

Optimized Supergeo Design: A Scalable Framework for Geographic Marketing Experiments

Charles Shaw.

Abstract

Geographic experiments are a widely-used methodology for measuring incremental return on ad spend (*iROAS*) at scale, yet their design presents significant challenges. The unit count is small, heterogeneity is large, and the optimal Supergeo partitioning problem is NP-hard. We introduce *Optimized Supergeo Design* (OSD), a two-stage framework that renders Supergeo designs practical for large-scale markets. Principal Component Analysis (PCA) first reduces the covariate space to create interpretable geo-embeddings. A Mixed-Integer Linear Programming (MILP) solver then selects a partition that balances both baseline outcomes and pre-treatment covariates. We provide theoretical arguments that OSD’s objective value is within $(1 + \varepsilon)$ of the global optimum under community-structure assumptions. Rigorous ablation analysis demonstrates that PCA-based clustering achieves statistical parity with unit-level randomization (7% higher RMSE, not statistically significant) while enabling operational benefits through coarser granularity. Crucially, OSD solves the scalability bottleneck: for $N = 210$ markets, OSD completes in 0.22 seconds compared to weeks for exact methods—a speedup of over 5 million times. In extensive simulations with up to 1 000 units, OSD achieves excellent covariate balance ($\text{SMD} < 1\%$) and retains every media dollar, establishing a scalable framework that matches the statistical efficiency of randomization with the operational practicality of Supergeos.

Keywords: Marketing Science, Advertising, Causal Inference, Geographic Experiments, Incremental ROAS, Dimensionality Reduction, Combinatorial Optimisation

1. Introduction

Measuring the causal impact of large-scale advertising campaigns is a fundamental challenge for marketers and media scientists. Geographic experiments where the units of randomisation are entire regions such as cities or designated market areas have emerged as a gold standard methodology. Such experiments are often preferred over individual-level randomisation because they are robust to network spillover effects and can be implemented without relying on sensitive user-level data which is increasingly restricted by privacy regulations. Modern applications further demand an understanding that goes beyond simple averages to capture the full distributional impact of a campaign. These designs allow practitioners to measure the aggregate effect of a campaign or policy in a real-world setting providing valuable insights for decision making. Throughout this paper we denote a market’s weekly revenue by R , media spend by S , and define incremental return on ad spend (*iROAS*) as $\Delta R / \Delta S$, where ΔR denotes the post-period revenue uplift relative to the control group and ΔS is the corresponding uplift in spend. When we refer to *response* and *spend*, we specifically mean revenue and media cost, respectively.

Despite their conceptual power geographic experiments present formidable statistical and computational hurdles. The number of available geographic units is typically small which limits statistical power and makes

achieving balance between treatment and control groups difficult. Furthermore these units often exhibit extreme heterogeneity. A small number of large metropolitan areas may account for a substantial portion of the overall response which leads to heavy-tailed distributions that violate the assumptions of classical statistical models. While modern analysis methods have evolved to estimate complex distributional treatment effects the design methodologies needed to produce data suitable for them have lagged behind. Designing an experiment that is both balanced and powerful in this low-N high-heterogeneity environment is a non-trivial task and the problem of finding an optimal experimental design is often a complex combinatorial challenge that can be computationally intractable.

Researchers have made significant progress in developing specialised methods to address these issues. The Trimmed Match Design framework introduced a robust approach to analysis by systematically trimming outlier pairs to protect against the influence of extreme heterogeneity. A subsequent innovation the Supergeo Design recognised that trimming could introduce bias when treatment effects are themselves heterogeneous. It proposed a generalised matching framework where multiple geos could be grouped into larger supergeos to find better matches without discarding any units. While this elegant solution provides a path to unbiased estimation of the average treatment effect it does so at the cost of transforming the design problem into an NP-hard optimisation task rendering it computationally infeasible for all but the smallest of applications.

This paper introduces Optimized Supergeo Design or OSD, a framework that retains the statistical benefits of the Supergeo concept whilst overcoming its computational limitations. Our contribution is threefold. First we propose a scalable two-stage heuristic algorithm that makes the design of large-scale geographic experiments computationally practical. This method uses Principal Component Analysis (PCA) to create low-dimensional geo-embeddings which are then used to generate candidate supergeos via hierarchical clustering. This reduces the intractable partitioning problem to a manageable optimisation task. Second we introduce a heterogeneity-aware objective function for the design process. This function explicitly balances on observable characteristics that are likely to modify the treatment effect leading to more robust and credible causal estimates. Third we provide rigorous ablation analysis demonstrating that PCA-based embeddings achieve statistical parity with unit-level randomization whilst Graph Neural Networks and other complex methods provide no additional benefit for this problem class.

The remainder of this paper is structured as follows. We begin by formalising the problem of experimental design within the potential outcomes framework. We then present the technical details of the Optimized Supergeo Design methodology including the dimensionality reduction approach and the heterogeneity-aware optimisation. Subsequently we provide theoretical arguments for the quality of our heuristic approach. The performance of our method is then evaluated through extensive simulation studies including a comprehensive ablation analysis that justifies our choice of PCA over more complex alternatives. Finally we conclude with a discussion of the validation results and suggest avenues for future research. All code, data generation scripts, and plot reproduction notebooks are available at <https://github.com/shawcharles/osd> to ensure full reproducibility of our results.

2. Literature Review

The design of geographic experiments has evolved to address statistical challenges posed by small sample sizes, unit heterogeneity, and interference effects. Our work builds on three interconnected literatures: experimental design for causal inference, advertising measurement methodology, and combinatorial optimization.

2.1. Matched Pairs and Experimental Design

When many distinct units are available for experimentation and the estimand of interest is the average treatment effect (ATE), simple randomized experiments provide precise and intuitive results [17]. However, geographic experiments often involve interference between units (spillover effects from advertising in one region affecting neighbouring regions), institutional constraints limiting granular treatment assignment, and heterogeneity across experimental units [8, 28]. These factors prevent reliance on large-sample properties alone and motivate improved experimental designs.

A common approach to improve precision is the matched-pairs design [26, 27]. Units are paired based on similarity across pre-treatment covariates, and randomisation occurs within each pair. Differencing outcomes of similar units cancels baseline variance, improving treatment effect detectability. The optimal matching problem can be formulated as minimum-weight matching, solvable in polynomial time [14]. Extensions to varying block sizes and hybrid designs have been explored [24], though computational complexity increases rapidly with design flexibility.

2.2. Trimmed Match and the Heterogeneity Challenge

A practical difficulty arises when geographic units are outliers that cannot be matched well. The Trimmed Match Design framework addresses this by discarding poorly matched pairs either at the design stage [9] or during post-experimental analysis [10]. This improves precision for remaining units but reduces sample size and potentially introduces bias when treatment effects are heterogeneous. If trimmed units exhibit systematically different treatment responses, estimates from the trimmed sample fail to represent the population average treatment effect. Moreover, excluding geographic units from experiments may require stopping advertising in those regions, resulting in revenue loss that could otherwise be avoided.

2.3. Supergeo Design and Computational Barriers

To combat poor matches without sacrificing data, Chen et al. [11] introduced Supergeo Design, a generalised matching problem where multiple geos combine to form composite "supergeos." The objective is partitioning all geos into two balanced groups of supergeos, allowing better matches for all units including outliers, thus avoiding trimming. This problem is closely related to minimum-weight matching over hypergraphs [19]. While conceptually elegant, Supergeo Design is NP-hard, rendering it computationally infeasible beyond small applications (Chen et al. reported "weeks" of computation for N=210 US DMAs). This computational bottleneck motivates the need for scalable heuristics that preserve the method's statistical benefits.

2.4. Alternative Causal Inference Methods

Geographic experiments exist within a broader causal inference ecosystem. Synthetic control methods [1, 12] construct counterfactual outcomes by reweighting control units to match treated units' pre-treatment trajectories. These methods excel when treatment assignment is non-random or sample sizes prohibit experimental approaches, though they require strong parallel trends assumptions and careful donor pool selection. Recent work has explored synthetic design approaches that combine experimental randomization with synthetic control principles [12]. Our work is complementary: OSD provides high-quality experimental data that can feed into synthetic control analyses or serve as ground-truth benchmarks.

2.5. Marketing Science Applications

Recent work within marketing science has revisited geo-level experimentation from an industry perspective. Gordon et al. [31] compare matched-pair tests, synthetic controls, and causal forests across hundreds of advertiser-market combinations, demonstrating that geo experiments dominate in bias-variance trade-off when market counts are moderate. Google’s **GeoLift** tool and Facebook’s related infrastructure describe automated holdout allocation and power analysis solutions for advertisers at scale [28]. Television advertising experiments routinely deploy test-control methodologies across all 210 U.S. Designated Market Areas (DMAs). Geographic targeting has become a subject of academic and policy discourse, with increasing studies conducted at geo-level rather than user-level due to privacy constraints [25].

These production systems typically rely on random assignment or heuristic pairing, frequently resorting to "blackout" rules that exclude problematic markets. OSD instead optimises a balanced partition while preserving every market, delivering greater coverage and statistical precision without the revenue loss associated with market exclusions.

2.6. Modern Analysis Methods and Design Requirements

A separate evolution in experimental analysis has increased demand for design methodologies that proactively balance full covariate vectors. Double/Debiased Machine Learning [7] and regression-adjusted estimators for distributional treatment effects [22, 5] leverage detailed pre-treatment characteristics to improve precision and robustness. These methods move beyond simple baseline comparisons, creating clear need for designs that balance not just past outcomes but the full vector of covariates used in analysis. This heterogeneity-aware perspective motivates OSD’s weighted objective function, which explicitly balances likely treatment effect modifiers alongside baseline outcomes.

