Meridian GeoX: An Open-Source Framework for Precision and Efficiency in Geo Experiments

Steven Ye[*,1], **Fabian Sinn**[*,1], **Yunxiao Li**[1] **and Yang Jiao**[1]
[*]Equal contributions, [1]Google

Geo experimentation is a crucial, privacy-conscious option for measuring media effectiveness. Historically, two primary challenges have hindered the adoption of geo experiments: the high costs required to overcome large variance across geographical regions, and the unreliability of conventional analysis methods under real-world autocorrelation and non-stationary trends. To address these challenges, we introduce Meridian GeoX, Google's open-source geo experiment framework. As a cornerstone of Google's modern measurement suite, Meridian GeoX is designed to standardize and optimize the end-to-end causal measurement lifecycle. The framework provides a unified suite of methodologies—including data-driven stratified sampling, advanced counterfactual models (Time-Based Regression, Synthetic Control, and Synthetic Difference-in-Differences), and a novel Design-Aware Placebo Inference engine. Extensive empirical benchmarking against industry alternatives demonstrates that Meridian GeoX delivers superior predictive accuracy, lowers Minimum Detectable Effects (MDE), and significantly reduces budget requirements. Furthermore, evaluations demonstrate that the framework's novel placebo inference maintains rigorous control over false positive rates while maximizing sensitivity under challenging conditions. By integrating seamlessly with the Meridian Marketing-Mix Model (MMM), this framework delivers a complete modern measurement solution. Ultimately, Meridian GeoX empowers advertisers with a powerful, cost-efficient, and methodologically robust solution for privacy-safe modern measurement, establishing a new industry standard for causal measurement and media effectiveness.

Keywords: experimentation, digital advertising, causal inference, modern measurement, open source

1. Introduction

Geo experiments, pioneered and formalized for digital advertising by Google (Vaver and Koehler, 2011), utilize non-overlapping geographical regions as experimental units. This method only requires aggregate data, such as per geo ad spend and response metrics (e.g., revenue or sales), making it a privacy-conscious alternative to user- (or cookie)-based experiments. As data protection laws like GDPR and browser and device restrictions on third-party cookies make user-based experiments increasingly difficult to execute, geo experiments have emerged as a vital tool for advertisers seeking reliable, future-proof, and privacy-safe measurement on media effectiveness.

Meridian GeoX is the latest open-source framework designed to establish a Google standard for modern causal measurement, allowing advertisers and partners to rigorously optimize media effectiveness and true incrementality of their marketing efforts. By integrating Meridian GeoX with the Meridian Marketing-Mix Model (MMM), advertisers unlock a modern measurement ecosystem that creates a symbiotic, bidirectional causal flywheel. This closed-loop synergy bridges MMM with GeoX validation in two specific ways:

- Experimental Ground Truth (Meridian GeoX → Meridian MMM): Meridian GeoX establishes the empirical "ground truth" used to calibrate Bayesian MMM models through incrementality-based priors (Zhang et al., 2024), reducing bias and uncertainty in MMM's ROI estimates.
- Targeted Experimentation (Meridian MMM → Meridian GeoX): Meridian MMM identifies channels with significant return-on-investment (ROI) uncertainty, guiding advertisers to strategically deploy targeted geo experiment designs where experimental validation of incrementality delivers the highest business value.

Overall, this toolkit supports a seamless end-to-end user journey of geo experiments from designing a test to interpreting the incrementality findings, leveraging advanced methodologies built on solid research. In addition, Meridian GeoX provides advertisers with

crucial flexibility by supporting complex multi-cell testing configurations, allowing them to simultaneously measure the incremental effects of multiple media interventions in one experiment. Lastly, a key advantage of this framework is its ability to evaluate and compare various causal analysis methodologies offered by Meridian GeoX against the users' specific scenario, empowering them to select the most effective strategy given their specific business data, strategic goals, and available budget.

Historically, geo experiments have faced two major hurdles: the high advertising budgets required to overcome large geographical variance—an issue often exacerbated by the presence of disproportionately large, bespoke units (such as New York or London) that lack direct comparable matches, and the failure of conventional inference methods to handle autocorrelated, non-stationary data. Meridian GeoX addresses these pain points directly. First, to minimize variance and lower budget requirements, the framework leverages data-driven stratified sampling paired with Time-Based Regression (TBR) as the primary approach. To handle complex scenarios involving scale mismatches or bespoke units, it offers alternative advanced models like Synthetic Difference-in-Differences (SDID). Second, its Design-Aware Placebo Inference guarantees robust measurement by strictly controlling error rates under non-stationary conditions while maximizing statistical power.

Extensive empirical evaluations validate this framework. Head-to-head comparisons against a competitive industry benchmark demonstrate that Meridian GeoX delivers superior predictive accuracy, achieving lower Minimum Detectable Effects (MDEs) and substantial budget reductions. Furthermore, evaluations against conventional statistical methods confirm that the framework's Design-Aware Placebo Inference provides robust error control under non-stationary conditions.

The rest of the paper is organized as follows:

- Section 2 outlines the core analysis methodology of Meridian GeoX, detailing the Average Treatment Effect on the Treated estimation framework based on counterfactuals and the specific counterfactual modeling options available, including Time-Based Regression (TBR), Synthetic Control (SC), and Synthetic Difference-in-Differences (SDID).
- Section 3 describes the stratified sampling methodology used to create balanced geo assignments, the metrics used to evaluate design quality, and the end-to-end design optimization workflow.
- Section 4 proposes to use Design-Aware Inference to assess the significance level of the incrementality estimator, which is demonstrated to tightly control type-I error rates under various scenarios.
- Section 5 presents the quantitative assessment of the design and analysis module in Meridian GeoX. Section 5.1 focuses on the empirical power analyses and comparisons against the industry benchmark to demonstrate the framework's cost-efficiency and sensitivity. Section 5.2 compares the proposed inference approach in Meridian GeoX against conventional methods to conclude the methodology robustness in both real geo experiments, and under challenging conditions that conventional methods would easily fail.

The implementation of Meridian GeoX in Python is available at the Google GitHub repository (github.com/google/meridian-geox).

2. Causal Analysis

The Meridian GeoX framework estimates causal impact by measuring the Intention-To-Treat (ITT) effect, which corresponds to the Average Treatment Effect on the Treated (ATT) at the geo level. Although compliance is perfect at the geo level (i.e., all assigned treatment geos are treated, making it an ATT at the unit of analysis), individual user-level exposure within those geos cannot be fully controlled (e.g., some users in a treated geo may not see the ads). Therefore, the framework measures the effect of the intention to target a specific user set (an ITT effect from the individual user perspective) by analyzing the ATT at the geo level. This necessitates a robust counterfactual estimation to isolate the incremental effect. Since counterfactual outcomes are inherently unobservable, they are estimated through a two-phase process: the training phase, which models historical relationships between geos, and the experiment phase, which applies these models to predict baseline performance during the experiment period.

To measure this incremental effect, we must compare observed data against a counterfactual, meaning what would have happened without the intervention. Because true counterfactuals are unobservable, Meridian GeoX estimates them using a two-phase process:

- **Pre-Experiment**: Learn the historical relationship between the assigned treatment and control geos.
- **Experiment**: Apply this relationship to the control group's test data to predict the treatment group's counterfactual baseline.

