisers often must impose on their experiments—such as the need to run a smaller scale experiment within a given budget, or the need to include certain geos in specific experimental groups. Consequently, in this section, we introduce a matched markets approach as an alternative option for advertisers who may be willing to forgo some of the benefits of randomization in favor of a more systematic way of assigning their geos to experimental groups. In particular, we propose a hill climbing algorithm that, subject to the advertiser's assignment constraints, greedily searches for experimental group assignments that appear to lead to valid and effective TBR models relative to the pretest period. If the assumptions of TBR do indeed hold, then the experimental designs that are recommended by our approach lead to straightforward estimates of the causal effects of the geo experiments that are run.

3.1 Assessing TBR Relative to the Pretest Period $\mathcal{T}_0$

For each geo $i = 1, \ldots, N$ that is available for experimentation, let the nonempty set

$$\mathcal{A}_i \subseteq \{treatment, control, unassigned\}$$

denote its set of possible experimental group assignments as stipulated by the advertiser. For instance, the advertiser would specify $\mathcal{A}_i = \{treatment\}$ if they require geo i to be in the treatment group, while they would set $\mathcal{A}_i = \{control, unassigned\}$ if they are allowing geo i to be assigned to either the control or unassigned group.

Given the set of allowable experimental group assignments $\mathcal{A}_i$ and the metric(s) of interest during some pretest period $\mathcal{T}_0$ for every geo i, the goal of our proposed matched markets approach is to recommend treatment and control groups for which a geo experiment can be run. In particular, this will involve explicitly specifying the group assignment $a_i \in \mathcal{A}_i$ for each geo $i = 1, \ldots, N$, which will then subsequently engender the experimental groups as defined in equation (1).

Because one of the motivating uses cases for our proposed matched markets approach is to facilitate the design of smaller scale geo experiments, we assume that TBR will be used to analyze the geo experiment since, unlike some other models, TBR is capable of being applied to experiments with a limited number of geos (Kerman et al., 2017). Consequently, we would like to be able to recommend experimental designs where we'd expect TBR to have the highest chance of being valid and effective.

In particular, recall from equation (4), that TBR postulates a linear relationship between the treatment and control group's potential outcomes that would have occurred in the absence of an intervention that holds for the entire duration of the geo experiment $t \in \mathcal{T}$. But because the treatment group's potential outcomes $y_t^{(0)}$ are no longer observable for the test period dates $t \in \mathcal{T}_1$ after the intervention occurs, TBR further assumes that the parameters for its hypothesized linear model coincide with the parameters for the linear model that it estimates from just the pretest period data $t \in \mathcal{T}_0$ as defined by equation (5). These estimated linear model parameters are then used in conjunction with the observed control time series x_t to derive posterior predictive distributions for the unobserved potential outcome $y_t^{(0)}$ on each test period date $t \in \mathcal{T}_1$.

Note, however, that the unobservable nature of the potential outcomes $y_t^{(0)}$ after the occurrence of an intervention does also complicate matters when assessing some of TBR's modeling assumptions. In particular, it is impossible to verify whether the linear relationship that is conjectured by TBR in equation (4) does indeed hold for the entirety of the geo experiment, which makes it difficult to assess whether TBR's estimated linear model in equation (5) accurately predicts the potential outcomes $y_t^{(0)}$ for each test period date $t \in \mathcal{T}_1$.

But when designing a geo experiment, it is still possible to evaluate certain aspects of TBR relative to the pretest period $\mathcal{T}_0$. As an example, consider the following generalization of TBR's estimated linear model as defined in equation (5) with time varying regression coefficients α_t and β_t:

$$y_t^{(0)} = \alpha_t + \beta_t x_t^{(0)} + \epsilon_t \quad (t \in \mathcal{T}_0), \tag{6}$$

where $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. Next, suppose that we would like to test the null hypothesis H_0 of "no structural breaks" in the regression coefficients against the alternative hypothesis H_1 of "at least one structural break" occurring, which can be mathematically expressed as

$$\begin{aligned} &H_0 : \forall t \in \mathcal{T}, \alpha_t = \alpha \text{ and } \beta_t = \beta, \\ &H_1 : \exists t \in \mathcal{T} \text{ such that } \alpha_t \neq \alpha \text{ or } \beta_t \neq \beta \end{aligned} \tag{7}$$

for some constants α and β (Zeileis et al., 2002). If the null hypothesis H_0 holds, then the linear model defined in equation (6) will reduce to TBR's estimated linear model as given by equation (5). However, if the null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1, then TBR may be misspecified and using it to predict the unobserved potential outcomes $y_t^{(0)}$ for the test period dates $t \in \mathcal{T}_1$ can be very costly—possibly resulting in unreliable inferences and misleading conclusions (Pesaran and Timmermann, 2004).

Several classes of structural break tests have been proposed for carrying out the hypothesis test described by equation (7). These include maximum likelihood scores, F-statistics, and fluctuation tests, with certain types of tests being more or less powerful depending on the specific pattern of deviation from the null hypothesis—we refer the reader to Zeileis et al. (2002) and Zeileis (2005) for a more extensive discussion of these methods. Consequently, these tests can be employed to elicit a p-value that indicates how incompatible the data are with the TBR's assumption of no structural breaks—with larger p-values suggesting higher degrees of compatibility between the data and the null hypothesis (Wasserstein and Lazar, 2016).