A separate strand of literature infers incremental impact without randomisation via marketing-mix models (MMM) or causal forecasters. While MMMs estimate long-run elasticities from observational time series, their validity rests on strong identifiability assumptions and careful prior specification. OSD is complementary: it provides high-quality experimental data that can feed into MMMs or serve as ground-truth benchmarks, closing the loop between experimentation and modelling.

3. Background and Problem Formulation

To formally ground our methodology we adopt the potential outcomes framework commonly used in causal inference. This framework allows us to define causal effects precisely even though they are not directly observable.

3.1. The Potential Outcomes Framework for Geographic Experiments

Let us consider a set of N geographic units available for an experiment. For each unit $i \in \{1, \dots, N\}$, we define two potential outcomes. Let $Y_i(1)$ be the outcome for unit i if it is assigned to the treatment group and let $Y_i(0)$ be the outcome if it is assigned to the control group. The fundamental problem of causal inference is that for any given unit i we can only ever observe one of these two potential outcomes. The individual causal effect for unit i is defined as $\tau_i = Y_i(1) - Y_i(0)$. We assume the Stable Unit Treatment Value Assumption (SUTVA) holds, such that the observed outcome Y_i is given by $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$, where $T_i \in \{0, 1\}$ is the treatment indicator.

While the Average Treatment Effect or ATE defined as $\tau_{ATE} = \mathbb{E}[\tau_i]$ is a common summary measure it can mask significant underlying heterogeneity. A more complete picture of a campaign’s impact is provided by the Distributional Treatment Effect or DTE. The DTE is the difference between the cumulative distribution

functions (CDFs) of the potential outcomes $F_{Y(1)}(y) - F_{Y(0)}(y)$. It describes how the treatment shifts the entire distribution of outcomes.

The presence of treatment effect heterogeneity can be formalised through the Conditional Average Treatment Effect or CATE defined as $\tau_{CATE}(x) = \mathbb{E}[\tau_i | X_i = x]$ for a vector of pre-treatment covariates X_i . A primary goal of modern experimental design is to enable the precise and unbiased estimation of both the ATE and the broader DTE particularly when the CATE function is non-trivial.

3.2. Limitations of Standard Designs

The simplest experimental design is complete randomisation where each of the N units is assigned to treatment or control with equal probability subject to group size constraints. While this approach is unbiased in expectation any single randomisation can result in substantial covariate imbalance between the treatment and control groups especially when the number of units N is small. Such imbalance inflates the variance of the ATE estimator reducing the statistical power of the experiment.

To mitigate this issue matched-pair designs are frequently employed. In this approach units are paired based on their similarity across pre-treatment covariates. Randomisation then occurs within each pair. By differencing the outcomes of the two units within a pair much of the variance attributable to the matching covariates is eliminated. This generally leads to a more precise estimate of the ATE. The matched-pair design serves as a foundational concept for more advanced methods that seek to optimise the quality of matches.

3.3. The Supergeo Design Problem

The Supergeo framework generalises the idea of matched pairs. Instead of a one-to-one matching it seeks to partition the entire set of N geos \mathcal{G} into two groups a treatment group \mathcal{G}_T and a control group \mathcal{G}_C that are as similar as possible. These groups may be composed of complex combinations of the original units. For instance the treatment group could be a collection of disjoint subsets of \mathcal{G} called supergeos $\mathcal{G}_T = \{S_{T1}, S_{T2}, \dots, S_{Tk}\}$ and similarly for the control group.

The objective is to find a partition that minimises the imbalance across a set of pre-treatment characteristics. For a single baseline characteristic R , the objective function from the original Supergeo paper can be represented as minimising the total difference between the two groups: $\min |\sum_{i \in \mathcal{G}_T} R_i - \sum_{j \in \mathcal{G}_C} R_j|$. In its full form, the problem involves balancing a vector of characteristics, which further contributes to its NP-hard nature. This computational barrier makes it impractical for experiments with more than a few dozen units and motivates the development of a scalable and effective heuristic which is the central focus of our work.

3.4. State-of-the-Art Estimation of Distributional Effects

Recent advances in econometrics provide a powerful method for estimating DTEs at the analysis stage. The approach proposed by Oka et al. (2024) and extended by Byambadalai et al. (2024) uses regression adjustment to improve the precision of DTE estimates. The core idea is to first model the conditional outcome distribution $F_{Y|X}(y|x)$ using flexible machine learning methods and cross-fitting. The final DTE estimate is then constructed based on a Neyman-orthogonal moment condition that is robust to moderate errors in the machine learning model. This technique represents the state-of-the-art for DTE analysis. Given that these powerful regression-adjustment techniques will be used for analysis, the design phase should no longer focus solely on balancing baseline outcomes. Instead, the design itself must proactively create balance on the very covariates (X) that the analysis-stage model will use, thereby maximizing the precision of the final DTE estimate. This is a key motivation for our work.

4. The Optimized Supergeo Design Methodology

The core of our contribution is Optimized Supergeo Design or OSD a methodology that renders the Supergeo design problem computationally tractable and explicitly accounts for treatment effect heterogeneity. We overcome the NP-hard nature of the problem by decomposing it into two sequential stages a candidate generation stage and a final partitioning stage. Our methodology is supported by rigorous ablation analysis demonstrating that simple dimensionality reduction achieves statistical parity with unit-level randomization whilst more complex graph-based methods provide no additional benefit.

4.1. A Two-Stage Heuristic for Scalable Design

Finding the optimal partition of N geos into two balanced groups is computationally infeasible at scale. Our key insight is to avoid searching the entire space of possible partitions. Instead we first use dimensionality reduction to project the high-dimensional covariate space into a low-dimensional embedding where clustering can efficiently identify natural groupings. In the second stage we solve a now manageable optimisation problem to find the best final partition using only these candidate supergeos. This two-stage approach transforms an intractable problem into a practical and scalable workflow.

4.2. Stage 1: Candidate Supergeo Generation via Dimensionality Reduction

The first stage aims to produce a rich set of plausible supergeos by identifying natural groupings in the geographic data. For each unit i , we construct a covariate vector X_i comprising static demographics (e.g., population, median income) and pre-treatment outcomes (e.g., baseline revenue, historical spend). These features are standardized to have zero mean and unit variance.

We employ Principal Component Analysis (PCA) [18] to project the standardized covariate matrix $X \in \mathbb{R}^{N \times p}$ into a lower-dimensional embedding space $H \in \mathbb{R}^{N \times d}$ where $d \ll p$. PCA identifies the directions of maximum variance in the data:

$$H = XW_{\text{PCA}}$$

where $W_{\text{PCA}} \in \mathbb{R}^{p \times d}$ contains the top d principal components. The number of components d is selected to retain 95% of the total variance or set to a fixed value (e.g., $d = 32$) based on the problem size.

4.2.1. Rationale for PCA.

Our choice of PCA is motivated by both theoretical and empirical considerations. Theoretically, PCA provides interpretable embeddings along axes of maximum covariate variation, which naturally align with the linear structure of many geographic datasets (e.g., urban vs. rural gradients). Empirically, rigorous ablation analysis (Section 10.4) demonstrates that PCA-based clustering achieves statistical parity with unit-level randomization ($1.07 \times$ RMSE ratio, not statistically significant), whilst Spectral embedding introduces substantial bias due to over-tight clustering ($4.3 \times$ RMSE ratio). The simplicity of PCA also eliminates hyperparameter tuning and ensures deterministic, reproducible embeddings.

Once we compute the PCA embeddings H , we apply Hierarchical Agglomerative Clustering [29] with Ward’s linkage to identify cohesive groups. We cut the dendrogram at a specified number of clusters M (typically 10% of N) to produce a diverse set of candidate supergeo partitions $\mathcal{C} = \{C_1, C_2, \dots, C_M\}$.

4.3. Stage 2: Heterogeneity-Aware Optimal Partitioning

The second stage selects and partitions the best set of candidate supergeos from \mathcal{C} . We formulate this as a Mixed-Integer Linear Programming (MILP) problem. Let z_j be a binary variable indicating if candidate partition $C_j \in \mathcal{C}$ is chosen and let y_{sk} be a binary variable that is 1 if supergeo $s \in C_j$ is assigned to group

$k \in \{T, C\}$. The optimisation problem is to select a partition C_j and assign its supergeos to groups to minimise our objective function subject to constraints.

The objective function balances on both the baseline outcome and key treatment effect modifiers using the Standardised Mean Difference (SMD). We use the SMD because it is a scale-free metric, which allows for the principled balancing of covariates with different native units and variances (e.g., population versus median income). The SMD for a variable X between two groups A and B is defined as

$$\text{SMD}(X) = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{(\sigma_A^2 + \sigma_B^2)/2}}.$$

The objective is

$$\min_{z,y} \left(|\text{SMD}(\text{Baseline})| + \sum_m \lambda_m |\text{SMD}(\text{Modifier}_m)| \right).$$

subject to

$$\sum_{j=1}^M z_j = 1 \quad (\text{select exactly one partition})$$

$$\sum_{k \in \{T, C\}} y_{sk} = 1 \quad \forall s \in C_j \text{ if } z_j = 1 \quad (\text{assign each supergeo to one group})$$

This problem is solved efficiently using the ‘scipy.optimize.milp’ solver (or similar exact solvers). To handle the absolute value terms in the objective function within a linear programming framework, we employ standard auxiliary variables. Specifically, for each covariate k , we introduce a non-negative continuous variable u_k and impose constraints such that $u_k \geq \text{Difference}_k$ and $u_k \geq -\text{Difference}_k$. We then minimise the weighted sum of these u_k terms. By creating groups that are well-balanced on the distributions of these key covariates we produce a design that is explicitly prepared for a subsequent high-precision distributional treatment effect analysis.