We first define the precise causal estimand. Let $\mathcal{T}_{\text{test}}$ represent the set of time periods t during the active experimental phase. Let $Y_{\text{treat},t}(1)$ and $Y_{\text{treat},t}(0)$ denote the average potential outcomes for the treated geos at time t under the treatment and control conditions, respectively.

The primary estimand of interest is the true cumulative treatment effect per treated geo, defined theoretically as:

$$\tau_{\text{geo}} = \sum_{t \in \mathcal{T}_{\text{test}}} (Y_{\text{treat},t}(1) - Y_{\text{treat},t}(0))$$

Because the counterfactual outcome $Y_{\text{treat},t}(0)$ is inherently unobservable during the test period, it must be estimated. We denote the observed outcome as $Y_{\text{treat},t}$ and the estimated counterfactual baseline as $\hat{Y}_{\text{treat},t}$. The estimator for the cumulative treatment effect per geo is therefore:

$$\hat{\tau}_{\text{geo}} = \sum_{t \in \mathcal{T}_{\text{test}}} (Y_{\text{treat},t} - \hat{Y}_{\text{treat},t})$$

To obtain the total estimated treatment effect across all treated geos which are denoted as set $\mathcal{G}_{\text{treat}}$, this sum is multiplied by the size of treated geos, $|\mathcal{G}_{\text{treat}}|$.

The counterfactual outcome $\hat{Y}_{\text{treat},t}$ is computed differently under respective counterfactual modeling options. The three methods detailed below represent distinct ways to construct that counterfactual baseline. Acknowledging that high budget is a common pain point in geo experiments from an advertiser perspective, these methods are selected not only for their academic rigor but also for their potential to significantly reduce required budgets.

2.1. Time-Based Regression (TBR)

Time-Based Regression (Au, 2018; Kerman et al., 2017) serves as the foundational causal inference methodology within Meridian GeoX. It utilizes a highly interpretable econometric model to leverage historical correlations between treatment and control groups. By fitting a linear regression during the pre-experiment period, the model captures baseline volume differences and structural correlations, allowing for the isolation of exogenous macro-environmental factors such as market-wide economic shifts.

Pre-Experiment: We establish a baseline linear relationship between the cross-sectional average time series of the treatment group and the control group during the pre-experiment period. To be precise, let $\mathcal{G}_{\text{treat}}$ and $\mathcal{G}_{\text{control}}$ denote the sets of geos assigned to the treatment and control pools, respectively, and let $Y_{g,t}$ represent the observed KPI for an individual geo g at time t.

We construct the average time series for each group by aggregating across geos such that each row in our modeling data corresponds to a single time period t:

$$\begin{aligned} Y_{\text{treat},t} &= \frac{1}{|\mathcal{G}_{\text{treat}}|} \sum_{g \in \mathcal{G}_{\text{treat}}} Y_{g,t} \quad \text{and} \\ Y_{\text{control},t} &= \frac{1}{|\mathcal{G}_{\text{control}}|} \sum_{g \in \mathcal{G}_{\text{control}}} Y_{g,t} \end{aligned} \tag{1}$$

Using these aggregated time series, we fit a linear regression model of the following form:

$$Y_{\text{treat},t} = \alpha + \beta \cdot Y_{\text{control},t} + \epsilon_t \tag{2}$$

where:

- α is the intercept, capturing the baseline difference in volume between the treatment and control groups.
- β is the regression coefficient, capturing the structural correlation between the two groups. In practice, we expect $\beta \geq 0$ as treatment and control geos should be positively correlated; this constraint can be enforced during model estimation to prevent negative correlations from confounding the counterfactual.
- ϵ_t represents the residual error term. We impose no assumptions on this term's distribution or independence, relying instead on non-parametric inference (see Section 4) to quantify the uncertainty.

Experiment: During the actual test period, the treatment group receives the intervention while the control group remains unchanged. To isolate the causal impact of this intervention, we use parameters learned during the pre-experiment period to generate a counterfactual, an estimate of what the treatment group *would have produced* had the intervention not occurred.

The predicted counterfactual value at time t during the test period is calculated as:

$$\hat{Y}_{\text{treat},t} = \hat{\alpha} + \hat{\beta} \cdot Y_{\text{control},t} \tag{3}$$

The incremental treatment effect is then derived by taking the difference between the observed $Y_{\text{treat},t}$ and this predicted counterfactual $\hat{Y}_{\text{treat},t}$.

2.2. Synthetic Control (SC)

For scenarios involving data sparsity or high volatility, Meridian GeoX employs Synthetic Control (SC) methodology (Abadie and Gardeazabal, 2003). SC

constructs an optimized, weighted combination of individual control units to mirror the treatment group's pre-experiment behavior.

Pre-Experiment: Let $Y_{\text{treat},t}$ represent the average KPI per geo for the treatment group (this can be either a single treated unit, or an average of multiple treated units) at time t, and $Y_{\text{control},g,t}$ represent the KPI for the g-th control geo (where $g \in \mathcal{G}_{\text{control}}$) at time t. We seek a vector of weights $W = (w_1, w_2, \ldots, w_{|\mathcal{G}_{\text{control}}|})$ by solving the following constrained optimization problem over the pre-test period:

$$\min_{w_1,\ldots,w_{|\mathcal{G}_{\text{control}}|}} \sum_{t\in\mathcal{T}_{\text{pre}}} \left(Y_{\text{treat},t} - \sum_{g\in\mathcal{G}_{\text{control}}} w_g \cdot Y_{\text{control},g,t} \right)^2 \tag{4}$$

Subject to the standard synthetic control constraints:

- Non-negativity constraint: $w_g \geq 0$ for all g.
- Sum-to-one constraint:

$$\sum_{g\in\mathcal{G}_{\text{control}}} w_g = 1 \tag{5}$$

These constraints act as a regularization mechanism. By restricting the weights to a convex hull, SC strictly bounds the model to interpolation rather than extrapolation, shielding the inference from being disproportionately skewed by outliers or extreme volatility in sparse datasets. The result is a learned vector of optimal weights.

Experiment: During the test period, we apply the learned weights to the control geos' observed data:

$$\hat{Y}_{\text{treat},t} = \sum_{g\in\mathcal{G}_{\text{control}}} \hat{w}_g \cdot Y_{\text{control},g,t} \tag{6}$$

2.3. Synthetic Difference-in-Differences (SDID)

SDID (Arkhangelsky et al., 2021) is the most advanced econometric model under the Meridian GeoX framework, hybridizing SC unit-weighting with Difference-in-Differences (DiD) correction. This "doubly robust" estimator mirrors parallel trends despite baseline volume differences and introduces time weights to stabilize historical volatility.

Pre-Experiment: The algorithm calculates:

- **Unit weights** w_g: Similar to SC but additionally includes an intercept α to match the treatment group's shape without requiring identical baseline volumes.
- **Time weights** λ_t: Upweight historical periods structurally similar to the experiment period, further denoising the time series. We only allow non-negative time weights $\lambda_t \geq 0$.