Similarly, the p-values obtained from performing other hypothesis tests on TBR's estimated linear model—such as tests for normality, homoscedasticity, or autocorrelation of the residuals—can also be used to quantify how incompatible an experimental design's pretest period data is with TBR. And although this does not necessarily guarantee that the assumptions of TBR will continue to hold for the entire duration of the geo experiment $t \in \mathcal{T}$ as posited by TBR in equation (4), it is reasonable to think that our proposed matched markets approach should favor designs that appear to be more compatible with the assumptions of TBR during the pretest period $\mathcal{T}_0$ since we may be more inclined to believe that these assumptions will continue to hold during the test period $\mathcal{T}_1$. Furthermore, if there is already evidence that its assumptions are violated during the pretest period $\mathcal{T}_0$, then TBR may already be misspecified, and using it may result in unreliable conclusions.

However, although these p-values can indicate how consistent the data are with some of TBR's assumptions, they do not assess the amount of statistical power that TBR provides. Kerman et al. (2017) discuss how various experimental design parameters (e.g., the volume of the treatment group, the correlation between the treatment and control groups, etc.) influence the estimation precision of TBR, which they define in terms of the half-width of the estimate's two-sided interval as this corresponds to the smallest effect size that can be detected at a given statistical significance level. Therefore, in addition to preferring experimental designs which appear to be more compatible with the assumptions of TBR during the pretest period $\mathcal{T}_0$, it is reasonable to think that our proposed matched markets approach should also favor designs that provide more statistical power.

To help formalize these experimental design preferences, we let $f(\mathcal{G}_{trt}, \mathcal{G}_{ctl})$ denote some objective function of the treatment and control groups to be maximized which quantifies the quality of TBR's estimated linear model as defined

by equation (5) relative to the pretest period $\mathcal{T}_0$ and with respect to the advertiser's requirements. Although the precise definition of $f(\mathcal{G}_{trt}, \mathcal{G}_{ctl})$ will vary from advertiser to advertiser and from experiment to experiment, in practice we've found that taking the minimum value amongst the p-value obtained from doing an OLS-based CUSUM structural breaks test (Ploberger and Krämer, 1992) on the estimated TBR model, the p-value obtained from doing a Breusch-Godfrey test for autocorrelation in the estimated TBR model's errors, and the estimated TBR model's R^2 to be a reasonable starting point for f since this penalizes experimental designs which appear to either be very inconsistent with the modeling assumptions of TBR or that lead to TBR model's that provide a low amount of statistical power. Meanwhile, for the situations where the advertiser is interested in using the TBR framework to estimate the iROAS, we've found that defining separate objective functions f_r and f_c for the response and cost metrics, respectively, and then setting $f = \min(f_r, f_c)$ to work well as a maximin strategy.

Unfortunately, although the objective function f helps to quantify the quality of the estimated TBR model with respect to the pretest period $\mathcal{T}_0$, it will typically not be tractable since the influence of an individual geo's experimental group assignment on equation (5) is generally not easily expressed. Depending on the total number of geos N and the specific constraints that an advertiser imposes on the set of allowable experimental group assignments $\mathcal{A}_i$ for each geo i, it may be possible to employ an exhaustive search to find the group assignments that maximize f. However, this is often not feasible in practice—for example, in a situation with N geos and no constraints placed on the set of allowable experimental group assignments $\mathcal{A}_i$, this brute force approach will be $O(3^N)$ since each individual geo can be assigned to either the treatment, control, or unassigned group. Consequently, an alternative optimization strategy is required, and we refer the reader to Michalewicz and Fogel (2004) for a comprehensive overview of some of the possible methods that can be used.

3.2 Hill Climbing Algorithm

Although several techniques exist for optimizing an objective function f, in this section, we propose our own variant of a hill climbing algorithm which aims to provide advertisers with several recommended experimental designs of varying treatment group sizes so that they can select the geo experiment that best suits their needs. In particular, this algorithm alternates between a "matching" routine which greedily looks for the best set of control geos given the current set of treatment geos, and an "augmentation" routine that greedily tries to add one new geo to the set of treatment geos given the current control group. The procedure is repeated until the treatment set reaches its maximum allowable size.

In addition to defining their objective function f, our proposed matched markets hill climbing algorithm also requires the advertiser to specify their set of allowable experimental group assignments $\mathcal{A}_i$ and to provide their metric(s) of interest $m_{i,t}$ for each geo $i = 1, \ldots, N$ on some historical pretest period dates

Algorithm 1: Proposed Matched Markets Hill Climbing Algorithm

Data: For each geo $i = 1, \ldots, N$ during some historical pretest period dates $t \in \mathcal{T}_0$:
- Response metrics $r_{i,t}$
- Cost metrics $c_{i,t}$
- Allowable group assignments $\mathcal{A}_i$

Objective function $f(\mathcal{G}_{trt}, \mathcal{G}_{ctl})$
Positive integer $K < N$ specifying the maximum allowable treatment group size

Define initial treatment and control groups

$$\begin{aligned} \mathcal{G}^*_{trt,k_0} &\leftarrow \{i \mid \mathcal{A}_i = \{treatment\}\} \\ \mathcal{G}_{ctl,k_0} &\leftarrow \{i \mid control \in \mathcal{A}_i\} \end{aligned} \tag{8}$$

where $k_0 \leftarrow \left|\mathcal{G}^*_{trt,k_0}\right|$ is the size of the initial treatment group

if k_0 *is* 0 **then**
  Set $\mathcal{G}^*_{ctl,k_0} \leftarrow \mathcal{G}_{ctl,k_0}$
  $needs\_matching \leftarrow FALSE$
else
  $needs\_matching \leftarrow TRUE$

$k \leftarrow k_0$
while $k \leq K$ **or** $needs\_matching$ *is* $TRUE$ **do**

  if $needs\_matching$ *is* $TRUE$ **then**
    Define the set of geos that can be reassigned to the control group:

$$\mathcal{R}_{ctl} \leftarrow \left\{i \notin \mathcal{G}_{ctl,k} \mid control \in \mathcal{A}_i, i \notin \mathcal{G}^*_{trt,k}, \textit{(optional constraints)}\right\} \tag{9}$$