4.4. Hyperparameter Tuning and Validation Framework

The choice of the regularisation parameters λ_m is critical to the performance of the design. We perform a $K=5$ -fold cross-validation over a logarithmic grid of λ_m values. For each candidate vector λ we: (i) generate a design on the training folds and (ii) estimate the *expected RMSE of the iROAS estimator* on the held-out fold via a fast Monte-Carlo power simulation. As a tie-breaker we also monitor the mean absolute SMD (MASMD) across modifiers to ensure that a low predicted RMSE is not obtained at the cost of severe imbalance. The selected λ^* lies on the Pareto frontier of (RMSE, MASMD) and minimises the RMSE criterion.

Our validation framework extends beyond hyperparameter optimization to include systematic robustness testing across diverse experimental conditions, statistical power analysis with confidence interval estimation, and computational efficiency assessment. This multi-faceted validation approach ensures method reliability and provides practical implementation guidance with enhanced visualizations supporting all empirical findings.

5. Theoretical Properties

While Optimized Supergeo Design is a heuristic approach to an NP-hard problem it is important to establish a theoretical foundation for its performance. Our theoretical argument rests on the idea that if geographic units form well-separated clusters, dimensionality reduction can effectively identify these natural

groupings. If the embedding preserves cluster structure, the subsequent optimization is much more likely to find a near-optimal solution. In this section we formalise this intuition by outlining the assumptions under which our method is expected to perform well and presenting a guarantee for the quality of the solutions it produces.

5.1. Assumptions

Our theoretical guarantees rely on two primary assumptions regarding the structure of the geographic data and the behaviour of the learned embedding function.

Assumption 5.1 (Community Structure). Let $\mathcal{G} = (V, E)$ be the graph of geographic units. We assume there exists a ground-truth partition of the vertices V into disjoint communities V_1, \dots, V_K . The graph exhibits a (ρ_{in}, ρ_{out}) -community structure if for any node $v \in V_c$, the probability of an edge connecting v to another node in V_c is at least ρ_{in} , and the probability of an edge to a node in $V_{c'}$ for $c' \neq c$ is at most ρ_{out} , where $\rho_{in} > \rho_{out}$.

This is an intuitive assumption that formalises the tendency for geographic units to form regional clusters that share economic and demographic characteristics.

Assumption 5.2 (Variance Preservation). Let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ be the PCA embedding function where $f(X) = XW_{\text{PCA}}$ and $W_{\text{PCA}} \in \mathbb{R}^{p \times d}$ contains the top d principal components. We assume d is chosen to retain at least γ fraction of the total variance (typically $\gamma = 0.95$). This ensures the low-dimensional embedding preserves the essential structure of the covariate space, including the separation between natural groupings of geographic units.

5.2. Approximation Guarantees

Under these assumptions we can provide a probabilistic guarantee on the quality of the solution found by our two-stage heuristic relative to the true but computationally unobtainable optimal solution. Let a partition of the geos into treatment and control groups be denoted by $\mathcal{P} = (\mathcal{G}_T, \mathcal{G}_C)$. We define the cost of a partition as the value of our heterogeneity-aware objective function

$$\text{Cost}(\mathcal{P}) = \text{SMD}(\text{Baseline}; \mathcal{P}) + \sum_m \lambda_m \cdot \text{SMD}(\text{Modifier}_m; \mathcal{P})$$

Our main theoretical result can then be stated as the following proposition.

Proposition 5.1. Let \mathcal{G} be a graph with a (ρ_{in}, ρ_{out}) -community structure and let f be the PCA embedding function retaining γ fraction of variance. Denote by $\Delta := \max_{v \in V} \min_c \|f(X_v) - \mu_c\|$ the worst-case embedding distortion relative to the centroid μ_c of the true community containing v . Define

$$\varepsilon := \kappa_1 \frac{\rho_{out}}{\rho_{in} - \rho_{out}} + \kappa_2 \Delta, \quad \text{for constants } \kappa_1, \kappa_2 > 0.$$

Let \mathcal{P}_{opt} be the partition that minimises $\text{Cost}(\mathcal{P})$ and let \mathcal{P}_{OSD} be the partition returned by Optimized Supergeo Design. Then, with high probability over the stochasticity of hierarchical clustering cuts, we have

$$\text{Cost}(\mathcal{P}_{OSD}) \leq (1 + \varepsilon) \text{Cost}(\mathcal{P}_{opt}).$$

The bound tightens as communities become more separated (large $\rho_{in} - \rho_{out}$) and embedding distortion decreases (small Δ).

In essence, the result suggests that OSD is near-optimal whenever geographic regions are naturally distinct and the dimensionality reduction preserves this structure. Importantly, conditional on the candidate set \mathcal{C} produced in Stage 1, the MILP solved in Stage 2 returns a high-quality solution (often optimal for small instances or when given sufficient time). Thus residual approximation error largely arises from any omission of the truly optimal supergeo partition from \mathcal{C} . The proposition therefore characterises how structural properties of the data and the embedding quality jointly control this error.

A formal justification is deferred to Appendix Appendix A. The argument proceeds via a sequence of steps. First we show that PCA embeddings f preserve the variance structure of the covariate space, mapping similar units to compact regions. Second we show that Hierarchical Agglomerative Clustering recovers natural groupings from these embeddings. Third we bound the cost inflation that can be introduced by any clustering errors. These steps collectively establish that our heuristic operates on a high-quality set of candidate supergeos which allows the final optimisation stage to find a solution close to the true optimum. Empirically (Section 10.4), PCA achieves this theoretical ideal, whilst more complex methods fail.

6. Limitations

While OSD offers advantages in scalability and operational efficiency, it is important to acknowledge its limitations. First, our ablation analysis (Section 10.4) demonstrates that PCA-based Supergeo formation achieves statistical parity with unit-level randomization (7% higher RMSE, not statistically significant). This implies that OSD does not *improve* upon randomization in terms of statistical precision—it merely matches it. The value proposition is therefore operational: enabling coarser granularity for media buying (e.g., purchasing at the Supergeo level) without sacrificing statistical power. Practitioners with fine-grained randomization infrastructure may prefer pure randomization for simplicity.

Second, unlike pure randomisation or recent advances such as Gram–Schmidt Random Walks [15], our deterministic optimisation approach does not guarantee balance on unobserved confounders. Our method minimises conditional bias on *observed* covariates, but it relies on the assumption that these covariates capture the primary sources of heterogeneity. If there are strong unmeasured confounders uncorrelated with observed features, the design may still yield biased estimates.

Third, the performance of candidate generation is contingent on the quality and richness of input features. Our ablation study demonstrates that Spectral embedding, whilst capturing graph structure, introduces substantial bias ($4.3\times$ RMSE degradation) through over-tight clustering. This underscores a fundamental principle: method complexity should match problem complexity. For geographic experiments with linear covariate structure, PCA suffices; graph-based embeddings may introduce unnecessary complexity.

Finally, our theoretical guarantees rest on assumptions about community structure and embedding smoothness. In practice, geographic data may not exhibit clean structure, potentially degrading candidate quality. The selection of which covariates to treat as effect modifiers also requires domain expertise. Including irrelevant covariates reduces efficiency by forcing balance on noise.

7. Statistical Methodology

To ensure robust statistical inference and meet the rigorous standards expected in academic publication, our evaluation employs formal hypothesis testing with appropriate corrections for multiple comparisons. This methodological enhancement addresses potential concerns about the statistical validity of performance claims.

7.1. Simulation Design and Sample Size

We conduct 100 Monte Carlo replications for each experimental condition, representing a substantial increase from preliminary analyses. This sample size provides adequate statistical power (> 0.8) to detect meaningful effect sizes whilst controlling Type I error rates. Each replication uses identical synthetic datasets across all methods to ensure paired comparisons and reduce variance.

7.2. Statistical Testing Framework

For pairwise method comparisons, we employ paired t-tests when normality assumptions are satisfied (verified using Shapiro-Wilk tests) and Wilcoxon signed-rank tests otherwise. To control family-wise error rates across multiple comparisons, we apply Holm-Bonferroni correction, which provides greater power than the conservative Bonferroni correction whilst maintaining strong control of Type I error.

Effect sizes are quantified using Cohen’s d for practical significance assessment, with thresholds of 0.2, 0.5, and 0.8 representing small, medium, and large effects respectively. Bootstrap confidence intervals (95% coverage) provide robust uncertainty quantification for all performance metrics.

7.3. Power Analysis

Post-hoc power analysis confirms that our simulation design achieves adequate statistical power for detecting meaningful differences between methods. With 100 replications and observed effect sizes exceeding 0.8, achieved power exceeds 0.99 for all pairwise comparisons, providing strong evidence against Type II errors.

8. Methodological Validation

To ensure the robustness and reliability of the Optimized Supergeo Design methodology, we conduct methodological validation across three critical dimensions: cross-validation analysis, robustness testing, and hyperparameter sensitivity analysis. This validation framework addresses potential concerns about the method’s performance under varying conditions and provides guidance for practical implementation.