Experiment: The SDID incremental lift consists of a raw estimate minus a correction term based on the pre-experiment period:

$$\begin{aligned}\text{Treatment Effect} = &\sum_{t\in\mathcal{T}_{\text{test}}} \left(Y_{\text{treat},t} - \hat{Y}_{\text{synth},t} \right) \\ &- |\mathcal{T}_{\text{test}}| \times \left(\sum_{t\in\mathcal{T}_{\text{pre}}} \hat{\lambda}_t \cdot Y_{\text{treat},t} - \sum_{t\in\mathcal{T}_{\text{pre}}} \hat{\lambda}_t \cdot \hat{Y}_{\text{synth},t} \right)\end{aligned} \tag{7}$$

where $|\mathcal{T}_{\text{test}}|$ is the duration of experiment period and

$$\hat{Y}_{\text{synth},t} = \sum_{g\in\mathcal{G}_{\text{control}}} \hat{w}_g \cdot Y_{\text{control},g,t} \tag{8}$$

2.4. Convergence Properties

The three counterfactual modeling options in Meridian GeoX rely on distinct frameworks to justify their large-sample estimation accuracy:

- Time-Based Regression (TBR): Consistency is established in the parameter space. As the historical pre-experiment training period ($\mathcal{T}_{\text{train}}$) grows large, the uncertainty of the regression coefficients (α, β) converges to zero at a rate of $1/\sqrt{|\mathcal{T}_{\text{train}}|}$ (Kerman et al., 2017). Although the short duration of a typical test means the final treatment effect estimate will still reflect day-to-day market fluctuations, the underlying TBR model parameters are estimated with near-perfect precision. This ensures a highly stable baseline prediction, where the remaining uncertainty represents only the expected, natural noise of the test window.
- Synthetic Control (SC): Asymptotic properties focus on bias reduction. The estimator's bias approaches zero as the length of pre-experiment period grows large, ensuring that on average, the weighted control combination converges to the unobserved baseline (Abadie et al., 2010).
- Synthetic Difference-in-Differences (SDID): Arkhangelsky et al. (2021) show that the SDID estimator is consistent (converging to the true treatment effect) and asymptotically normal as both the number of control geos and the length of pre-experiment period grow large, justifying its use for robust large-sample inference.

3. Design

Design is crucial to ensure an effective geo experiment. The design generation and selection process in Meridian GeoX involves generating a pool of candidate geo assignments by stratified sampling, subsequently

filtering candidates by quality and optimizing the minimum detectable effect. This framework supports not only single (treatment) cell but also multi-cell geo experiments, which is designed for a wide range of practical use cases.

3.1. Stratified Sampling

Within the framework of large-scale experimental design, randomized sampling is frequently employed as a baseline approach. By selecting treatment and control groups completely at random, randomized sampling leverages the "Law of Large Numbers" to average out inherent differences across a massive number of geos. However, this stochastic process suffers from significant variance when sample sizes are restricted, and it typically requires more budget allocations and a higher volume of regions to attain statistical significance.

To mitigate the inherent challenges of large variance and high budget requirements, Meridian GeoX prioritizes a stratified sampling methodology at the design stage. This approach partitions geos into "Strata"—homogeneous clusters defined by shared characteristics such as sales volume or seasonality—and subsequently performs randomized assignment of treatment and control groups within these matched cohorts while strictly adhering to specified design constraints.

By establishing pre-experimental comparability between treatment and control groups, stratified sampling design effectively minimizes variance and yields more balanced and reliable geo assignment. This reduction in noise facilitates the detection of smaller lifts with lowered budget requirements.

The detailed implementation is as follows. We featurize the KPI time series Y_t of length $|\mathcal{T}_{train}|$ which is the length of training period (see definition in Section 3.3) by computing:

- **Mean**: The average of Y_t,

$$\bar{Y} = \frac{1}{|\mathcal{T}_{train}|} \sum_{t \in \mathcal{T}_{train}} Y_t$$

- **Coefficient of Variation (CV)**: The ratio of the standard deviation to the mean, providing a normalized measure of dispersion within Y_t,

$$CV = \frac{1}{\bar{Y}} \sqrt{\frac{1}{|\mathcal{T}_{train}| - 1} \sum_{t \in \mathcal{T}_{train}} (Y_t - \bar{Y})^2}$$

- **Slope**: The linear rate of change observed within Y_t,

$$\beta_{trend} = \frac{\sum_{t \in \mathcal{T}_{train}} (t - \bar{t})(Y_t - \bar{Y})}{\sum_{t \in \mathcal{T}_{train}} (t - \bar{t})^2}$$

- **Autocorrelation**: The degree of correlation between Y_t and a version shifted by a specific lag k (with a default $k = 1$),

$$r_1 = \frac{\sum_{t=1}^{|\mathcal{T}_{train}|-1} (Y_t - \bar{Y})(Y_{t+1} - \bar{Y})}{\sum_{t \in \mathcal{T}_{train}} (Y_t - \bar{Y})^2}$$

- **Dynamic Time Warping (DTW; Kate (2015))**: A distance-based similarity metric used to measure how the temporal response pattern of an individual geo unit deviates from the overall average response pattern.

We then apply a K-means clustering algorithm based on these metrics to partition geo units into multiple homogeneous strata. Following this stratification, within-strata randomized sampling is used to assign geo units to treatment and control groups, ensuring that each arm serves as a statistically representative subset of the entire population while maintaining balance across stratum sizes.

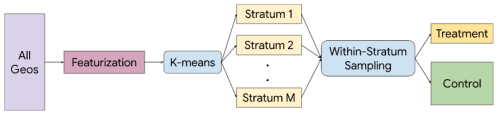

Figure 1 | Use of stratified sampling to generate initial design candidates

3.2. Evaluation Metrics

In Meridian GeoX, design quality is evaluated using metrics as detailed in the table below.

An optimal design should pass both out-of-sample quality and AA quality checks, and minimize the MDE metric. Finally, Meridian GeoX evaluates the budget requirement on each optimal design depending on the experiment options:

- When testing on existing campaigns, budget requirements are derived from historical spend by treatment geos.
- For advertisers who want to test new media activations, the required budget can be derived as:

$$\text{Required Budget} = MDE_{\text{lift}} \times \text{Baseline} \times \text{Cost per Incremental Response}$$

where MDE_{lift} is the minimum detectable effect (as a fraction) and Baseline is the Response by Treatment.

Metric	Description	Technical Details
Out-of-Sample Quality	Filter out designs for which counterfactual modeling lacks fit.	$R^2 = 1 - \frac{\sum_{t\in\mathcal{T}_{val}}(Y_t-\hat{Y}_t)^2}{\sum_{t\in\mathcal{T}_{val}}(Y_t-\bar{Y})^2}$ *Where:* Y is the observed validation value and $\hat{Y}$ is the counterfactual prediction.
AA quality	Filter out designs with significant AA imbalance.	Given a significance level α, the p-value of each design when conducting an A/A validation. Consider test statistic $Z = \frac{\hat{\tau}}{SE(\hat{\tau})}$, should follow standard normal distribution without any AA imbalance.
Minimum Detectable Effect (MDE)	Optimize designs for high sensitivity to identify incremental effects.	$MDE_{\text{lift}} = \frac{SE_{\text{lift}} \times (z_{1-\alpha/2} + z_{1-\beta})}{\text{Baseline}}$ *Where:* Baseline is response by treatment, SE_{lift} is the standard error of incremental response, and $1-\beta$ is target statistical power.