    Define the set of geos that can be reassigned to the unassigned group:

$$\mathcal{R}_{uad} \leftarrow \left\{i \in \mathcal{G}_{ctl,k} \mid unassigned \in \mathcal{A}_i, i \notin \mathcal{G}^*_{trt,k}, \textit{(optional constraints)}\right\} \tag{10}$$

    Define the "neighboring" control group:

$$\mathcal{G}^{'}_{ctl,k} \leftarrow \mathcal{G}_{ctl,k} \oplus \left\{ \underset{i \in \mathcal{R}_{ctl} \cup \mathcal{R}_{uad}}{\arg\max} f\left(\mathcal{G}^*_{trt,k}, \mathcal{G}_{ctl,k} \oplus \{i\}\right) \right\} \tag{11}$$

    where $\oplus$ denotes the symmetric difference set operation.
    if $f(\mathcal{G}^*_{trt,k}, \mathcal{G}^{'}_{ctl,k}) > f(\mathcal{G}^*_{trt,k}, \mathcal{G}_{ctl,k})$ **then**
      Update the control group: $\mathcal{G}_{ctl,k} \leftarrow \mathcal{G}^{'}_{ctl,k}$
    else
      Define the recommended control group: $\mathcal{G}^*_{ctl,k} \leftarrow \mathcal{G}_{ctl,k}$
      $needs\_matching \leftarrow FALSE$

  if $needs\_matching$ *is* $FALSE$ **and** $k < K$ **then**
    Define the set of geos that can be reassigned to the treatment group:

$$\mathcal{R}_{trt} \leftarrow \left\{i \notin \mathcal{G}^*_{trt,k} \mid treatment \in \mathcal{A}_i, \textit{(optional constraints)}\right\} \tag{12}$$

    Define the recommended treatment group $\mathcal{G}^*_{trt,k+1}$ by augmenting $\mathcal{G}^*_{trt,k}$:

$$\mathcal{G}^*_{trt,k+1} \leftarrow \mathcal{G}^*_{trt,k} \cup \left\{ \underset{i \in \mathcal{R}_{trt}}{\arg\max} f(\mathcal{G}^*_{trt,k} \cup \{i\}, \mathcal{G}^*_{ctl,k}) \right\} \tag{13}$$

    Define the control group: $\mathcal{G}_{ctl,k+1} \leftarrow \mathcal{G}^*_{ctl,k}$
    $k \leftarrow k + 1$
    $needs\_matching \leftarrow TRUE$

Result: A recommended treatment group $\mathcal{G}^*_{trt,k}$ and its recommended "matching" control group $\mathcal{G}^*_{ctl,k}$ for each treatment group of size $k = \max(k_0, 1), \ldots, K$

$t \in \mathcal{T}_0$. Letting $k_0 = |\{i \mid \mathcal{A}_i = \{treatment\}\}|$ denote the number of geos that the advertiser has required to be assigned to the treatment group, our proposed algorithm also requires advertisers to specify some positive integer $K \geq k_0$ which indicates the maximum number of geos in the treatment group that they are willing to allow for their geo experiment.

Given these inputs, the goal of our proposed matched markets hill climbing algorithm is to provide the advertiser with several different experimental design choices—one for each treatment group of size $k = \max(k_0, 1), \ldots, K$. In particular, for each k, this will entail specifying a recommended treatment group $\mathcal{G}^*_{trt,k}$ and its recommended "matching" control group $\mathcal{G}^*_{ctl,k}$, where an asterisk is used in the superscript to differentiate these recommended groups from other non-recommended groups. Furthermore, note that $k = \left|\mathcal{G}^*_{trt,k}\right|$ since the recommended treatment group $\mathcal{G}^*_{trt,k}$ will always, by definition, contain exactly k geos. However, $\mathcal{G}^*_{trt,k}$'s corresponding matching control group $\mathcal{G}^*_{ctl,k}$ will not necessarily be of size k—that is, the subscript k for the recommended control group $\mathcal{G}^*_{ctl,k}$ only indicates which recommended treatment group $\mathcal{G}^*_{trt,k}$ it is paired with.

To achieve its goal, our proposed hill climbing algorithm begins by initializing all of the geos to the experimental groups as defined by equation (8) in line 1 of Algorithm 1. In particular, the initial recommended treatment group $\mathcal{G}^*_{trt,k_0}$ contains all of the geos which the advertiser has *required* to be assigned to the treatment condition, while the initial control group $\mathcal{G}_{ctl,k_0}$ consists of all of the geos which the advertiser has *allowed* to be assigned to the control condition. Afterwards, our proposed matched markets hill climbing algorithm will then repeatedly alternate between a matching routine and an augmentation routine until the stopping rule is reached, and where lines 2-6 of Algorithm 1 determine which routine is used first—a decision that depends on whether or not the advertiser has required any of its geos to be in the treatment group.

