



Latest updates: <https://dl.acm.org/doi/10.1145/3746252.3761532>

RESEARCH-ARTICLE

Augmenting Limited and Biased RCTs through Pseudo-Sample Matching-Based Observational Data Fusion Method

KAIRONG HAN, Zhejiang University, Hangzhou, Zhejiang, China

WEIDONG HUANG, DiDi Chuxing, Beijing, China

TAIYANG ZHOU, DiDi Chuxing, Beijing, China

PENG ZHEN, DiDi Chuxing, Beijing, China

KUN KUANG, Zhejiang University, Hangzhou, Zhejiang, China

Open Access Support provided by:

DiDi Chuxing

Zhejiang University



PDF Download
3746252.3761532.pdf
23 December 2025
Total Citations: 0
Total Downloads: 45

Published: 10 November 2025

[Citation in BibTeX format](#)

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
November 10 - 14, 2025
Seoul, Republic of Korea

Conference Sponsors:

SIGWEB
SIGIR

Augmenting Limited and Biased RCTs through Pseudo-Sample Matching-Based Observational Data Fusion Method

Kairong Han

Zhejiang University
Hangzhou, China
zju_handso@163.com

Weidong Huang

Didi Chuxing
Beijing, China

Peng Zhen
Didi Chuxing
Beijing, China
zhenpeng@didiglobal.com

huangweidong@didiglobal.com

Taiyang Zhou

Didi Chuxing
Beijing, China

zhoutaiyang@didiglobal.com

Kun Kuang[†]

Zhejiang University
Hangzhou, China
kunkuang@zju.edu.cn

Abstract

In the online ride-hailing pricing context, companies often conduct randomized controlled trials (RCTs) and utilize uplift models to assess the effect of discounts on customer orders, which substantially influences competitive market outcomes. However, due to the high cost of RCTs, the proportion of trial data relative to observational data is small, which only accounts for 0.65% of total traffic in our context, resulting in significant bias when generalizing to the broader user base. Additionally, the complexity of industrial processes reduces the quality of RCT data, which is often subject to heterogeneity from potential interference and selection bias, making it difficult to correct. Moreover, existing data fusion methods are challenging to implement effectively in complex industrial settings due to the high dimensionality of features and the strict assumptions that are hard to verify with real-world data. To address these issues, we propose an empirical data fusion method called pseudo-sample matching. By generating pseudo-samples from biased, low-quality RCT data and matching them with the most similar samples from large-scale observational data, the method expands the RCT dataset while mitigating its heterogeneity. We validated the method through simulation experiments, conducted offline and online tests using real-world data. In a week-long online experiment, we achieved a 0.41% improvement in profit, which is a considerable gain when scaled to industrial scenarios with hundreds of millions in revenue. In addition, we discuss the harm to model training, offline evaluation, and online economic benefits when the RCT data quality is not high, and emphasize the importance of improving RCT data quality in industrial scenarios. Further details of the simulation experiments can be found in the GitHub repository here.

[†]Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '25, Seoul, Republic of Korea

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2040-6/2025/11
<https://doi.org/10.1145/3746252.3761532>

CCS Concepts

- Information systems → Data cleaning; • Computing methodologies → Causal reasoning and diagnostics.

Keywords

Uplift Model, Random Control Trial, Data Fusion

ACM Reference Format:

Kairong Han, Weidong Huang, Taiyang Zhou, Peng Zhen, and Kun Kuang. 2025. Augmenting Limited and Biased RCTs through Pseudo-Sample Matching-Based Observational Data Fusion Method. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3746252.3761532>

1 Introduction

Ride-hailing, such as Uber [6] and DiDi, is a great example of the creative economy in which owners of private vehicles provide a convenient service to the public for a modest fee [12], which has transformed the way people travel. A well-designed pricing strategy can enhance user loyalty, increase the user base, improve the customer experience, and maximize revenue while minimizing costs [37]. To determine optimal pricing strategies, ride-hailing companies conduct Randomized Controlled Trials (RCTs) [31], also known as A/B tests. Specifically, when users request a ride, companies randomly apply discounts from a set of alternatives and observe user behavior, allowing them to calculate the heterogeneous causal effect of discounts on order rates. This information is crucial for selecting discounts, budget control, cost optimization, and profit calculation.