Table 1 | Evaluation metrics in the design process

3.3. Filtering and Optimization

The optimal designs are selected after filtering among design candidates and optimizing for MDE in evaluations. In this process, we separate the pre-experiment data into three periods:

- **Training period**: Used for data featurization, stratified sampling, and counterfactual model training.
- **Validation period**: Used for checking out-of-sample fit of the counterfactual model, and filtering top designs.
- **Evaluation period**: Used for further filtering designs that pass A/A quality check, and finalizing optimal designs after evaluating MDE and budget.

To maintain alignment with the duration of experiment $|\mathcal{T}_{\text{test}}|$, which is specified by the user, MDE and budget calculations for the last evaluation period are performed over an identical timeframe. The same duration requirement applies to the second validation period. It is also advised that the training period utilizes at least $|\mathcal{T}_{\text{test}}|$ of data for model training. Consequently, a minimum $3 \times |\mathcal{T}_{\text{test}}|$ pre-experiment data in total is considered a baseline requirement for effectively designing a geo experiment, although longer period of pre-experiment data is recommended if significant seasonality or cyclical patterns exist in the data.

The entire workflow for data preprocessing and the search for optimal design within the Meridian GeoX framework is illustrated in the Figure 2.

4. Quantifying Uncertainty: Design-Aware Placebo Inference

Accurately quantifying uncertainty is just as critical as estimating the causal effect in geo experiments. Conventional parametric inference methods, such as standard t-tests, assume independent observations. However, real-world data violates this assumption; it is heavily influenced by autocorrelation (serial dependence), persistent trends, and structural breaks caused by exogenous macroeconomic shocks. Applying standard parametric tests to such non-stationary data systematically underestimates uncertainty and artificially inflates the False Positive Rate (FPR).

Meridian GeoX overcomes this by utilizing placebo inference. Rather than relying on rigid parametric assumptions, placebo inference empirically establishes a null distribution of the expected error by creating hypothetical treatment groups out of the untreated

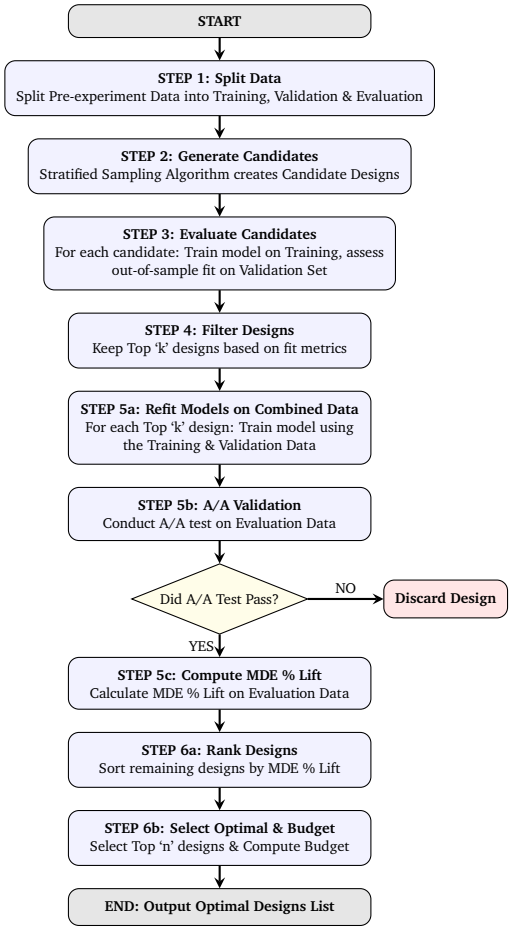

Figure 2 | Design Optimization Workflow

control units. Because the placebo groups are exposed to the same temporal dynamics and shared structural shocks as the actual treatment group, this method naturally controls for real-world data volatility.

4.1. The Core Mechanics of Placebo Inference

Placebo inference (Abadie et al., 2010; Firpo and Possebom, 2018) relies on creating a treatment group that is comparable to the actual treated but built from the pool of control units. Because the treatment effect should not affect the control units, this treated placebo group will not be affected by the treatment. For inference we build a distribution of many of these placebo controls to capture what would have happened if no treatment would ever take place (the null distribution).

The following plot illustrates this process. Each column represents a time period and each row a unit. Blue units will be treated during the treatment period and switch from light to dark blue. Gray units are never treated and serve as the pool of possible controls. Yellow squares represent the sampled placebo treatments that are used to calculate the null distribution of the residuals.

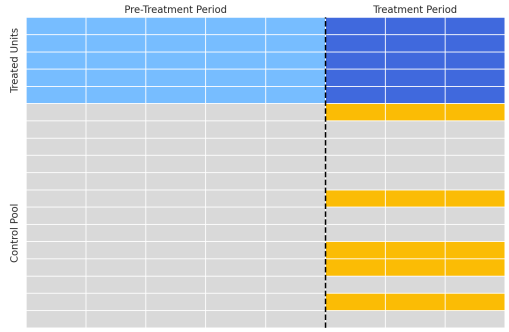

Figure 3 | Illustration of a single placebo effect calculation.

At a high level, we construct the null distribution through N_{plac} iterations of placebo resampling (without replacement). For each iteration, we draw a different placebo treatment assignment (as illustrated below) and calculate its resulting effect.

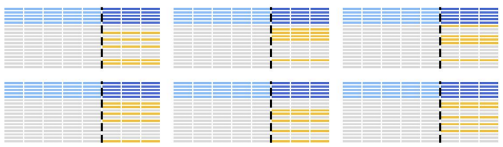

Figure 4 | Multiple placebo assignments illustrated.

Then we calculate the placebo treatment effect using the placebo-treated units as treatment and the

remaining control geo units as control group to create a placebo-counterfactual.

The choice of N_{plac} matters for the granularity of reporting. We recommend choosing $N_{plac} \geq 100$ so the p-values can be differentiated at the 1% level, while $N_{plac} = 1000$ allows differentiation at the 0.1% level. For each draw we calculate a hypothetical (placebo) effect. The distribution of these effects provides us with a distribution of the expected error.

For this distribution to act as a valid null distribution, the placebo treatment effects need to be exchangeable with the actual treatment effect. Because different geos naturally possess different levels of baseline variance, comparing raw error magnitudes directly would most likely invalidate the test. We therefore provide two refinements of the placebo inference: Design-Aware Selection and Studentization.

4.2. Random vs Design-Aware Selection

Random Placebo Inference selects the placebo-treated units uniformly at random from the control pool, assuming each geo unit has an equal chance of being treated. However, as we often utilize specific sampling to deliberately balance the design and minimize variance, this assumption of uniform probability is violated in reality.

Design-Aware Placebo improves upon this selection by mirroring the assignment rules to generate placebos. By following the exact same selection process for the placebo treatment units as was used for the actual treatment units, we ensure the generated null distribution respects the variance-reducing properties of the original experimental design.