In the matching routine outlined by lines 9-16 of Algorithm 1, the matching control group $\mathcal{G}^*_{ctl,k}$ for a given recommended treatment group $\mathcal{G}^*_{trt,k}$ is found by incrementally updating an existing nonrecommended control group $\mathcal{G}_{ctl,k}$ until a local optimum is reached. This is accomplished by first finding the sets $\mathcal{R}_{ctl}$ and $\mathcal{R}_{uad}$ as defined by equations (9) and (10) which contain the geos that are eligible to be reassigned to either the control or unassigned groups, respectively. Afterwards, as defined by equation (11), a "neighboring" control group $\mathcal{G}^{'}_{ctl,k}$ is derived from $\mathcal{G}_{ctl,k}$ by reallocating the geo whose reassignment—either from the control group to the unassigned group, or from the unassigned group to the control group—maximizes f when used in conjunction with the recommended treatment group $\mathcal{G}^*_{trt,k}$. Then, as described by lines 12-13 of Algorithm 1, if $f(\mathcal{G}^*_{trt,k}, \mathcal{G}^{'}_{ctl,k}) > f(\mathcal{G}^*_{trt,k}, \mathcal{G}_{ctl,k})$—that is, if $\mathcal{G}^{'}_{ctl,k}$ appears to lead to a higher quality TBR model than $\mathcal{G}_{ctl,k}$ when paired with $\mathcal{G}^*_{trt,k}$—then the algorithm will update its definition of the control group $\mathcal{G}_{ctl,k}$ to coincide with $\mathcal{G}^{'}_{ctl,k}$, and this updated control group will then be used in the next iteration of the matching routine. However, if $f(\mathcal{G}^*_{trt,k}, \mathcal{G}^{'}_{ctl,k}) \leq f(\mathcal{G}^*_{trt,k}, \mathcal{G}_{ctl,k})$—that is, if a local optimum has been reached—then as lines 14-16 of Algorithm 1 indicate, the

hill climbing algorithm will take the existing set $\mathcal{G}_{ctl,k}$ to be the recommended matching control group $\mathcal{G}^*_{ctl,k}$ for its recommended treatment group $\mathcal{G}^*_{trt,k}$ of size k.

Meanwhile, the augmentation routine detailed in lines 17-22 of Algorithm 1 is used to derive a larger recommended treatment group $\mathcal{G}^*_{trt,k+1}$ of size $k+1$ from an existing recommended treatment group $\mathcal{G}^*_{trt,k}$ of size k. To accomplish this, the algorithm first finds the set of geos $\mathcal{R}_{trt}$ that are eligible to be reassigned to the treatment group as defined by equation (12). Afterwards, as can be seen from equation (13), the recommended treatment group $\mathcal{G}^*_{trt,k+1}$ of size $k+1$ is then constructed by simply augmenting the recommended treatment group $\mathcal{G}^*_{trt,k}$ of size k with the geo whose reassignment to the treatment group maximizes f when used in combination with the recommended control group $\mathcal{G}^*_{ctl,k}$. Finally, as lines 20-22 of Algorithm 1 show, the recommended control group $\mathcal{G}^*_{ctl,k}$ is then taken to be the starting point for the subsequent matching routine will be used to find the matching control group $\mathcal{G}^*_{ctl,k+1}$.

As indicated by line 8 of Algorithm 1, our proposed matched markets hill climbing algorithm continues to alternate between the augmentation and matching routines until it has determined a recommended treatment group $\mathcal{G}^*_{trt,k}$ and its corresponding matching control group $\mathcal{G}^*_{ctl,k}$ for experimental designs with a treatment group of size $k = \max(k_0, 1), \ldots, K$. Moreover, each of these recommended designs will locally optimize the advertiser's objective function f in terms of the advertiser's requirements, and if the assumptions of TBR do indeed hold, then the geo experiments that are recommended by our matched markets approach lead to straightforward causal estimates. Furthermore, as discussed in Section 2.1, a power calculation can be done for each of these recommended experimental designs in order to obtain an estimate of each design's experimental cost. In particular, Kerman et al. (2017) showed that the cost of experimentation tends to proportionally increase as the volume of the treatment group increases. Therefore, because the volume in the treatment groups recommended by our proposed hill climbing algorithm increases with k, our proposed matched markets approach is able to provide advertisers with several geo experiment design options of varying experimental costs.

However, in addition to any budgetary constraints that they may have, it is also important for advertisers to consider the limitations of our proposed matched markets approach before deciding on which of the recommended experimental design to use for their geo experiment, if any. First, it may be more difficult to generalize the results a nonrandomized matched markets geo experiment since the treatment group may not necessarily be representative of the target population of interest. Second, although the recommended matched markets designs are locally optimal with respect to the advertiser's experimental group assignment constraints $\mathcal{A}_i$ and objective function f, it is imperative that advertisers verify the feasibility of the experimental design prior to running their geo experiment since the local optima are not necessarily guaranteed to lead to viable experiments—for example, it is probably not possible to recommend a workable geo experiment if an advertiser requires New York City, Chicago, and Los Angeles to all be in the treatment group. Finally, because our proposed

matched markets approach assumes that TBR will be used for the analysis, we also emphasize that the causal estimates may no longer be valid if the modeling assumptions of TBR do not hold throughout the entire duration of the geo experiment.

4 Examples

In this section, we present two simulated examples and one real data example that further motivate our proposed matched markets approach. Our first simulated example demonstrates the importance of designing an experiment that is consistent with the methodology that will be used to analyze it. Meanwhile, our second and third examples use simulated and real data, respectively, to compare how our proposed matched markets approach fares against randomization when determining an appropriate control group for TBR analysis.

4.1 Example 1

As a simple but instructive example highlighting the importance of designing an experiment that is consistent with the methodology that will be used to analyze it, suppose that an advertiser has three geos whose metrics on the dates $t \in \mathcal{Z}^+$ are, unbeknownst to the advertiser, generated according to the following noiseless processes:

$$\begin{aligned} m_{1,t} &= \frac{3}{2} + sin\left(\frac{\pi t}{27} - \frac{\pi}{4}\right), \\ m_{2,t} &= \frac{1}{4}\left(m_{1,t}\right)^2, \\ m_{3,t} &= 6 - m_{1,t}, \end{aligned} \tag{14}$$

where we note that data generating processes for geos 2 and 3 are both functions of geo 1. Plots of these three time series for $t \in 1, \ldots, 35$ can be seen in the top panel of Figure 1.