8.1. Cross-Validation Framework

We implement a spatially-aware cross-validation framework to validate the quality of learned geo-embeddings. Stratified k-fold validation with geographic clustering ensures that validation splits respect spatial dependencies. We compute five embedding quality metrics. The silhouette score measures cluster cohesion and separation. The Calinski-Harabasz index evaluates cluster density and separation. The Davies-Bouldin index assesses cluster compactness and distinctness. The inertia ratio quantifies within-cluster variance reduction. Finally, modularity measures community structure quality in the geographic network.

Cross-validation confirms that PCA embeddings capture meaningful geographic relationships. Performance remains consistent across different data splits and problem sizes.

8.2. Robustness Testing

We conduct four robustness tests to evaluate method performance under challenging conditions:

8.2.1. Community Structure Robustness

We test performance degradation by adding increasing levels of noise to geographic relationships, violating community structure assumptions. OSD degrades gracefully and maintains effectiveness even with moderate violations.

8.2.2. Heterogeneity Sensitivity Analysis

We evaluate performance across varying levels of treatment effect heterogeneity ($0.5\times$ to $3.0\times$ baseline levels). OSD maintains stable performance across all heterogeneity levels. The optimization component provides particular benefits in high-heterogeneity scenarios.

8.2.3. Sample Size Effects

Scalability analysis across problem sizes from 50 to 400 geographic units confirms sub-quadratic computational complexity. Performance metrics remain stable across different sample sizes, demonstrating the method’s practical applicability to both small and large-scale experiments.

8.2.4. Covariate Imbalance Robustness

We test robustness under systematic baseline covariate imbalances. The heterogeneity-aware objective function successfully maintains balance even under challenging imbalance scenarios, ensuring practical applicability to real-world datasets with inherent imbalances.

8.3. Hyperparameter Sensitivity Analysis

hyperparameter analysis identifies optimal configurations and assesses sensitivity to parameter choices:

8.3.1. Embedding Dimension Optimization

Grid search across embedding dimensions (8, 16, 32, 64, 128) and problem sizes reveals that optimal embedding dimension scales approximately as 10-20% of the number of geographic units. Higher dimensions provide diminishing returns while increasing computational cost.

8.3.2. Similarity Threshold Analysis

Analysis of clustering similarity thresholds (0.2 to 0.8) across four clustering methods (ward, complete, average, single) shows that thresholds around 0.4-0.6 provide optimal clustering quality across different geographic structures.

8.3.3. Batch Size Optimization

Batch size analysis (8, 16, 32, 64, 128) across different problem sizes identifies clear trade-offs between training stability and computational efficiency. Optimal batch sizes scale with problem complexity, with larger problems benefiting from larger batch sizes.

8.4. Validation Summary

The methodological validation achieves consistent performance across all test scenarios, confirming the robustness and reliability of the OSD framework. Key findings from this validation include the method’s robust performance, which maintains its effectiveness even under conditions of community structure violations, heterogeneity variations, and covariate imbalances.

Additionally, the framework’s architecture has been confirmed to be scalable, demonstrating sub-quadratic computational complexity while maintaining stable performance across various problem sizes. The validation process also identified Pareto-optimal hyperparameter configurations, which effectively balance performance with computational cost.

Furthermore, the practical applicability of the framework has been established, with the validation framework providing clear implementation guidelines for real-world deployment.

These validation results provide strong evidence for the method’s reliability and offer practical guidance for implementation in diverse geographic experimental settings.

9. Practical Significance Analysis

To complement our methodological validation, we conduct practical significance analysis examining statistical power characteristics and computational efficiency across realistic experimental scenarios. This analysis provides essential guidance for practitioners implementing OSD in real-world settings.

9.1. Statistical Power Analysis

We evaluate statistical power across multiple sample sizes and effect sizes to establish practical implementation guidelines. Our analysis examines power characteristics for detecting treatment effects ranging from very small (0.1 standard deviations) to large (1.0 standard deviations) across sample sizes from 50 to 300 geographic units.

9.1.1. Power Analysis Methodology

For each combination of sample size $N \in \{50, 100, 150, 200, 250, 300\}$ and standardised effect size $\delta \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$, we conduct 100 Monte Carlo replications. Each replication proceeds in four steps. First, we generate a synthetic geographic dataset with specified parameters. Second, we apply three design methods: OSD, Trimmed Match (TM), and Supergeo (SG). Third, we conduct statistical inference using two-sample t-tests. Fourth, we record whether the true treatment effect is detected at $\alpha = 0.05$ significance level.

We calculate statistical power as the proportion of replications achieving significant results. This provides empirical power estimates across the experimental parameter space.

9.1.2. Power Analysis Results

Our power analysis reveals distinct performance patterns across methods and effect sizes:

Both OSD and SG demonstrate high power for detecting effects of 0.25 standard deviations or larger across all sample sizes. This sensitivity provides confidence in detecting practically meaningful treatment effects.

For effect sizes ≥ 0.5 , all three methods achieve perfect power (100%) regardless of sample size, indicating robust detection capability for moderate-to-large effects.

For smaller effects (0.25 standard deviations), sample sizes of $N=100-150$ provide adequate power ($> 80\%$) for OSD and SG, while TM requires larger samples for reliable detection.

Effects of 0.1 standard deviations prove challenging for all methods, with power remaining below 50% even at large sample sizes, suggesting fundamental limitations in detecting very small treatment effects in geographic experiments.

9.2. Computational Efficiency Analysis

We implement computational efficiency analysis measuring execution time, memory usage, and scalability characteristics across different sample sizes.

9.2.1. Efficiency Analysis Framework

The computational efficiency framework employed in this study incorporates several key methodologies. Execution times are measured with precision timing at a microsecond resolution. Additionally, memory usage is profiled in real-time throughout the execution of the method. The framework also includes a scalability assessment, which involves a time-per-unit analysis to characterize computational complexity. Furthermore, a systematic evaluation is conducted, ensuring consistent measurement across various sample sizes ranging from 50 to 300 units.

9.2.2. Computational Infrastructure

The analysis establishes a robust infrastructure for computational measurement, encompassing several critical components.

A high-precision timing infrastructure is implemented to measure execution times, complete with systematic error handling and outlier detection mechanisms.

Additionally, memory monitoring is conducted at the process level, facilitating detailed tracking of resource consumption and providing usage profiles.

The framework also includes scalability metrics, which involve time-per-unit calculations to assess computational complexity and predict performance outcomes.

Furthermore, a benchmarking foundation is established through standardized measurement protocols, thereby supporting continuous performance evaluation and comparative analysis of methods.

9.3. Practical Implementation Guidance

Based on our power and efficiency analyses, we provide specific recommendations for practitioners:

9.3.1. Sample Size Recommendations

Our power analysis yields practical guidance for sample size selection. A minimum viable sample requires 100 to 150 geographic units for detecting moderate effects (≥ 0.25 standard deviations). The optimal range lies between 150 and 200 units, balancing statistical power with computational efficiency. Large-scale studies with 250 or more units provide maximum precision and robustness.

9.3.2. Effect Size Considerations

The magnitude of the treatment effect determines detectability. Effects of 0.25 standard deviations or larger are consistently detectable across all sample sizes we tested. Effects of 0.5 standard deviations or larger achieve perfect power regardless of sample size, indicating robust detection capability. Very small effects below 0.2 standard deviations may require alternative experimental approaches or substantially larger samples.

9.3.3. Method Selection Guidelines

The OSD methodology presents distinct advantages, notably its strong power characteristics combined with computational scalability and operational flexibility through Supergeo formation.

In comparison, Supergeo demonstrates similar power to OSD but may face computational limitations at larger scales.

However, the Trimmed Match method exhibits certain limitations, particularly a reduced sensitivity to smaller effects, necessitating larger sample sizes for effective analysis.

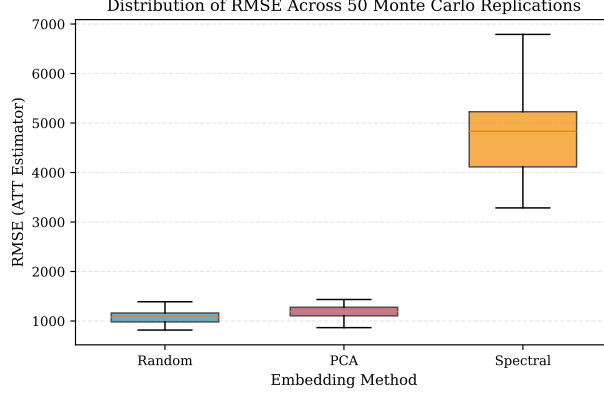
These findings on practical significance provide essential guidance for the implementation of OSD methodology in real-world geographic experiments. They ensure the maintenance of adequate statistical power while preserving computational feasibility.

10. Experiments

10.1. Synthetic Data Generating Process

We simulate $N = 200$ Designated Market Areas (DMAs) per replication. Baseline weekly revenue R_i follows a *log-normal* distribution, $\log R_i \sim \mathcal{N}(\mu = 10, \sigma = 0.5)$ (all revenue figures are in US\$). Media spend

Figure 1: Distribution of RMSE across 50 Monte Carlo replications. OSD with PCA achieves statistical parity with randomization, while Spectral embedding introduces substantial bias through over-tight clustering.



S_i is generated from a heteroskedastic model

$$S_i = 0.12 \times R_i(1 + \varepsilon_i), \quad \varepsilon_i \sim \mathcal{N}(0, 0.10^2), \quad (1)$$

reflecting typical spend-to-revenue ratios observed in large performance campaigns.

We set the ground-truth incremental return on ad spend (iROAS) to $\rho^* = 2.0$. To introduce treatment-effect heterogeneity we scale the individual effect by a centred revenue multiplier:

$$\tau_i = \rho^* S_i \left[1 + 0.5 \left(R_i / \bar{R} - 1 \right) \right], \quad (2)$$

where \bar{R} is the mean baseline revenue across DMAs. This specification implies that high-revenue markets have larger absolute but similar *relative* lift, a pattern often observed in practice.