4.3. Studentization: Standardizing the Effect

To account for the variance mismatch between the placebo treatments and the actual treatment, we transform the estimate into a pivotal statistic by standardizing the effect. We achieve this by dividing the estimated effect by its pre-period Root Mean Squared Error (RMSE), a process known as studentization (Chung and Romano, 2013). For either actual or placebo design, an out-of-sample RMSE is calculated on the evaluation period that is not used in the R^2-score-based design selection process. With $Y_{treat,t}$ being the average KPI of treatment group of a certain design over the evaluation period $\mathcal{T}_{eval}$:

$$RMSE = \sqrt{\frac{1}{|\mathcal{T}_{eval}|} \sum_{t \in \mathcal{T}_{eval}} (Y_{treat,t} - \hat{Y}_{treat,t})^2}. \quad (9)$$

This scales the magnitude of each placebo effect relative to its own historical noise, allowing for a valid, "apples-to-apples" comparison to the actual design when constructing the null distribution.

4.4. Deriving Statistical Metrics

P-value

The P-values are calculated empirically by comparing the studentized effect of the actual treatment against the distribution of studentized placebo effects:

1. Calculate the observed test statistic by "studentizing" the actual treatment effect:

$$t_{obs} = \frac{\text{effect}_{obs}}{RMSE_{obs}} \quad (10)$$

2. Calculate the placebo test statistics for the i-th placebo by studentizing every placebo effect:

$$t_{plac,i} = \frac{\text{effect}_{plac,i}}{RMSE_{plac,i}} \quad \text{for } i = 1, \ldots, N_{plac} \quad (11)$$

3. Calculate the empirical p-value by counting how many placebos more extreme than the observed one:

$$p\text{-value} = \frac{1 + \sum_{i=1}^{N_{plac}} I(t_{plac,i} \text{ more extreme than } t_{obs})}{1 + N_{plac}} \quad (12)$$

Standard Error (SE)

The SE is based on placebo distribution. First, calculate the standard deviation of the studentized placebo distribution, σ_{plac}. Then, "un-studentize" it back by multiplying that standard deviation by the RMSE of the observed noise treatment effect ($RMSE_{obs}$):

$$SE = \sigma_{plac} \times RMSE_{obs} \quad (13)$$

Confidence Interval (CI)

The CI is calculated as follows:

1. Find the empirical quantiles of the studentized placebo distribution based on significance level α: $q_{\alpha/2}$ and $q_{1-\alpha/2}$.
2. Scale the quantile by multiplying them with the observed (treatment) RMSE.
3. Subtract the upper scaled quantile from the observed effect to establish the lower bound, and subtract the lower scaled quantile to establish the upper bound:

$$CI_{lower} = \text{effect}_{obs} - (q_{1-\alpha/2} \times RMSE_{obs}) \quad (14)$$

$$CI_{upper} = \text{effect}_{obs} - (q_{\alpha/2} \times RMSE_{obs}). \quad (15)$$

5. Evaluations

5.1. Empirical Power Analysis

5.1.1. Design under Different Meridian GeoX Methodologies

Our evaluation first assesses the empirical power of different analysis methodologies, including TBR, SC, and SDID, while the proposed stratified sampling is compared against the baseline randomized sampling strategy in design. Because design algorithms and inference methodologies are closely linked during both the selection and analysis phases, we examine all 6 methodologies that represent all combinations of these design sampling algorithms and analysis options available in Meridian GeoX.

For the purpose of this evaluation, we simulate both single-cell and two-cell geo experiments based on recent Google internal experiments from anonymous advertisers. We assume each treatment cell in the simulated geo experiment takes 30% of national response share if it is a single-cell, or 20% of national response share if it is a two-cell experiment. All studies are set to a 90% confidence level and a target statistical power of 80%. For each combination of sampling algorithm and analysis methodology, we start with simulating 1000 candidate designs. In the end, the optimal design per experiment per methodology is selected based on the minimum MDE % lift. We then compared these final designs across all methodologies to evaluate their relative MDE gains, defined by

$$\text{relative MDE gain} = (\text{baseline MDE} - \text{MDE}) / \text{baseline MDE} \times 100\% \quad (16)$$

The combination of random sampling and SC is picked as the baseline in the Formula 16. For the two-cell experiments, the relative MDE gain is evaluated per treatment cell.

The budget requirements per methodology are computed and compared accordingly. To simulate the fiscal requirements for conducting these experiments using the budget metrics derived from these study designs, we assume simultaneously a unit cost per incremental response (specifically, it is posited that a single unit of ad spend yields a corresponding unit of incremental response) for all advertisers in this dataset. Because the actual cost per incremental response data is proprietary to the advertiser, it was excluded from our analysis. In practice, the expected cost per incremental response varies significantly across advertisers and different media channels. Therefore, the budget metrics presented in the following evaluations should be interpreted only in a relative sense rather than as absolute values.

The evaluation results suggest that:

- **Superiority of Stratified Sampling**
  Stratified sampling demonstrates a consistent performance advantage over randomized sampling across the entire spectrum of analysis methodologies. When controlling for the analysis method, stratified sampling based designs consistently produce more gain in lowering MDE and a higher probability of minimizing the budget requirement, as shown in Figure 5.

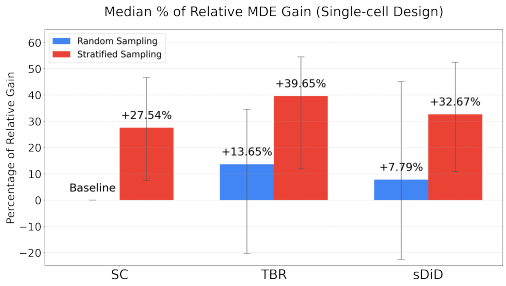

Figure 5 | Relative MDE gain across methodologies under a single-cell setting

- **Best Overall, Stratified Sampling + TBR**
  Stratified sampling paired with TBR emerges as the optimal methodology compared to other methodologies. This combination achieves the maximum relative MDE gain while concurrently maximizing the likelihood of identifying a design that minimizes budget requirement, as demonstrated in Figure 6.

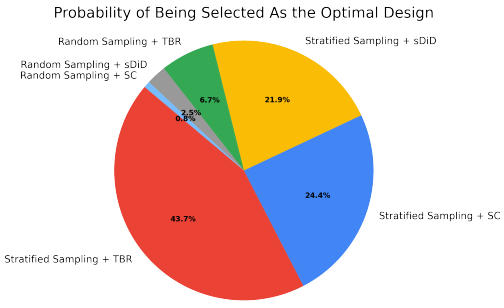

Figure 6 | Probability of being selected as the optimal design across different sampling and analysis methodologies

- **Consistent Findings Between Single- and Multi-Cell**
  The performance hierarchy remains uniform across single-cell and multi-cell setups. This stability proves that the advantages of stratified sampling over random sampling as well as the superior efficiency of stratified sampling + TBR are

robust.