Suppose further that, with a pretest period of $\mathcal{T}_0 = \{1, \ldots, 28\}$, the advertiser would like to run an "A/A" geo experiment where no intervention actually occurs during the test period $\mathcal{T}_1 = \{29, \ldots, 35\}$—that is, the true cumulative incremental effect of the intervention is $\Delta_m = 0$—and where there is only one treatment geo and one control geo. Since they know that it can be applied to experiments with a limited number of geos (Kerman et al., 2017), the advertiser plans on analyzing their geo experiment using TBR. However, although the advertiser insists on geo 1 being included in the experiment as either the treatment or control geo, they are unsure of which of the other two geos to include in their experiment.

Matching prior to randomization is one approach that can be used to help make this decision. That is, the advertiser could first define some distance measure to compare the similarity between geos 1 and 2 against the similarity between geos 1 and 3, with the more similar geo forming a "matched pair" with

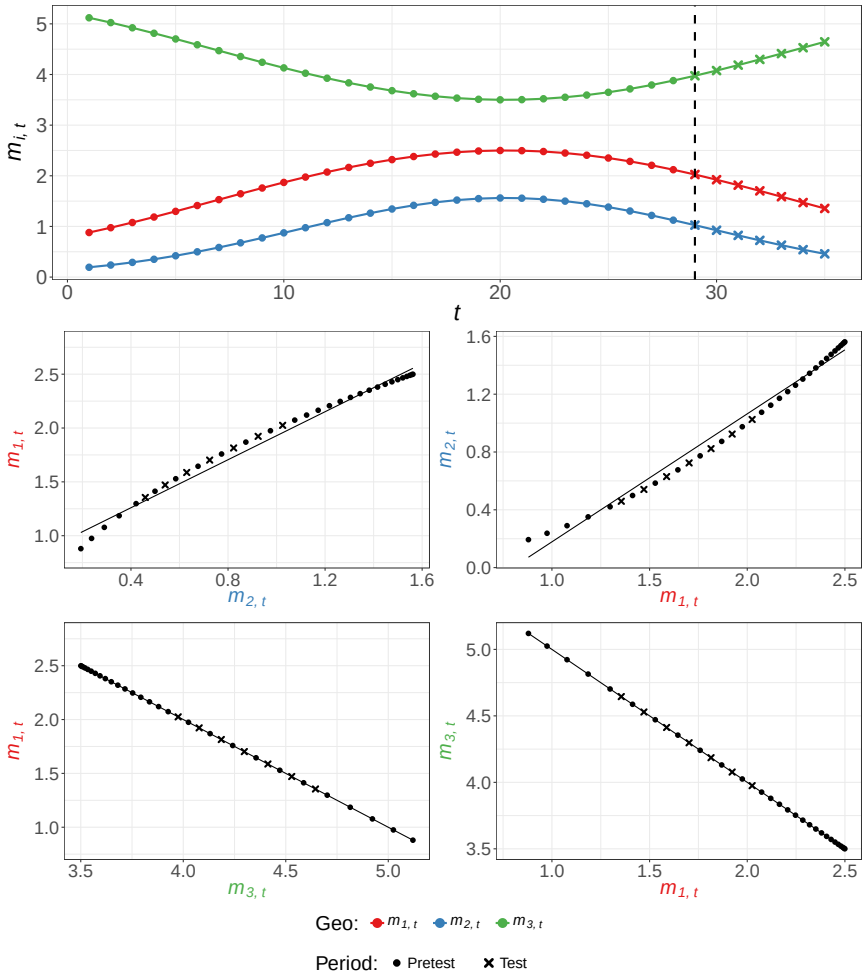

Figure 1: Top Panel: The time series for the three geos as defined in equation (14), where the vertical dashed line indicates the start the A/A experiment's test period. Middle Panels: If we pair geos 1 and 2, then TBR's estimated linear model will always be misspecified since the two geos have an underlying nonlinear relationship. Consequently, the advertiser will incorrectly infer that there is a treatment effect regardless of how the two geos are allocated to the experimental groups. Bottom Panels: If we pair geos 1 and 3, then TBR's estimated linear model will always be correctly specified since the two geos have an underlying linear relationship. Therefore, the advertiser will always be able to correctly infer that there is a null treatment effect.

geo 1. Then, after this match has been determined, one of the geos in the pair can be randomly allocated to the control group, leaving the other geo in the pair to be assigned to the treatment group.

If additional information on the three geos were available and thought to be important (e.g., consumer demographics, population size/density, customer acquisition rates, etc.), then they could be used to calculate these similarity scores—perhaps by using a method analogous to what was done in Ye et al. (2016) or by using the "optimal non-bipartite matching" algorithm proposed by Greevy et al. (2004). But recall from equations (4) and (5), that TBR only considers the time series for the metrics of interest—it does not require or use any other additional information. Therefore, although matching on this additional information may help to improve the balance of the covariates, such an approach does not necessarily help to ensure that the treatment and control groups lead to a correctly specified TBR model.