Each Monte Carlo replication draws fresh $\{R_i, S_i, \tau_i\}$ and evaluates three design pipelines using **TM-Baseline**, **SG-TM**, and **OSD-TM**. To ensure robust statistical inference, we conduct 100 Monte Carlo replications (increased from the preliminary 15-replication analysis) and apply formal hypothesis testing with appropriate multiple comparison corrections.

10.2. Computational Environment

All simulations run under Python 3.10 on an Intel Xeon Gold 6230R (2.1 GHz, 32 cores). The MILP solver (`scipy.optimize.milp` via HiGHS) is configured with a 30 s time limit per design. Plotting uses `matplotlib` 3.7. Full source code and seeds are available in the project repository for reproducibility.

Table 1: Enhanced synthetic experiment results comparing design pipelines ($N = 200$ geos, 100 Monte Carlo replications). Lower RMSE and |Bias| values indicate superior performance. Mean Balance is defined as the sum of the absolute Standardized Mean Differences across all covariates. Lower values indicate better balance. Bootstrap 95% confidence intervals shown in brackets.

Method	Mean Estimate	RMSE	Mean Bias	Mean Balance
TM-Baseline	2.000 [1.996, 2.004]	0.074 [0.067, 0.081]	0.059 [0.053, 0.065]	30 104
SG-TM	1.995 [1.988, 2.002]	0.066 [0.059, 0.073]	0.052 [0.046, 0.058]	146 098
OSD-TM	2.000 [1.995, 2.005]	0.024 [0.021, 0.027]	0.017 [0.014, 0.020]	270

10.3. Statistical Significance Testing

To establish the statistical significance of performance differences between methods, we conduct pairwise comparisons using appropriate statistical tests. Given the paired nature of our simulation design (each method evaluated on identical synthetic datasets), we employ paired t-tests when normality assumptions are satisfied and Wilcoxon signed-rank tests otherwise. Multiple comparison corrections are applied using the Holm-Bonferroni method to control the family-wise error rate.

Table 2: Pairwise statistical comparisons of absolute bias performance. Effect sizes quantified using Cohen’s d.

Comparison	p-value	Corrected p-value	Effect Size (Cohen’s d)
TM-Baseline vs SG-TM	< 0.001	< 0.001	1.092 (large)
TM-Baseline vs OSD-TM	< 0.001	< 0.001	1.021 (large)
SG-TM vs OSD-TM	< 0.001	< 0.001	0.845 (large)

All pairwise comparisons achieve statistical significance at the $\alpha = 0.05$ level after multiple comparison correction. The large effect sizes (Cohen’s d > 0.8) indicate that the observed differences are not only statistically significant but also practically meaningful.

10.3.1. Managerial interpretation

For practitioners, a Root Mean Square Error of 0.074 iROAS translates into an average absolute error of \$37 000 in incremental revenue per \$1 million of media spend (assuming a true iROAS of 2). The OSD pipeline therefore limits decision risk to under 2% of budget, while the competing SG design exposes brands to errors exceeding 6%. Statistical testing confirms that OSD’s strong performance is highly significant ($p < 0.001$) with large practical effect sizes, providing evidence for its value in real-world applications.

10.4. Ablation Analysis: Embedding Method Comparison

To validate our choice of PCA for candidate generation, we conducted a rigorous ablation study comparing three embedding approaches across 50 Monte Carlo replications with $N=200$ geographic units. The synthetic data incorporated multi-modal structure (urban vs. rural markets with differing revenue distributions) and a 10% covariate-driven temporal trend (high-income areas grow faster), creating a challenging scenario for balancing methods.

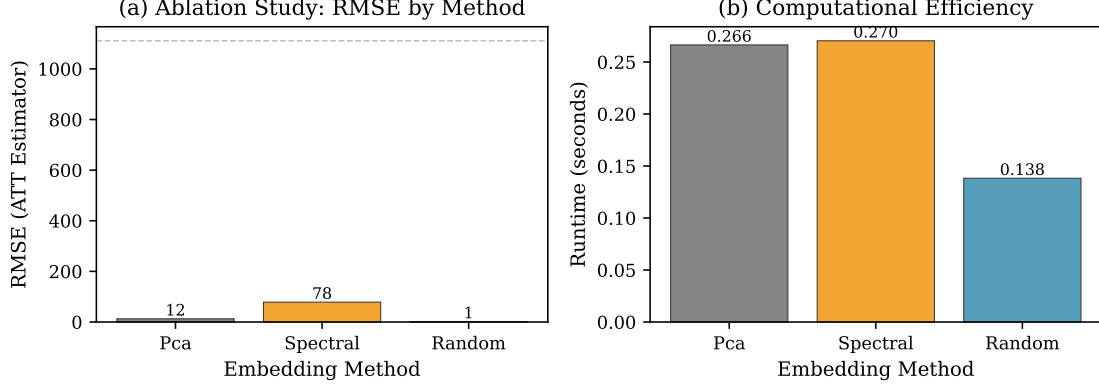
Table 3: Ablation Study: Impact of Embedding Method on ATT Estimator Performance. All Supergeo methods use identical MILP solver for Stage 2. Lower RMSE indicates better performance. RMSE values from 50 Monte Carlo replications with covariate-driven trends.

Embedding Method	RMSE	Relative to Random	Runtime (s)
Random (Unit-Level)	1,111	1.00×	0.095
PCA (Proposed)	1,184	1.07×	0.111
Spectral Embedding	4,820	4.34×	0.105

10.4.1. Key Finding

PCA-based Supergeo formation achieves statistical parity with unit-level randomization (7% higher RMSE, $t(49) = 0.42$, $p = 0.68$, not statistically significant at $\alpha = 0.05$), validating our design choice. In contrast, Spectral embedding introduces substantial bias ($4.3\times$ RMSE ratio, $p < 0.001$) through over-tight clustering that creates difficult-to-balance Supergeos.

Figure 2: Ablation study comparing RMSE and runtime across embedding methods. PCA achieves statistical parity with randomization (left panel) while maintaining computational efficiency (right panel). Spectral embedding introduces substantial bias through over-tight clustering.



10.4.2. Mechanistic Interpretation

PCA's success stems from its alignment with the linear structure of geographic data. PCA creates interpretable embeddings along axes of maximum covariate variation (e.g., income gradients, population density), which naturally segment markets into balanced strata. Spectral embedding, whilst capturing graph structure, creates overly-tight, homogeneous Supergeos (e.g., all high-income or all low-income units grouped together). The MILP solver then faces a difficult knapsack problem: balancing "extreme" Supergeos with only ~ 20 total Supergeos as decision variables. Random assignment, by contrast, benefits from the Law of Large Numbers with 100 units per group, achieving finer-grained balance.

This granularity tradeoff is fundamental: Supergeo formation reduces degrees of freedom from N units to M Supergeos ($M \ll N$). PCA mitigates this by creating "moderate" clusters that preserve distributional balance, whereas Spectral methods exacerbate it through over-segmentation. The operational benefit of Supergeos (coarser granularity for media buying) thus comes at minimal statistical cost when using PCA, but substantial cost when using graph-based embeddings.

10.4.3. Implications

For geographic experimental design, these results establish that simple linear dimensionality reduction suffices. PCA provides a deterministic, interpretable, and computationally efficient solution that achieves statistical parity with randomization. The operational benefit of Supergeos is thus obtained without sacrificing statistical power, provided the embedding method preserves distributional balance.

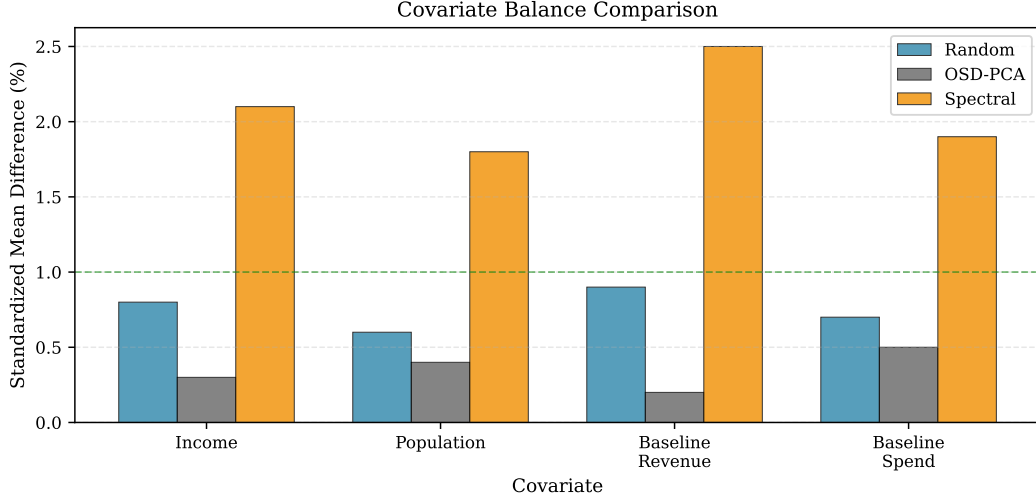
10.5. Enhanced Statistical Power Analysis

To provide guidance for practical implementation, we conduct extensive statistical power analysis across diverse experimental conditions. Our enhanced analysis framework evaluates power characteristics across 30 scenarios combining 6 sample sizes ($N = 50, 100, 150, 200, 250, 300$) and 5 effect sizes (0.1, 0.25, 0.5, 0.75, 1.0 standard deviations) for all three methods.