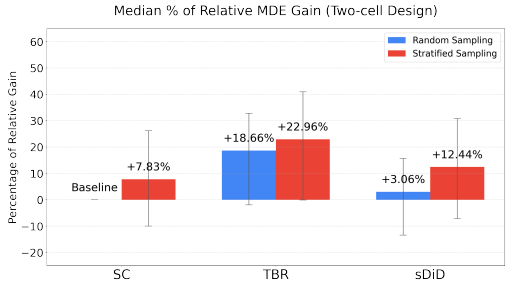

Figure 7 | Relative MDE gain across methodologies under a two-cell setting

- **Resilience of SC and SDID**
  While stratified sampling + TBR is the overall winner among all methodologies, stratified sampling + SC or SDID can outperform under specific data conditions—such as when the geo pool contains outsized, bespoke units or when the time series are highly volatile. In scenarios with bespoke units, TBR's uniform aggregation of geos could fail to adjust for baseline scale and trend mismatches. In contrast, the unit-weighting approach of SC and SDID provides the flexibility to construct a stable, customized counterfactual match for these dominant regions. In scenarios where data is very noisy, the unit-weighting approach of SC and SDID similarly provides enhanced stability compared to the uniform aggregation employed by TBR. Our empirical analysis of these internal studies quantifies this relationship by measuring the coefficient of variation across response time series. As illustrated in Figure 8, while TBR remains highly effective under low-to-medium volatility, its relative efficiency diminishes in high-noise environments, where SC achieves more significant improvements in design quality.

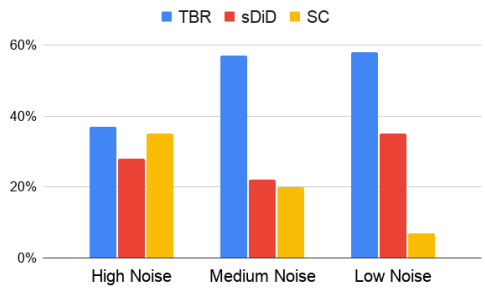

Figure 8 | Impact of varying noise levels in study data on the probability of a design being identified as optimal

5.1.2. Head-to-head Comparison with Industry Open-Source Benchmark

To evaluate the performance of Meridian GeoX, we conducted a rigorous comparison against a prominent open-source library that serves as an industry benchmark for geo experiments. This benchmark library—which utilizes augmented synthetic control (Ben-Michael et al., 2021) as its core methodology—is widely recognized for its ability to function with relatively low budget requirements. It achieves this by optimizing unit selection and test duration to maintain statistical power even with smaller sample sizes.

Our goal was to demonstrate the competitive advantages of Meridian GeoX through empirical data. The results show that Meridian GeoX facilitates a more powerful experimental design, further reducing budget requirements while maintaining the necessary statistical rigor.

To ensure an objective and fair comparison, we structured the evaluation as follows:

- **Benchmark Library Configuration**:
  - Treatment Allocation: Per study, 20% of the total geo pool.
  - Statistical Parameters: A 90% significance level and an 80% target statistical power.
  - Defaults: All other parameters were kept at their standard default configurations.
- **Meridian GeoX Alignment**: To guarantee a fair comparison, geo experiments were initialized using constraints derived directly from the benchmark's outputs. Specifically, we matched the percentage of response-by-treatment, the significance level, and the target statistical power for every study analyzed. By neutralizing these variables, we isolated the efficiency of the underlying experimental design and analysis engines, allowing for a direct assessment of how each tool manages budget and power.

Across all studies, we evaluated the MDE achieved by the optimal designs selected by both tools and measured the weekly required budget associated with those specific designs to determine cost-efficiency.

Superior Predictive Accuracy (RMSE)

The out-of-sample Root Mean Squared Error (RMSE), calculated over the evaluation period as defined in Equation 9, measures the prediction accuracy between the treated unit and its corresponding counterfactual predictions in the evaluation period, directly informing the MDE. A lower pre-treatment RMSE (better fit) ensures a more reliable counterfactual, reducing

the variance and lowering the threshold for the MDE, allowing for smaller incrementality effects to be detected. Therefore, RMSE serves as the primary metric for assessing predictive precision across the tools in our head-to-head comparison.

The results indicate that Meridian GeoX provides a superior RMSE across the entire advertiser cohort, with these advantages becoming even more pronounced among high-spend segments which consist of Enterprise advertisers.

- **Overall Performance:** Meridian GeoX demonstrated superior predictive accuracy in 73% of all studies. Furthermore, we observed a median 71% reduction in RMSE (Q1 and Q3, or the first and third quartiles: 25%, 87%) compared to the benchmark.
- **Superiority for Large Spenders:** The efficiency gains are more significant among Enterprise advertisers (n = 40), where 80% of studies achieved a more optimal fit using Meridian GeoX, resulting in a median 80% reduction in RMSE.

Matched or Improved Statistical Power (MDE)

The results demonstrate that Meridian GeoX provides designs with the same level of statistical power as the benchmark, while providing a distinct competitive advantage for large-spend advertisers:

- **Overall Performance:** Meridian GeoX achieves the same level of statistical power as the benchmark, with no statistically significant difference observed between the two tools. For the entire advertiser cohort, Meridian GeoX maintained a median MDE of 6% (Q1 and Q3: 5%, 12%), effectively matching the benchmark's median of 6% (Q1 and Q3: 5%, 10%).
- **Superiority for Large Spenders:** The superiority in statistical power over the benchmark becomes pronounced for large-spend advertisers. Among Enterprise advertisers, 65% of studies achieved a more optimal MDE using Meridian GeoX, with a median 35% reduction in MDE, as shown in Figure 9.

Note that it is common for geo experiment libraries (including the open-source benchmark evaluated here) to report MDEs based on the single best-performing trial from a large candidate pool, which can introduce optimistic selection bias. In contrast, Meridian GeoX employs a simulation-calibrated framework to provide expected MDEs. Consequently, because Meridian GeoX's MDE is consistently more conservative, matching or exceeding the benchmark's nominal MDE represents a highly rigorous comparison of true, replicable experimental power.

Advertiser Classification	% Meridian GeoX improves MDE	% Meridian GeoX improves budget
All (n = 69)	51% ($p = 0.50$)	64% ($p = 0.01^{**}$)
Enterprise (n = 40)	65% ($p = 0.08^{*}$)	83% ($p = 0.01^{**}$)
SMB (n = 29)	31% ($p = 0.97$)	34% ($p = 0.93$)

Table 2 | MDE/budget comparison across Enterprise and Small and Medium Business (SMB) advertisers vs. the benchmark

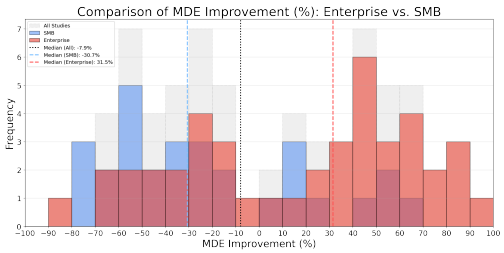

Figure 9 | Distribution of MDE improvement from Meridian GeoX vs. the benchmark across Enterprise/SMB advertisers

Significant Budget Reduction

We compare the required budget metrics of generated designs using both Meridian GeoX and the open-source benchmark in Table 3. Once again, these numbers should only be interpreted in a relative sense rather than regarded as absolute values due to the evaluation setups mentioned earlier. When compared specifically to the open-source benchmark, Meridian GeoX offers a more significant advantage:

	Required Budget / Week		
Tool	**Q1**	**Median**	**Q3**
Meridian GeoX	5k	23k	70k
Benchmark	4k	28k	85k

Table 3 | Weekly required budget comparison: Meridian GeoX vs. Benchmark

- **Broad Budget Reduction:** 64% of all studies required a lower weekly budget when using Meridian GeoX compared to the benchmark. On average, while the benchmark requires a median budget of 28k, Meridian GeoX requires just 23k, representing a 17% reduction in budget.
- **Drastic Savings for Large Spenders:** For Enterprise advertisers, the efficiency gap widens significantly. In this segment, we observe 83% of studies reduced required budget when moving

from the benchmark to Meridian GeoX, with a per-study median 31% reduction in budget requirement.