Alternatively, because TBR posits that the treatment group's time series is a function of the control group's time series, similarities in the historical behavior of the two time series (e.g., their trends, seasonality, etc.) can possibly be used by the advertiser to determine which geo to pair with geo 1. Many distance measures $d(x, y)$ have been proposed to evaluate the similarity between two time series x and y, and Table 1 contains the results from applying some of the more popular distance measures to this example for $t \in \mathcal{T}_0$ after the three time series have been normalized to have zero mean and unit variance—we refer the reader to Mori et al. (2016) for a more in-depth discussion of how these distance measures are defined and differ from one another. Despite their differences, however, from this table we see that all of them strongly favor creating a matched pair between geos 1 and 2. This is perhaps not too surprising given that, from the top panel of Figure 1, $m_{2,t}$ does appear to be visually more similar to $m_{1,t}$ than $m_{3,t}$ does. But from equation (14), we know that pairing geos 1 and 2 together would lead to a misspecified TBR model since they share a nonlinear relationship. In particular, as can be seen from the middle panels in Figure 1, if TBR were used to evaluate the A/A geo experiment, then the advertiser would incorrectly infer either a positive or negative incremental treatment effect depending on whether geo 1 is assigned to treatment or control.

On the other hand, since the advertiser knows ahead of time that they will be using TBR to analyze their geo experiment, they could define a similarity score that explicitly specifies the need for a stable and predictive linear relationship between the treatment and control time series—perhaps similar to the one that we suggested in Section 3.1. Such a similarity score would then choose to pair geo 3 with geo 1 since we know from equation (14) that these two geos do share an underlying linear relationship. Consequently, as can be seen from the bottom panels in Figure 1, if TBR were used to evaluate the A/A geo experiment, then the advertiser would be able to correctly recover a null incremental treatment effect regardless of the specific experimental groups that geos 1 and 3 are randomized to.

Time Series Distance Measure	$d(m_{1,t}, m_{2,t})$	$d(m_{1,t}, m_{3,t})$
Euclidean	**0.59**	10.39
Manhattan	**2.65**	47.24
Pearson Correlation	**0.11**	2.00
Dynamic Time Warping	**3.05**	61.74
Fourier	**2.21**	38.92
Wavelet	**0.58**	10.36

Table 1: Results from applying some popular time series distance measures to Example 1 after the three time series have been normalized to have zero mean and unit variance. Lower distances, which are highlighted in bold, indicate a higher degree of similarity. Therefore, all of the distance measures evaluated suggest that geos 1 and 2 should form a matched pair. However, we know from equation (14), that this would lead to a misspecified TBR model since they have a nonlinear relationship—suggesting that these distance measures may not be appropriate when analyzing a geo experiment with TBR.

4.2 Example 2

In our second example, we use simulated data to compare how our proposed matched markets hill climbing algorithm fares against randomization when determining an appropriate control group for TBR analysis. In particular, consider a scenario where an advertiser has geos $i = 1, \ldots, 6$ whose metrics on each date $t \in \mathbb{Z}^+$ are, unbeknownst to the advertiser, generated as follows:

$$m_{i,t} = \begin{cases} 100 + 25sin\left(\frac{\pi t}{105}\right) + \epsilon_{i,t} & \text{if } i \text{ is odd} \\ 100 + 25sin\left(\frac{\pi t}{105} - \frac{\pi}{8}\right) + \epsilon_{i,t} & \text{if } i \text{ is even} \end{cases}, \tag{15}$$

and where $\epsilon_{i,t} \overset{iid}{\sim} N(0, 1)$. Figure 2 depicts one example data set which was simulated according to equation (15).

Suppose further that the advertiser would like to use TBR to analyze an A/A geo experiment with pretest period $\mathcal{T}_0 = \{1, \ldots, 84\}$ and test period $\mathcal{T}_1 = \{85, \ldots, 91\}$. Furthermore, although the advertiser requires exactly 2 geos in the treatment group, they place no other constraints on the composition of the treatment and control groups.

Randomization provides one way of allocating the geos to the treatment and control groups. That is, the experimental groups can be determined by randomly assigning two of the geos to the treatment group, with the remaining four geos being assigned to the control group. But as discussed in Section 3, any particular randomized experiment may exhibit imbalances between the treatment and control groups. Specifically, for this example, we would intuitively expect the most suitable experimental groups to be the ones containing the same ratio of odd indexed to even indexed geos. However, consider the situation where randomization leads to a treatment group of two odd indexed geos.

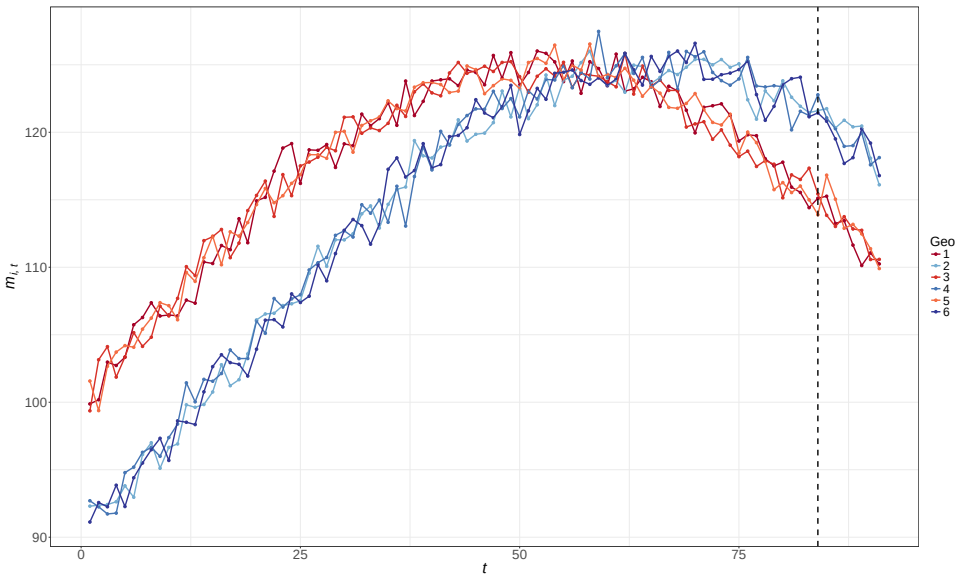

Figure 2: An example data set simulated from the data generating process defined in equation (15), where the vertical dashed line indicates the start of the A/A experiment's test period.