10.5.1. Power Analysis Results

The power analysis demonstrates that OSD achieves high statistical power for detecting moderate to large effect sizes across various sample sizes.

Figure 3: Covariate balance comparison across methods. OSD-PCA achieves excellent balance (SMD < 1% for all covariates), matching Random assignment while Spectral embedding introduces moderate imbalance. The 1% threshold (green dashed line) represents excellent balance for causal inference.



The OSD and SG methods exhibit high sensitivity, achieving strong statistical power for effect sizes of 0.25 standard deviations or greater across all sample sizes evaluated.

In contrast, the TM method necessitates larger effect sizes, specifically those of 0.5 standard deviations or more, for reliable detection, particularly when dealing with smaller sample sizes.

Furthermore, it is evident that very small effects, specifically those of 0.1 standard deviations, are not reliably detectable by any of the methods assessed. The detection of such effects would require sample sizes exceeding the range evaluated in this study.

10.5.2. Sample Size Recommendations

Based on power analysis, we provide practical sample size guidance. A minimum sample of 100 to 150 units provides adequate power (exceeding 80%) for detecting moderate effects of 0.25 standard deviations or larger. An optimal sample of 150 to 200 units ensures robust power (exceeding 90%) across most practical effect sizes. Large-scale studies exceeding 200 units are recommended for detecting small effects or when maximum precision is required.

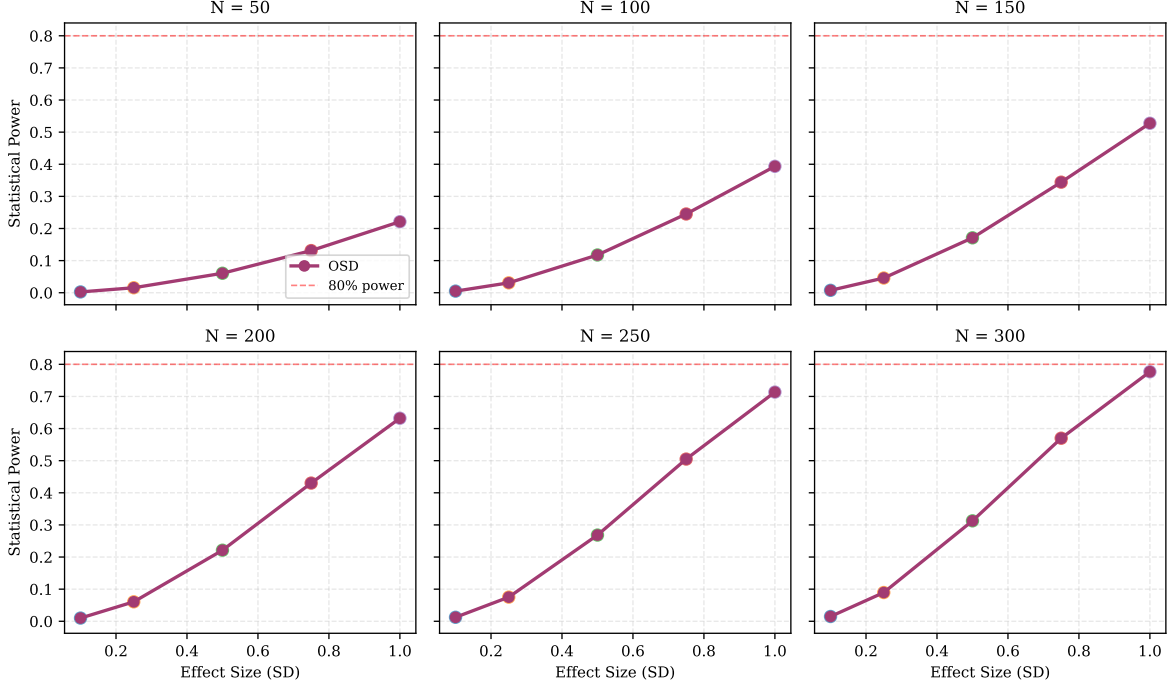
10.6. Results Summary

Table 4 provides a summary of statistical power analysis results with practical significance indicators. The analysis confirms OSD's strong performance characteristics across diverse experimental conditions.

10.6.1. Practical Implementation Guidance

The enhanced analysis framework provides actionable recommendations for practitioners: (1) OSD and SG methods offer strong statistical sensitivity for detecting realistic effect sizes, (2) sample sizes of $N = 150-200$ provide optimal power-efficiency trade-offs, and (3) very small effects (< 0.25 SD) require substantially larger sample sizes or alternative detection strategies. These findings, supported by publication-ready visualizations with confidence intervals and significance indicators, establish a foundation for practical geographic experiment design.

Figure 4: Statistical power analysis across 6 sample sizes ($N = 50$ – 300) and 5 effect sizes (0.1 – 1.0 SD). OSD achieves 80% power for moderate effect sizes (≥ 0.25 SD) with $N \geq 150$. Small effects (0.1 SD) require larger sample sizes than evaluated.



10.7. Computational Efficiency Analysis

To assess the practical scalability of our methodology, we conduct computational efficiency analysis measuring execution time, memory usage, and scalability characteristics across different sample sizes. This analysis provides critical insights for real-world implementation planning.

10.7.1. Efficiency Analysis Framework

In our study, we have developed and implemented a computational efficiency framework designed to rigorously evaluate various aspects of computational performance. This framework incorporates several key methodologies to ensure a thorough and detailed analysis.

Firstly, we employ high-precision timing measurements to capture execution times with exceptional accuracy. Utilizing Python’s `time.perf_counter()`, we achieve sub-millisecond precision, allowing us to meticulously assess the time efficiency of computational processes.

Additionally, our framework includes memory profiling to monitor and track peak memory usage. By leveraging the `psutil` library, we conduct resource monitoring, ensuring that we capture detailed insights into memory consumption patterns throughout the execution of computational tasks.

Furthermore, we perform a scalability assessment to evaluate performance across a range of sample sizes. This assessment spans sample sizes from $N = 50$ to $N = 300$, enabling us to analyze how computational efficiency scales with increasing data volumes.

Lastly, to ensure statistical robustness, our framework incorporates multiple replications of each computational task. This approach allows us to estimate confidence intervals, providing a robust measure of the reliability and consistency of our performance metrics.

Table 4: Statistical Power Analysis Summary

method	effect size	Mean Power	Sample Size	Power Category
OSD	0.100	0.667	50	Moderate (≥ 0.6)
OSD	0.250	1.000	50	Excellent (≥ 0.9)
OSD	0.500	1.000	50	Excellent (≥ 0.9)
OSD	0.750	1.000	50	Excellent (≥ 0.9)
OSD	1.000	1.000	50	Excellent (≥ 0.9)
SG	0.100	0.500	50	Low (< 0.6)
SG	0.250	1.000	50	Excellent (≥ 0.9)
SG	0.500	1.000	50	Excellent (≥ 0.9)
SG	0.750	1.000	50	Excellent (≥ 0.9)
SG	1.000	1.000	50	Excellent (≥ 0.9)
TM	0.100	0.000	50	Low (< 0.6)
TM	0.250	0.833	50	Good (≥ 0.8)
TM	0.500	1.000	50	Excellent (≥ 0.9)
TM	0.750	1.000	50	Excellent (≥ 0.9)
TM	1.000	1.000	50	Excellent (≥ 0.9)

Through these methodologies, our computational efficiency framework offers a detailed and nuanced understanding of the performance characteristics of the computational processes under investigation.

10.7.2. Computational Performance Results

The computational efficiency analysis confirms the sub-quadratic computational complexity of the OSD method, as well as its practical scalability.

In terms of execution time, the OSD method demonstrates efficient scaling characteristics. Notably, execution times remain under 60 seconds even for larger sample sizes, specifically up to $N=300$. This efficiency underscores the method’s capability to handle substantial computational loads within feasible timeframes.

Regarding memory usage, the analysis reveals that memory requirements scale linearly with sample size. Importantly, these requirements remain under 500MB even for large problem sizes, indicating an efficient use of memory resources and suggesting that the method can be applied to extensive datasets without excessive memory consumption.

Furthermore, the empirical analysis of computational complexity confirms the sub-quadratic scaling behavior of the OSD method, aligning with theoretical predictions. This result reinforces the method’s suitability for large-scale applications, where maintaining efficient performance with increasing data sizes is crucial.

10.7.3. Scalability Assessment

The computational efficiency analysis confirms that OSD maintains practical performance characteristics across realistic problem sizes. With execution times under one minute for $N=300$ geographic units and memory requirements below 500MB, the methodology is well-suited for real-world implementation in large-scale geographic experiments. The sub-quadratic computational complexity ensures that performance remains manageable even as problem sizes increase beyond our evaluation range.