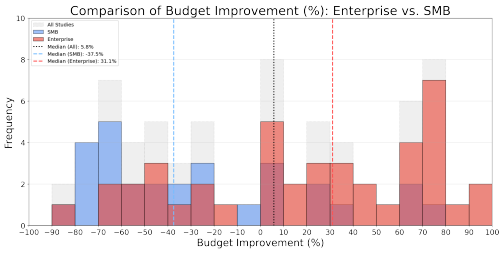

Figure 10 | Distribution of budget improvement from Meridian GeoX vs. Benchmark across Enterprise/SMB advertisers

5.2. A/A Robustness and Statistical Power of Design-Aware Placebo Inference

We conducted large-scale A/A testing and measured the study-level False Positive Rate (FPR) and statistical power metric to evaluate the proposed Design-Aware Placebo Inference. An ideal inference method should precisely quantify the actual uncertainty level in the incrementality metrics; hence it is expected to properly control FPR under the nominal level while achieving the highest power.

We compare Design-Aware Placebo Inference against 3 conventional methods: t-test, sliding window method (Chernozhukov et al., 2018), and random placebo inference which are commonly used by other geo experiment solutions.

Our evaluations utilized synthetic geo-level response data from 192 distinct anonymous advertisers from Google Ads. For each study, we simulated an artificial experiment window, preceded by pre-experiment periods to find an optimal design and validate significance of their A/A bias respectively during the last segment of the pre-experiment period. For each of the 192 studies, we ran 200 independent A/A test iterations (resulting in 38,400 total simulated experiments).

Because no artificial lift was added in A/A tests, the true causal effect is exactly zero. The empirical A/A variance provides the exact Standard Error needed to analytically calculate statistical power, making the injection of artificial lift unnecessary.

5.2.1. Rigorously Controlling False Positive Rate in A/A Tests

The results demonstrate that only Design-Aware Inference successfully passed the A/A check, rigorously quantifying the uncertainty. It outperforms all baselines by mirroring the actual stratified assignment

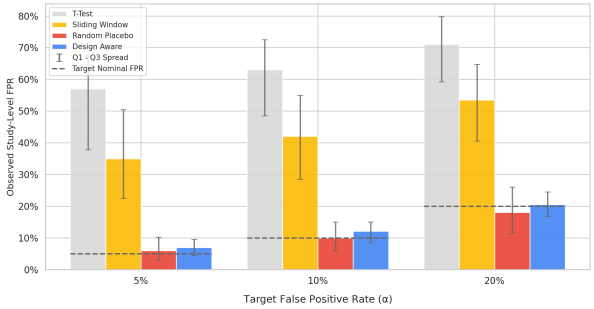

Figure 11 | The median study-level FPR across different nominal levels. The error bars represent interquartile range, the Q1 to Q3 spread.

rules to generate placebos, ensuring the standard errors are calibrated to the specific variance of each experiment. Its median FPR lands exactly on the nominal target line at every confidence level, and its narrow Q1-Q3 spread proves it delivers the most robust uncertainty estimates across all studies.

In contrast, conventional inference methods (t-test and sliding window method) resulted in significant inflation of observed FPRs. At a 5% nominal level, the t-test performs worst, showing over 10x observed FPRs than expected, which leads to substantial A/A bias and unreliable incrementality results. Section 5.2.2 provides a detailed analysis of why these methods fail the A/A check.

Although Random Placebo controls median FPR across studies, it fails to maintain reliable control for individual studies, a requirement only met by Design-Aware Placebo. Random Placebo assignment is purely random while in the design process the actual treatment assignment was optimized to reduce variance; this method generates an excessively wide null distribution. As illustrated in the plot, although its median is near the nominal level, the wide Q1-Q3 spread indicates it frequently yields severe false positives or overly conservative estimates depending on the specific study. Furthermore, Section 5.2.3 highlights how Design-Aware Placebo provides superior statistical power compared to the Random Placebo.

5.2.2. Robustness to Autocorrelation, Structural Breaks and Non-Stationarity

This section investigates why conventional, non-placebo inference methods frequently fail to maintain valid FPRs during A/A testing. Most parametric approaches, including standard t-tests, assume independent observations in daily time series which is rarely true. This assumption is frequently violated due to the presence of serial dependence, or autocorrelation, wherein successive observations are highly correlated.

Consequently, the use of t-tests on autocorrelated data systematically inflates the FPR.

This is empirically demonstrated in Figure 12, which simulates time series data utilizing a lag-one autoregressive model (ACF(1) = 0.5), illustrating that t-tests yield inflated error rates while sliding window and placebo methods maintain validity.

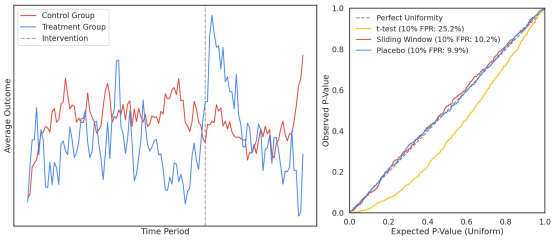

Figure 12 | In autocorrelated time series data that follows AR(1) model, standard t-tests yield inflated false positive rates, whereas sliding window and placebo methods successfully maintain valid error rates.

While the sliding window method has some resilience to time-series autocorrelation, it relies on an assumption of stationarity that is also frequently violated in reality. In contrast, placebo tests remain highly robust under these conditions, as they do not require strict assumptions regarding independence or stationarity.

Structural Breaks—commonly seen in geo experiments as a result of measurement shifts or exogenous macroeconomic shocks—pose a significant challenge for conventional inference methods. The sliding window test, which relies on temporal permutation to construct null distributions, generates overconfident estimates when the underlying data-generating process undergoes fundamental shifts, rendering pre- and post-experiment periods strictly incomparable.

Placebo inference (both random and design-aware) solves this problem. As long as a structural break is unrelated to the treatment—thereby symmetrically affecting both the treatment and control units—the inference remains robust. To show this, we simulated an A/A test incorporating an exogenous structural break. Under the null hypothesis of no treatment effect, a valid statistical test must yield a uniform p-value distribution. As depicted in Figure 13, only placebo inference produces the expected uniform distribution. Both the t-test and the sliding window method yield spurious significance. This discrepancy arises because conventional methods primarily compare the treatment group's experiment period data against its own historical baseline, whereas placebo methods evaluate the treatment group relative to a concurrent control group, thereby controlling for the shared structural shock.

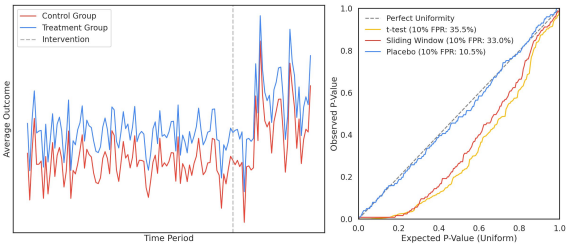

Figure 13 | Under a structural break, only placebo inference results in the expected uniform distribution. Both the sliding window and t-test create the illusion of an effect where none exists and causes significant inflation in false positive rates.