Although the optimal corresponding control group in this case would appear to be the one that consists of only the third odd indexed geo, randomization will yield a control group containing both odd and even indexed geos coming from both data generating processes, which would then subsequently result in a misspecified TBR model.

To evaluate how randomization fares in this example, we simulated 1000 randomized A/A geo experiment situations. In particular, for each experimental replication j, we randomly generated the six geos according to equation (15) and then randomized two of the geos to the treatment group and four of the geos to the control group. Finally, TBR was trained using the pretest period data $t \in \mathcal{T}_0$ to obtain an estimate of the intervention's cumulative causal effect $\hat{\Delta}_{m,j}$ and a 95% interval for this estimate $\left[\hat{\Delta}^L_{m,j}, \hat{\Delta}^U_{m,j}\right]$ during the test period $\mathcal{T}_1$.

Recall that since the advertiser is running an A/A geo experiment, the true cumulative incremental effect of the intervention is $\Delta_m = 0$. As a result, we can summarize the performance of randomization across all 1000 experimental replications in terms of its root mean squared error (RMSE), its 95% interval

Performance Measure	Randomization	Matched Markets
RMSE	72.312	**5.563**
CIHW	14.994	**9.202**
TypeI	0.423	**0.087**

Table 2: Comparison of TBR's performance across all 1000 experimental replications when either randomization or our proposed matched markets approach is used to determine the A/A geo experiment design in Example 2. Specific definitions of the performance measures are given in equation (16), with the better assignment scheme for each performance measure highlighted in bold.

half-width size, and its Type I error rate at the 0.05 significance level as follows:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}\hat{\Delta}_{m,j}^2}, \\ CIHW &= \frac{1}{1000}\sum_{j=1}^{1000}\frac{\hat{\Delta}_{m,j}^U - \hat{\Delta}_{m,j}^L}{2} \\ TypeI &= \frac{1}{1000}\sum_{j=1}^{1000} I\left(0 \notin \left[\hat{\Delta}_{m,j}^L, \hat{\Delta}_{m,j}^U\right]\right), \end{aligned} \tag{16}$$

and where $I(\cdot)$ is the indicator function. Results are shown in Table 2, where it can be seen that TBR's 95% intervals have a much higher than expected Type I error rate when randomization is used to allocate the geos—suggesting that randomization may be an inadequate method of creating an appropriate control group for TBR in this particular example.

Consequently, an alternative method of determining experimental groups may be preferred, and we evaluate how our proposed matched markets approach compares to randomization in this particular example. Specifically, for each experimental replication j, we first take its randomized treatment group as fixed. Afterwards, we apply our hill climbing algorithm with the objective function f suggested in Section 3.1 in order to determine which subset of its randomized control geos to use as the randomized treatment group's matching control group. Afterwards, TBR was used to analyze the matched markets A/A geo experiment, with results summarizing its performance across all 1000 experimental replications also appearing in Table 2. From this table, we see that our proposed matched markets approach appears to be more effective at determining a suitable control group than randomization—it achieves a lower RMSE, and shorter 95% intervals. Furthermore, on average, it attains a Type I error rate that is much more in line with expectations, with the situations where our proposed matched markets approach was unable to find a satisfactory matching control group partially responsible for this small inflation from the

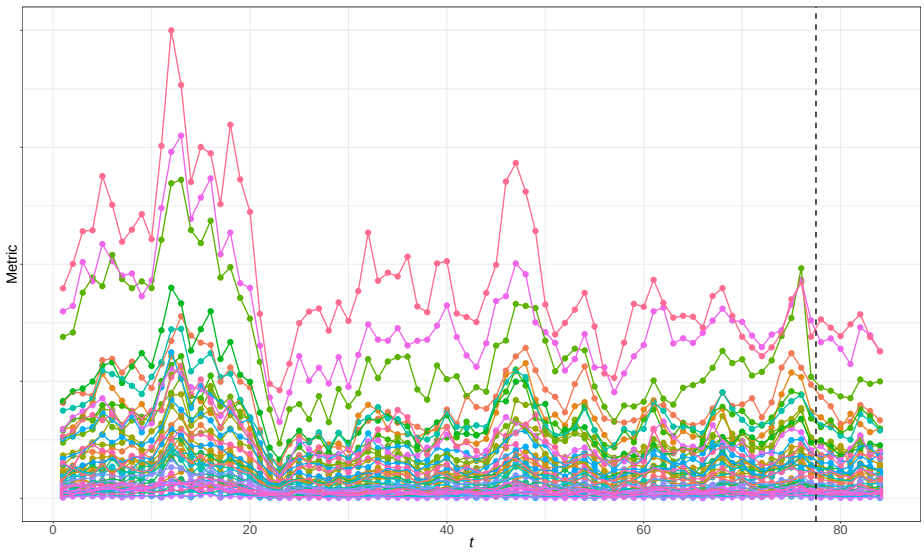

Figure 3: Time series plot of the 50 control geos that we consider in Example 3, where the vertical dashed line indicates the start of the A/A experiment's test period.

nominal Type I error rate.

Finally, we note that measurement error in the control group's time series may also be undermining TBR's performance in both the randomized and matched markets contexts. Therefore, if an advertiser has concerns about these measurement errors compromising the validity of their analysis, then they may want to consider using an "errors-in-variables" model instead, and we note that one such model—an orthogonal regression extension of TBR—is discussed in Kerman et al. (2017).