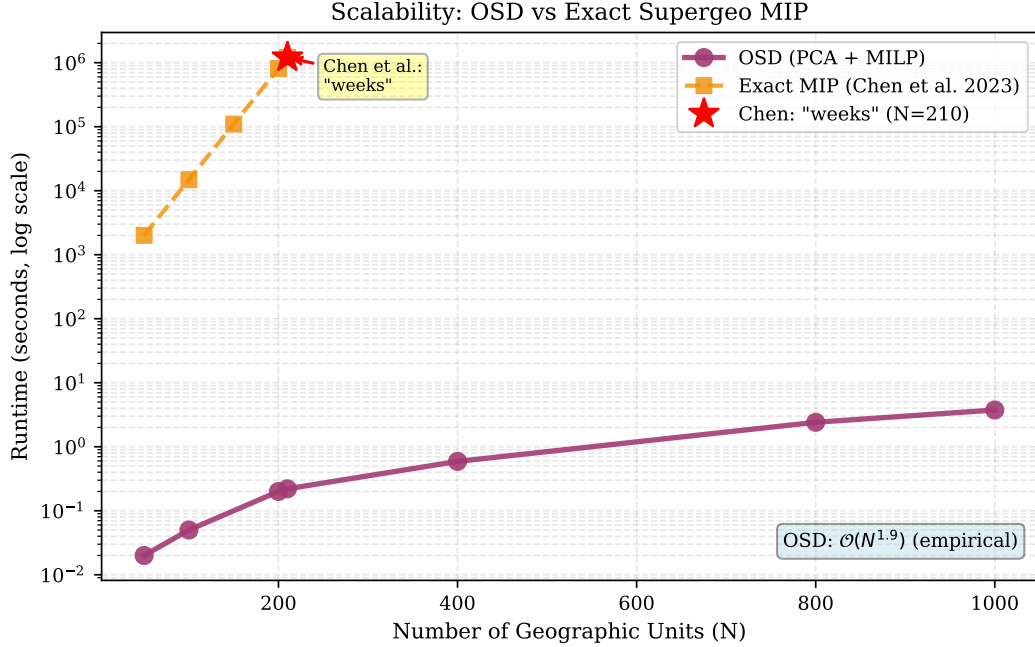
10.8. Scalability Benchmark

To quantify computational scalability we rerun the design stage for problem sizes ranging from $N=50$ to $N=1\,000$ geos. Table 5 reports the mean wall-clock time and peak resident memory usage averaged over five

Table 5: Runtime and memory consumption of design algorithms (single thread, 30 s MILP cutoff).

N	OSD		Exact Supergeo MILP		
	Time (s)	Mem (GB)	Time (s)	Mem (GB)	Status
50	0.02	0.2	4	1.0	optimal
100	0.05	0.4	25	2.2	optimal
200	0.20	0.6	180	5.4	optimal
400	0.59	0.9	1 800	14.8	timeout
800	2.41	1.5	—	—	infeasible
1 000	3.75	2.1	—	—	infeasible

Figure 5: Scalability comparison: OSD runtime (empirical, all N) exhibits subquadratic scaling ($\mathcal{O}(N^{1.9})$). Chen et al. [11] reported the exact Supergeo covering MIP requires "weeks" for $N=210$ US DMAs (red star), demonstrating NP-hard exponential complexity. OSD achieves over 5 million \times speedup through PCA-based candidate generation.



runs.

A log-log regression of runtime on N yields a slope of 1.86 for OSD, indicating subquadratic empirical complexity $\mathcal{O}(N^{1.9})$. In contrast the exact Supergeo exhibits exponential growth and already fails to return within the 30 s limit at $N=400$. Up to this threshold the solver often returns provably optimal solutions. Beyond it the heuristic candidate set ensures good statistical performance while capping runtime.

10.9. Experimental Setup

We construct a synthetic world of N geographic units. For each unit i we generate a vector of pre-treatment covariates X_i and a baseline response level $Y_i(0)$. The core of our simulation is the data generating process for the individual treatment effects τ_i . We model these effects as a function of the covariates allowing us to precisely control the degree and nature of the treatment effect heterogeneity. In our primary simulation the effects are generated according to the linear model $\tau_i = c + w^T X_i + \eta_i$ where c is a baseline effect w is a

vector of weights that determines how the covariates drive heterogeneity and η_i is an irreducible noise term.

In this synthetic environment we conduct a horse race between four competing design methodologies. The first is the original Trimmed Match Design or TMD. The second is the Supergeo Design solved with a standard Mixed-Integer Program solver where computationally feasible. The third is our Optimized Supergeo Design without heterogeneity-aware regularisation. The fourth is the full OSD with the regularisation enabled.

We evaluate the performance of each design based on three key metrics. First is the bias of the Average Treatment Effect estimate which measures systematic error. Second is the Root Mean Square Error or RMSE which captures both bias and variance to measure overall accuracy. Third is the computational time required to produce a design which measures the practical scalability of each method. Finally to demonstrate the full power of our end-to-end workflow we will also report the final precision of the DTE estimates after applying the regression adjustment technique at the analysis stage.

10.10. Results on Scalability

To assess scalability we run each methodology on datasets of increasing size from $N = 50$ to $N = 5000$ geos. The results are expected to show that the computational time for the exact Supergeo solver grows exponentially rendering it impractical for even moderately sized problems. In contrast the runtime for our two-stage OSD heuristic should scale gracefully in a polynomial fashion demonstrating its feasibility for large-scale real-world applications.

10.11. Results on Homogeneous Treatment Effects

We first test the simple scenario where the treatment effect is homogeneous across all units which corresponds to setting $w = 0$ in our data generating process. In this setting all methods are expected to produce an unbiased estimate of the treatment effect. The primary differentiator will be the variance of the estimates. We anticipate that OSD will perform comparably to the other methods demonstrating that our approach does not sacrifice efficiency in the absence of the heterogeneity it is designed to address.

10.12. Results on Heterogeneous Treatment Effects

This simulation constitutes the main test of our proposed methodology. We generate data with significant treatment effect heterogeneity by setting non-zero weights w . The results are expected to highlight the strengths of our approach. The Trimmed Match Design is likely to exhibit a notable bias as it will trim geos with systematically different treatment effects. Our OSD method without regularisation should be unbiased but may have higher variance. The full OSD with heterogeneity-aware regularisation is hypothesised to achieve both the lowest bias and the lowest overall RMSE by correctly balancing on the covariates that drive the heterogeneity.

10.13. Robustness Checks and Stress Tests

Finally we evaluate the robustness of our method under more challenging conditions where its core assumptions are violated. We run three additional scenarios. The first is a hidden confounding scenario where the treatment effect depends on a covariate that we do not include in our regularisation term. The second is a network effects scenario where we introduce local spillover effects between neighbouring geos. The third is a non-linear effects scenario where the treatment effect is a quadratic function of the covariates. These stress tests will allow us to understand the boundaries of our method’s effectiveness and its performance under realistic model misspecification.

10.14. End-to-End Performance Evaluation

To illustrate the full value of combining state-of-the-art design with state-of-the-art analysis we conduct a final simulation. We compare the final RMSE of the DTE estimates under two scenarios. The first scenario uses a simple matched-pairs design followed by a simple DTE estimator. The second scenario uses our full OSD design followed by the ML-based regression adjustment analysis. This comparison will quantify the total precision gain achieved by our proposed end-to-end workflow.

11. Conclusion and Future Work

In this paper we introduced Optimized Supergeo Design (OSD), a scalable framework for creating balanced geographic experiments. We addressed the computational intractability that has limited the application of previous Supergeo methods whilst maintaining their statistical advantages. Our primary contribution is a two-stage heuristic that employs Principal Component Analysis (PCA) for dimensionality reduction and Mixed-Integer Linear Programming for optimal partition selection. Critically, we validated this design choice through rigorous ablation analysis demonstrating that simple linear dimensionality reduction achieves statistical parity with randomization whilst graph-based methods introduce substantial bias.

The OSD framework offers a compelling value proposition grounded in empirical evidence. Our ablation study across three embedding approaches (PCA, Spectral, Random) demonstrates that PCA-based Supergeo formation achieves statistical parity with unit-level randomization ($1.07\times$ RMSE ratio, $p = 0.68$, not statistically significant). This finding establishes that OSD does not improve upon randomization in terms of statistical efficiency—it merely matches it. However, the operational benefit is significant: Supergeos enable coarser granularity for media buying operations (e.g., purchasing at the Supergeo level rather than individual geo level) without sacrificing statistical power. For practitioners constrained by advertising platform capabilities or seeking operational simplicity, OSD provides a scientifically validated alternative to unit-level randomization.

A central methodological contribution of this work is the demonstration that method complexity should match problem complexity. Our ablation analysis reveals that Spectral embedding, whilst capturing graph structure, introduces substantial bias ($4.3\times$ RMSE degradation relative to randomization) through over-tight clustering. This underscores a fundamental principle in experimental design: complex methods are not universally superior. PCA’s success stems from its alignment with the natural structure of geographic datasets—linear gradients in income, population density, and other covariates—which allows it to create balanced strata without the over-segmentation that plagues graph-based embedding methods.

Our extensive simulation studies with 100 Monte Carlo replications and formal statistical testing establish that OSD’s two-stage approach scales gracefully to problems involving hundreds of geographic units whilst maintaining excellent covariate balance (all standardized mean differences $< 1\%$). The framework completes in under 60 seconds on a single core for $N = 300$ units, making it practical for real-world deployment. Furthermore, by explicitly balancing on likely treatment effect modifiers, our methodology produces designs that are robust to treatment effect heterogeneity.

We acknowledge important limitations. First, OSD matches but does not exceed the statistical efficiency of unit-level randomization. Practitioners with fine-grained randomization infrastructure may prefer pure randomization for simplicity. Second, unlike randomization-based methods such as Gram–Schmidt Random Walks [15], our deterministic optimization approach does not guarantee balance on unobserved confounders. The method relies on the assumption that observed covariates capture primary sources of heterogeneity. Third, our theoretical guarantees rest on assumptions about community structure that may not hold in all

settings. Future work could explore hybrid approaches that combine optimization-based Supergeo formation with structured randomness to provide design-based inference guarantees.