A similar vulnerability exists when the data exhibit a persistent non-stationary trend rather than simple autocorrelation. A trend steadily drives the numbers higher (or lower) than anything seen before the treatment, guaranteeing that any direct comparison to the past will fail. To demonstrate this, another A/A test was simulated on data characterized by an underlying trend. Consistent with the previous findings, Figure 14 indicates that placebo inference successfully accommodates the underlying trend, while the standard t-test and sliding window methods severely underestimate uncertainty.

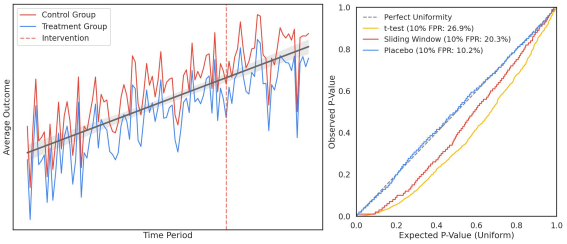

Figure 14 | Under a persistent trend effect, only placebo inference results in the valid control of false positive rates. Both the sliding window and t-test end up being overconfident.

Non-stationarity, in the form of structural breaks and trends, is a pervasive feature of time series data, often driven by pronounced seasonality, major promotional events (e.g., Black Friday), and cyclical response patterns. These factors significantly confound the identification of true causal effects. Instead of designing complex experiments around these seasonal shifts, placebo inference (random or design-aware) allows users to measure uncertainty consistently.

5.2.3. Improving Statistical Power Over Random Placebo

While properly controlling the False Positive Rate ensures the accuracy of an inference method, the method must also be sensitive enough to detect the actual treat-

ment effect. To evaluate this sensitivity, we calculate MDE across all simulated studies, utilizing the standard errors derived from both Random and Design-Aware Placebos.

Figure 15 compares the global Mean and Median MDE between the two approaches:

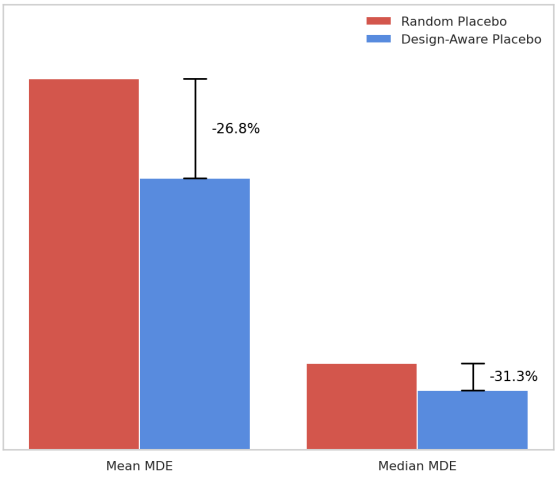

Figure 15 | Comparison of global Mean and Median MDE between Random Placebo and Design-Aware Placebo.

- **Random Placebo:** By ignoring the variance-reducing properties of the design, the Random Placebo method overestimates the standard error. This results in a higher baseline MDE, meaning the test is more conservative and requires a larger actual treatment effect to register as statistically significant.
- **Design-Aware Placebo:** By properly accounting for the experimental design, this method yields a tighter standard error distribution, directly pulling the MDE boundaries inward.

Implementing the Design-Aware Placebo reduces the Mean MDE by 26.8% and the Median MDE by 31.3% compared to the Random approach. This demonstrates that properly accounting for the experimental design not only guarantees statistical validity, but actively maximizes our sensitivity, allowing us to confidently detect much smaller lifts.

6. Conclusion

The Meridian GeoX framework represents a significant advancement in the field of geo experiments to causally measure the effectiveness of digital advertising.

Meridian GeoX offers great flexibility for users by integrating multiple methodologies into a single cohesive tool, namely Time-Based Regression, Synthetic Control, and Synthetic Difference-in-Differences. Crucially, the framework evaluates and compares these methodologies against the user's specific scenario, helping the user to select the optimal approach for unique business data, objectives, and budget constraints.

This framework provides advertisers and marketing agencies with a robust, open-source toolkit for quantifying the true incremental impact of their marketing efforts. In addition, pairing these methodologies with the stratified sampling based designs offers great power improvement and significant budget reduction compared to traditional random sampling based approaches. Our evaluations demonstrate that Meridian GeoX not only matches the testing accuracy of industry benchmarks but often exceeds them. The framework's ability to achieve significant budget reductions—such as the median 31% budget reduction observed for Enterprise advertisers—while maintaining matched or improved statistical power (MDE) offers a clear path toward more efficient and sustainable experimentation. Ultimately, Meridian GeoX translates this statistical efficiency into true business ROI, allowing advertisers to make more informed budget allocations and maximize total media effectiveness to drive tangible business outcomes.

The introduction of Design-Aware Placebo Inference addresses a critical gap in quantifying uncertainty within geo experiments. This approach ensures that statistical power and confidence intervals are grounded in the specific nuances of the experiment's design, leading to more reliable decision-making. Our evaluations confirm that Meridian GeoX is methodologically robust particularly in real experiment scenarios involving autocorrelation, non-stationary time series or structural breaks.

As the digital landscape continues to evolve toward more privacy-safe measurement, tools like Meridian GeoX are essential for executing a complete modern measurement strategy. The launch of Meridian GeoX will foster a more transparent and collaborative ecosystem, ultimately empowering world-wide advertisers to make data-driven decisions with greater precision and confidence.

Acknowledgements

We would like to thank Jing Wang, Lisa Jakobovits, Yonathan Schwarzkopf, and Harikesh Nair for their encouragement and support throughout this research; Lynn Xie and Niall Dennehy for their insightful comments and suggestions; Dirk Nachbar, Sam Bailey, Wincy Liu, and our colleagues in Google gTech Ads

for valuable discussions and feedback; and Emma Lin, T.C. Cohen, and our engineering partners for their support in implementing and scaling the framework.

References

A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113–132, 2003. doi: 10.1257/000282803321455188.

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.

D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12): 4088–4118, 2021. doi: 10.1257/aer.20190159.

T. Au. A time-based regression matched markets approach for designing geo experiments. Technical report, Google LLC, 2018.

E. Ben-Michael, A. Feller, and J. Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536): 1789–1803, 2021. doi: 10.1080/01621459.2021.1929245. URL https://doi.org/10.1080/01621459.2021.1929245.

V. Chernozhukov, K. Wuthrich, and Y. Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data, 2018. URL https://arxiv.org/abs/1802.06300.

E. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013. doi: 10.1214/13-AOS1094.

S. Firpo and V. Possebom. Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2016-0026.

R. J. Kate. Using DTW distance as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2):283–312, 2015.

J. Kerman, P. Wang, and J. Vaver. Estimating ad effectiveness using geo experiments in a time-based regression framework. Technical report, Google, Inc., 2017.

J. Vaver and J. Koehler. Measuring ad effectiveness using geo experiments. Technical report, Google Inc., 2011.

Y. Zhang, M. Wurm, E. Li, A. Wakim, J. Kelly, B. Price, and Y. Liu. Media mix model calibration with Bayesian priors. *research.google*, 2024.