However, the real-world ride-hailing pricing scenario is highly complex, and the idealized process described above may not perform well in practical industrial applications. Specifically, companies encounter two core challenges:

(1) **The generalizability of a small amount of RCT data is limited.** The target customer base is vast, and the feature space is highly dimensional [22]. And the cost of conducting large-scale RCT experiments is also prohibitively high. In our scenario, the RCT data represents only about 0.65% of the total traffic. Confounding factors such as different scenarios, weather conditions, passenger demographics, and supply-demand relationships all influence the

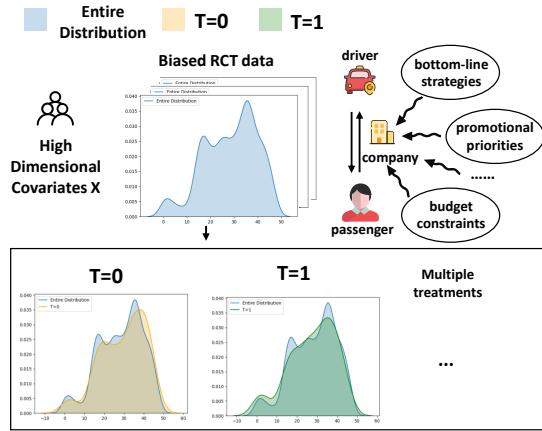


Figure 1: Heterogeneity diagram from real-world data. Complex link interference and potential selection bias lead to biased RCTs. The user's high-dimensional covariate X has different distributions across the different treatment groups T in the biased RCT.

causal effect of discounts on order rates. Consequently, when applied to real-world outcomes, the uplift model trained on this small portion of RCT data exhibits significant bias.

(2) Complex relationships introduce inherent heterogeneity in randomized controlled trials. In real-world scenarios, the experimental connections are intricate, making idealized random experiments impractical, as shown in Figure 1. When conducting RCTs, heterogeneity in the data can arise from several factors, including budget constraints, promotional priorities in specific cities, implementation errors in complex engineering systems, bottom-line strategies, potential changes in population characteristics over time, and even interdepartmental influences within the company. These potential interferences and selection biases result in groups in RCTs that are not entirely random, leading to discrepancies in data distribution across different treatment groups. Consequently, the inherent heterogeneity results in significant deviations in the final trained uplift model.

To address these challenges, data fusion [11] methods are employed to incorporate biased but more comprehensive observational data on population characteristics, thereby enhancing the generalization of RCT data. However, current data fusion techniques face three significant challenges in industrial scenarios: (1) **The assumptions are often overly stringent and difficult to verify with real-world, high-dimensional data.** For instance, most existing methods require positive assumptions and Stable Unit Treatment Value (SUTV) assumptions [39, 43], which are challenging to validate in practical settings. (2) **Traditional approaches treat RCTs as the gold standard.** However, RCT data may exhibit certain biases and heterogeneity due to potential selection bias and uncontrollable interference in complex industrial scenarios. Overlooking these biases may amplify estimation errors in downstream tasks, ultimately affecting the conclusions. (3) **Some depth-based data fusion algorithms are prohibitively expensive and complex.** For example, the CorNet [20] necessitates training an additional correction model for each intervention group. However, in industrial

scenarios, there may be many treatment groups. For example, there are 33 discrete treatments in our ride-hailing. Training a separate correction model for each intervention group is too expensive and difficult to iterate. Moreover, the high dimensionality of the data further poses challenges to existing methods.

Therefore, we propose a data fusion method based on pseudo-sample matching. It simply matches similar samples in observational data by constructing pseudo-samples of biased RCT data, thereby enhancing the effect of the uplift model. We first proved the effectiveness of the method in simulation experiments and then applied the method to offline and online tests on real-world data, which achieved positive results in online experiments. Our contributions are summarized as follows:

- To address the issues of potential heterogeneity and significant generalization bias due to insufficient sample sizes in industrial RCTs, we propose a method for combining RCT and observational data based on pseudo-sample matching.
- We validated and analyzed the effectiveness of the proposed method through simulation experiments, offline and online experiments. In a week-long online experiment, we achieved a 0.41% improvement in profit, which is a considerable gain when scaled to industrial scenarios with hundreds of millions in revenue.
- We discuss the biases present in RCT data within industrial settings and their potential risks. Our method is grounded in real-world scenarios, offering strong practical applicability. With flexible sample bucket and fusion ratio selection, the method can be easily and effectively integrated into a wide range of industrial applications.