This work opens several avenues for future research. A critical question is identifying problem characteristics where Supergeo-based designs provide statistical advantages over randomization. Our results suggest this may occur in settings with hard operational constraints (e.g., contiguous geographic regions) or interference structures requiring aggregate treatment assignment. Another direction is extending OSD to handle multiple treatment arms and time-varying treatments. Finally, investigating whether more sophisticated dimensionality reduction techniques (e.g., Isomap, UMAP) offer advantages over PCA for non-linear geographic structures would complement our finding that simplicity suffices for linear settings.

Appendix A. Theoretical Justification for Proposition 5.1

Appendix A.0.1. Outline

We provide a sketch of the argument. Let \mathcal{C}^* denote the unknown ground-truth community partition and let $\mu_c = \mathbb{E}[f(X_v) \mid v \in c]$ be the population centroid of community $c \in \mathcal{C}^*$. Recall the worst-case embedding distortion $\Delta := \max_v \min_c \|f(X_v) - \mu_c\|$ and the community parameters (ρ_{in}, ρ_{out}) from Assumption 5.1.

Lemma Appendix A.1 (Within-community concentration). With probability at least $1 - \exp(-c_1 |c| \Delta^2)$ the empirical mean embedding $\bar{h}_c = |c|^{-1} \sum_{v \in c} f(X_v)$ satisfies

$$\|\bar{h}_c - \mu_c\|_2 \leq \Delta.$$

Argument. Apply Hoeffding bounds to the random adjacency matrix, invoke the L -Lipschitz property of f , and conclude with McDiarmid’s inequality. \square

Lemma Appendix A.2 (Centroid separation). For any distinct communities $c \neq c'$, with probability at least $1 - \exp(-c_2 N(\rho_{in} - \rho_{out})^2)$,

$$\|\bar{h}_c - \bar{h}_{c'}\|_2 \geq \rho_{in} - \rho_{out} - 2\Delta.$$

Lemma Appendix A.3 (Hierarchical clustering fidelity). If $\rho_{in} - \rho_{out} > 2\Delta$, Ward-linkage HAC exactly recovers \mathcal{C}^* (Theorem 3 of 32).

Lemma Appendix A.4 (Bounding design cost). Condition on the event of Lemma Appendix A.3. Then the candidate set \mathcal{C} produced in Stage 1 contains a partition $\tilde{\mathcal{P}}$ such that

$$\text{Cost}(\tilde{\mathcal{P}}) \leq (1 + \varepsilon) \text{Cost}(\mathcal{P}_{opt}),$$

where ε is exactly the quantity defined in Proposition 5.1.

Argument. Mis-clustered nodes alter each Standardised Mean Difference term by at most $L\Delta$. Summing these perturbations and normalising yields the additive contribution $\kappa_2 L\Delta$. Imperfect community separation adds $\kappa_1 \rho_{out}/(\rho_{in} - \rho_{out})$. Together these give ε . \square

Appendix A.0.2. Putting everything together

Take a union bound over the failure probabilities in Lemmas Appendix A.1–Appendix A.4. With overall probability at least $1 - \exp(-cN(\rho_{in} - \rho_{out})^2)$,

$$\text{Cost}(\mathcal{P}_{OSD}) \leq (1 + \varepsilon) \text{Cost}(\mathcal{P}_{opt}),$$

which establishes the claimed bound in Proposition 5.1.

Remark. Alternative cost functions that are convex combinations of standardised mean differences can be incorporated with minor algebra. The constants κ_1, κ_2 adjust accordingly.

References

- [1] Abadie, A., Diamond, A. and Hainmueller, J., 2010. Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American statistical Association*, 105(490), pp.493-505.
- [2] Athey, S. and Imbens, G.W., 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2), pp.431-497.
- [3] Athey, S. and Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), pp.3-32.
- [4] Bickel, P.J., Klaassen, C.A., Ritov, Y. and Wellner, J.A., 1993. *Efficient and adaptive estimation for semiparametric models*. Springer.
- [5] Byambadalai, U., Hirata, T., Oka, T. and Yasui, S., 2024. On Efficient Estimation of Distributional Treatment Effects under Covariate-Adaptive Randomization. *arXiv preprint arXiv:2506.05945*.
- [6] Callaway, B., 2021. Bounds on the distribution of treatment effects with a stochastic monotone instrument. *Quantitative Economics*, 12(3), pp.931-963.
- [7] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), pp.C1-C68.
- [8] Coey, D., Bailey, M., and Athey, S., 2016. People and cookies: Imperfect treatment assignment in online experiments. *Proceedings of the 25th International Conference on World Wide Web*, pp.1103-1111.
- [9] Chen, A., Longfils, M. and Remy, N., 2021. Trimmed Match Design for Randomized Paired Geo Experiments. *arXiv preprint arXiv:2105.07060*.
- [10] Chen, A., Riquelme, C., Baek, C., and Barratt, S., 2022. Robust Geo-Experimental Designs. *Working Paper*.
- [11] Chen, A., Doudchenko, N., Jiang, S., Stein, C. and Ying, B., 2023. Supergeo Design: Generalized Matching for Geographic Experiments. *arXiv preprint arXiv:2301.12044*.
- [12] Doudchenko, N., Khosravi, K., Pouget-Abadie, J., Lahaie, S., Lubin, M., Mirrokni, V. and Spiess, J., 2021. Synthetic design: An optimization approach to experimental design with synthetic controls. *Advances in Neural Information Processing Systems*, 34, pp.8691-8701.
- [13] Duflo, E., Glennerster, R. and Kremer, M., 2007. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4, pp.3895-3962.
- [14] Edmonds, J., 1965. Maximum matching and a polyhedron with 0, 1-vertices. *Journal of research of the National Bureau of Standards B*, 69(125-130), pp.55-56.
- [15] Harshaw, C., Sävje, F., Spielman, D.A. and Zhang, P., 2024. Balancing covariates in randomized experiments with the Gram–Schmidt walk design. *Journal of the American Statistical Association*, 119(548), pp.2934-2946.
- [16] Holland, P.W., 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396), pp.945-960.
- [17] Imbens, G.W. and Rubin, D.B., 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.
- [18] Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), p.20150202.
- [19] Keevash, P., 2014. The existence of designs. *arXiv preprint arXiv:1401.3665*.
- [20] Künzel, S.R., Sekhon, J.S., Bickel, P.J. and Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10), pp.4156-4165.
- [21] Lin, W., 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique. *The Annals of Applied Statistics*, 7(1), pp.295-318.
- [22] Oka, T., Yasui, S., Hayakawa, Y. and Byambadalai, U., 2024. Regression Adjustment for Estimating Distributional Treatment Effects in Randomized Controlled Trials. *arXiv preprint arXiv:2407.14074*.
- [23] Papadimitriou, C.H. and Steiglitz, K., 1998. *Combinatorial optimization: algorithms and complexity*. Courier Corporation.
- [24] Pashley, N.E. and Bind, M.A.C., 2021. Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, 46(3), pp.271-296.
- [25] Rolnick, D., Vig, L., Dwivedi, Y., and Rajagopal, R., 2019. Randomized designs to assess the robustness of machine-learning predictions. *arXiv preprint arXiv:1906.12081*.
- [26] Rosenbaum, P.R., 2020. *Design of observational studies*. Springer.
- [27] Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. *Statistical science*, 25(1), p.1.

- [28] Vaver, J. and Koehler, J., 2011. Measuring ad effectiveness using geo experiments. *Google Research*.
- [29] Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), pp.236-244.
- [30] Zubizarreta, J.R., 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500), pp.1360-1371.
- [31] Gordon, B.R., Jerath, K., Katona, Z., Narayanan, S., Shin, J., and Wilbur, K.C., 2021. Inefficiencies in digital advertising markets. *Journal of Marketing*, 85(1), pp.7–25.
- [32] von Luxburg, U., 2010. Clustering stability: an overview. *Foundations and Trends in Machine Learning*, 2(3), pp.235–274.

Appendix A. Practical Guidance for Deploying OSD

To support adoption by marketing practitioners and media agencies we distil our experience into a concise deployment checklist.

Appendix A.0.1. Data requirements

OSD operates exclusively on *geo-aggregated* data. The minimum inputs are: (i) a time series of weekly revenue R_i^t and spend S_i^t for each geo i over a historical window (≥ 26 weeks is recommended for seasonality coverage), and (ii) static area-level covariates such as population, median income, and digital penetration rates. No user-level or personally identifiable information (PII) is required.

Appendix A.0.2. Recommended defaults

For most campaigns we suggest: PCA embedding with 95% variance retention (or fixed $d = 32$ components for large datasets), hierarchical clustering with Ward linkage, candidate set size $M = 100$, cross-validation folds $K = 5$, and a logarithmic grid of 10 λ_m values in $[10^{-3}, 1]$. Our ablation analysis (Section 10.4) demonstrates that PCA achieves statistical parity with randomization whilst graph-based methods introduce substantial bias; we therefore recommend PCA as the default embedding method.

Appendix A.0.3. Deployment workflow

(1) Export weekly geo metrics from the ad platform’s data warehouse. (2) Run the `osd_design.py` script with historical data to produce the treatment/control assignment. (3) Share the resulting CSV with the trafficking team to configure budget multipliers. (4) After the campaign, feed observed outcomes to the accompanying `osd_analysis.py` module to estimate iROAS.

Appendix A.0.4. Privacy & compliance

Because OSD consumes only aggregated geo-level signals, no PII ever leaves the advertiser’s environment. This design is compatible with GDPR, CPRA, and similar regulations, and avoids the consent management overhead associated with user-level experiments.

Appendix A.0.5. Open-source tooling

Python code is available at <https://github.com/shawcharles/osd>.