4.3 Example 3

For our third and final example, we now use real data to evaluate how our proposed matched markets hill climbing algorithm fares against randomization when determining an appropriate control group for TBR analysis. In particular, the data set that we consider originates from a geo experiment that was executed in the United States by one of Google's advertisers using 50 control geos and 50 treatment geos in order to measure the effectiveness of their Google Paid Search advertising efforts. However, for the purposes of establishing an approximate "ground truth" of no cumulative incremental effect from an A/A test, in this example we only consider the 50 control geos from the experiment. Time series data for each of these 50 control geos can be seen in Figure 3.[1]

[1] To anonymize the actual geo experiment, all of its numbers have been rescaled and all of its dates have been relabeled.

Performance Measure	Randomization	Matched Markets
RMSE	0.563	**0.312**
CIHW	0.629	**0.554**
TypeI	0.253	**0.067**

Table 3: Comparison of TBR's performance across all 1000 experimental replications when either randomization or our proposed matched markets approach is used to determine the A/A geo experiment design in Example 3. Specific definitions of the performance measures are given in equation (16), with the better assignment scheme for each performance measure highlighted in bold.

Similar to what was done previously in Example 2, we simulated and evaluated 1000 A/A geo experiment situations with pretest period $\mathcal{T}_0 = \{1, \ldots, 77\}$ and test period $\mathcal{T}_1 = \{78, \ldots, 84\}$. Specifically, for each experimental replication j, we first simulate a randomized design by randomly assigning 25 of the geos to the treatment group and 25 of the geos to the control group. Afterwards, within each experimental replication j, we then simulate a matched markets design by applying our hill climbing algorithm with the objective function f suggested in Section 3.1 in order to determine which subset of its randomized control geos to use as the randomized treatment group's matching control group. Finally, TBR was used to evaluate both of the experimental designs.

In Table 3, we summarize the 1000 experimental replications in terms of the performance measures defined previously in equation (16). From this table we see that TBR's 95% intervals have a much higher than expected Type I error rate when randomization is used to create the experimental design, while the matched markets design achieves a Type I error rate that is more in line with nominal expectations. Moreover, we see that the matched markets design also leads to causal estimates that are, on average, more accurate terms of its RMSE and more precise in terms of its interval half-width size.

5 Conclusions

Although randomized controlled trials are regarded as the gold standard for causal inference, it may not always be feasible for an advertiser to rely on randomization when designing their geo experiment. In this paper, we introduced a TBR matched markets approach as an alternative method for advertisers who may be willing to forgo some of the benefits of randomization in favor of a more systematic way of assigning their geos to experimental groups. In particular, we proposed a hill climbing algorithm that provides several geo experiment design options which locally optimize TBR's modeling assumptions subject to the advertiser's constraints. Consequently, advertisers can choose to run the geo experiment that best suits their needs, and if the assumptions of TBR do indeed hold, then the experimental designs that are recommended by our matched

markets approach lead to straightforward estimates of the causal effects of the geo experiments that are run.

Acknowledgments

The author would like to thank Jouni Kerman, Nicolas Remy, and Jim Koehler for many helpful discussions and comments.

References

Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274, 2015.

Miriam Bruhn and David McKenzie. In pursuit of balance: Randomization in practice in devlopment field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232, 2009.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

Brett Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measuremnet: Evidence from big field experiments at Facebook. White paper, 2017.

Robert Greevy, Bo Lu, Jeffrey H. Silber, and Paul Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275, 2004.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

G.W. Imbens and D.B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015.

Jouni Kerman, Peng Wang, and Jon Vaver. Estimating ad effectiveness using geo experiments in a time-based regression framework. Technical report, Google, Inc., 2017. URL `https://research.google.com/pubs/pub45950.html`.

Merton S. Krause and Kenneth I. Howard. What random assignment does and does not do. *Journal of Clinical Psychology*, 59(7):751–766, 2003.

Randall A. Lewis and Justin M. Rao. The Unfavorable Economics of Measuring the Returns to Advertising. *The Quarterly Journal of Economics*, 130(4): 1941–1973, 2015.

Zbigniew Michalewicz and David B. Fogel. *How to Solve It: Modern Heuristics*. Springer Berlin Heidelberg, New York, second, revised and extended edition edition, 2004.

Kari L. Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(20):1263–1282, 2012.

Usue Mori, Alexander Mendiburu, and Jose A. Lozano. Distance measures for time series in R: The TSdist package. *R journal*, 8(2): 451–459, 2016. URL `https://journal.r-project.org/archive/2016/RJ-2016-058/index.html`.

M. Hashem Pesaran and Allan Timmermann. How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting*, 20(3):411–425, 2004.

Werner Ploberger and Walter Krämer. The CUSUM test with OLS residuals. *Econometrica*, 60(2):271–85, 1992.

Donald B. Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484): 1350–1353, 2008.

Peter Urbach. Randomization and the design of experiments. *Philosophy of Science*, 52(2):256–273, 1985.

Jon Vaver and Jim Koehler. Measuring ad effectiveness using geo experiments. Technical report, Google Inc., 2011. URL `https://research.google.com/pubs/pub38355.html`.

Jon Vaver and Jim Koehler. Periodic measurement of advertising effectiveness using multiple-test-period geo experiments. Technical report, Google Inc., 2012. URL `https://research.google.com/pubs/pub38356.html`.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

Quinn Ye, Saarthak Malik, Ji Chen, and Haijun Zhu. 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, 2016. ACM.

Achim Zeileis. A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals. *Econometric Reviews*, 24(4):445–466, 2005.

Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002. URL `http://www.jstatsoft.org/v07/i02/`.