2 Related Work

2.1 Uplift Model

Uplift modeling [15, 27, 41] is a predictive technique that estimates the incremental impact of a treatment (e.g., a marketing action) on individual behavior, which is widely used in recommendation systems [4, 17, 18]. It is closely related to the estimation of heterogeneous treatment effects in causal inference [29, 35, 36, 44]. A key distinction, however, lies in their assumptions: uplift modeling methods are typically designed for data from randomized experiments and often omit explicit assumptions. In contrast, methods from the treatment effect heterogeneity literature specify assumptions. [41]. Many machine learning methods have been adapted for estimating heterogeneous treatment effects, such as random forest-based methods [34], Bayesian algorithms [1, 42], and deep learning algorithms [24, 38, 40]. To evaluate the performance of the uplift model, when the Ground-truth causal effect is known, the researchers use the PEHE [21] and MAPE [14] metrics. When the Ground-truth causal effect is unknown, the researchers construct the Qini curve by ranking the sample estimated causal effect and then obtaining the Qini coefficient for quantitative evaluation.

2.2 Combining Randomized Trials and Observational Studies

Although RCT is the golden standard of causal inference and the uplift model [7], the cost of RCT is too high, so it is difficult to

obtain a large amount of RCT data, resulting in a small amount of RCT data and an inability to generalize well to the target population [8, 11]. Therefore, some studies focus on combining RCT and observational data, either to ensure the unconfoundedness of the observational analysis or to improve (conditional) average treatment effect estimation (ATE). When observational data have no treatment and outcome information, to identify the ATE on the target population, most existing methods rely on direct modeling of the selection score previously introduced. The selection score adjustment methods include IPSW [5, 10, 32], stratification [28, 33], and leverage regression formulation. The IPSW estimator of the ATE is defined as the weighted difference of average outcomes between the treated and control group in the trial, to account for the shift of the covariate distribution from the RCT sample to the target population. Stratification is proposed as a solution to mitigate the risks of extreme weights in the IPSW formula. Regression formulation, known as plug-in g-formula estimators [13], fits a model of the conditional outcome mean among trial participants, rather than modeling the probability of trial participation. When observational data have treatment and outcome information, efficiency improvements can be obtained [23], and we could deal with unmeasured confounders in observational data. One way is using an assumption on secondary outcomes or surrogates [2]. The other way is using the bias function to de-bias confounding. For instance, Kallus et al. [25] use observational data to estimate a flexible, but biased function for the heterogeneous treatment effect and then aim to remove the bias using the randomized data. Cheng et al. [9] estimate two separate estimators: one biased estimator on observational data, and another unbiased on randomized data. Hatt et al. [20] introduce a two-step framework. First, they use observational data to learn a shared structure, and then use randomized data to learn the data-specific structures. They named this sample-efficient algorithm CorNet. In this paper, we consider the quality risks of RCT in complex real-world scenarios, while the above methods all use RCT as the gold standard.

3 Empirical Method

3.1 Motivation and Intuition

Given the high cost and limited availability of RCTs, increasing data volume is a practical strategy. Our setting involves 33 treatment groups with varying discount values and slight distributional differences. Directly matching each RCT group to observational data introduces bias at the group level. In contrast, the overall RCT distribution better captures the target population and reflects trade-offs across interventions. Therefore, we represent the high-dimensional data distribution using the mean and apply a bucketing strategy for fine-grained control of key features. Since inter-group differences are small, we introduce a subtle delta adjustment per sample to align intervention groups with minimal distortion. The following section shows details about how our method is conducted.

3.2 Method Description

In the online ride-hailing pricing scenario, our treatments are 33 discrete discount types: $T = \{T_i\}_{i=1}^{N_t}$, where $N_t = 33$. Under a specific discount T_i , the user's price will be $P * T_i \in R$ against the original price $P \in R$. After the user gets the discounted price, they

may choose to take an online ride-hailing service. Therefore, the outcome is a binary variable $Y \in \{1, 0\}$ to mark whether the online ride-hailing order is successfully issued. If the order is successfully issued, $Y = 1$, and if it is not issued, $Y = 0$. In the above process, the enterprise will comply with the user privacy agreement and record a series of desensitized features, such as the time of the current order, the supply and demand relationship of nearby vehicles, and some statistical features. These variables together constitute the high-dimensional feature X .

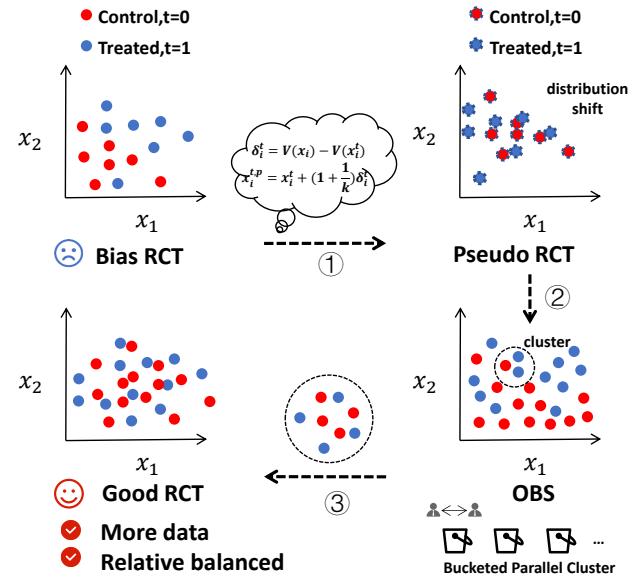


Figure 2: (1) Construct pseudo-samples based on the fusion ratio. (2) Cluster pseudo-samples from the observational data. (3) Add selected observational data back into biased RCT data to obtain good RCT.

To obtain the impact of discounts on order issuance, we use the uplift model to estimate the heterogeneous causal effect of T_i on Y of x compared to T_0 , that is,

$$E(Y|T = T_i, X = x) - E(Y|T = T_0, X = x).$$

However, due to the small amount of RCT data, insufficient generalization, and the inherent heterogeneity of RCT data, this leads to biased results. That means if f is a probability density function, then the probability density function of X is different for different T .

$$\exists T_i, T_j \in T \quad \text{let } f(X | T = T_i) \neq f(X | T = T_j).$$

In our scenario, the RCT data $R = \{(T_i, X_i, Y_i)\}_{i=0}^{N_r}$ is about 0.65% of the observational data $O = \{(T_i, X_i, Y_i)\}_{i=0}^{N_o}$. To constrain the heterogeneity between different treatments and expand the sample size, we will take a part of the observational data as a supplement to the RCT data. Specifically, the overall distribution of RCT data is closer to the target distribution than an individual RCT group. Since our data is high-dimensional and the time complexity of the algorithm needs to be considered in the industry, we calculate the feature mean of the selected j partially important features $X_s \subseteq X$

dimensions for the RCT data to obtain the mean vector V_{avg} of N_r samples:

$$\begin{aligned} V_{avg} &= \left(\frac{1}{N_r} \sum X_1, \frac{1}{N_r} \sum X_2, \dots, \frac{1}{N_r} \sum X_j \right) \\ &= (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_j). \end{aligned}$$

For every RCT treatment group, we obtain separate mean vectors V :

$$V = \{V_1, V_2, \dots, V_{N_t}\}.$$

In data fusion, for each treatment group of RCT, we first calculate the deviation δ between its feature mean vector and the overall distribution mean vector:

$$\begin{aligned} \delta_i &= V_{avg} - V_i \\ &= (\bar{X}_1 - \bar{X}_{1,t=i}, \dots, \bar{X}_j - \bar{X}_{j,t=i}). \end{aligned}$$

So we get a set of bias vectors δ :

$$\delta = \{\delta_1, \dots, \delta_{N_t}\}.$$

According to the preset fusion ratio k , to make the fused mean of the treatment equal to the overall distribution mean, the data mean we need to extract from the observational data of treatment i is:

$$V_i^{obs} = V_i + \left(1 + \frac{1}{k}\right) * \delta_i.$$

Therefore, for each sample in the RCT, we construct a pseudo sample to be fused with the target, that is

$$R_i^{pseudo} = \{(T_i, [X_{s_i} + (1 + \frac{1}{k}) * \delta_i; X/X_{s_i}], Y_i)\}.$$

Then we perform column normalization on the pseudo sample and the observational data so that different dimensions are scaled to the same range to avoid the interference of numerical values in the selection of observational samples. After normalization, a corresponding k data point with the closest feature to the pseudo sample is extracted from the observational data. We use the bucket first and the L2 distance:

$$d_{m,n} = \begin{cases} \infty & \text{if } \text{bucket}_m \neq \text{bucket}_n \\ ||X_s^m - X_s^n||_2 & \text{if } \text{bucket}_m = \text{bucket}_n \end{cases}.$$

In industrial scenarios, several important variables have a huge impact on the final effect of the model. For such variables, we hope to ensure that the heterogeneity of the features in fused data is aligned well between different treatments. Therefore, we will first bucket the observational data and pseudo-sample data according to these variables, and then calculate the similarity of the observational samples and pseudo-samples in the same bucket. In addition, the bucketing operation can make full use of the cluster environment, allowing each bucket to be calculated on a distributed machine, maximizing the concurrency of the algorithm and optimizing the algorithm time under large-scale data. According to the different importance of features and business needs, calculating L2 in the distance measurement can add corresponding weights to different features, so that the fusion process pays more attention to the heterogeneity of these important features:

$$\sum_{k=1}^j w_k * ||X_{s_k}^m - X_{s_k}^n||_2.$$

Through the above process, we can implement the data fusion algorithm at any fusion ratio k , select appropriate features in high-dimensional scenarios, and choose different distance measurement methods to adapt to complex industrial scenarios. The pseudo-code of the algorithm is shown in Algorithm 1. The overall workflow of the method is shown in Figure 2.

Algorithm 1 Pseudo-Sample Matching Algorithm

Require: RCT data R , Observational data O , Fusion ratio k

Ensure: Fused dataset S

Function DATAFUSION(R, O, k)

$$V_{avg} \leftarrow (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_j)$$

$$S \leftarrow \emptyset$$

for $i \leftarrow 1$ **to** N_t **do**

$$V_i \leftarrow (\bar{X}_{1,t=i}, \dots, \bar{X}_{j,t=i})$$

$$\delta_i \leftarrow V_{avg} - V_i$$

$$R_i^{pseudo} \leftarrow (T_i, [X_{s_i} + (1 + \frac{1}{k}) * \delta_i; X/X_{s_i}], Y_i)$$

$$R_i^{pseudo}, O_i \leftarrow \text{normalization}(R_i^{pseudo}, O_i)$$

for $q \leftarrow 1$ **to** number of samples in R_i^{pseudo} **do**

$$d \leftarrow \text{DISTANCE}(R_{i,q}^{pseudo}, O_i)$$

$S \leftarrow S \cup k$'s nearest samples in O_i

end for

end for

return S

3.3 Time Complexity Analysis

Assume that the cluster has enough computing units, and each computing unit can process a bucket in parallel. Given a total of N_b buckets with a relatively balanced distribution of samples across buckets, where the number of RCT samples is N_r and the number of observational samples is N_o , the expected number of RCT and observational samples in each bucket is approximately $\frac{N_r}{N_b}$ and $\frac{N_o}{N_b}$, respectively. For each sample, the time complexity of KNN [19] clustering in the observational data is $O(\frac{N_o}{N_b})$, and the total time complexity is $O(\frac{N_o * N_r}{N_b^2})$. In actual industrial scenarios, the KNN algorithm can also be accelerated using algorithms such as KD trees [3, 16] to optimize the time complexity to $O(N_r * \log N_o)$.

4 Experimental Results

In Section 4.1, we provided a discussion of the evaluation metrics used in our study. Section 4.2 detailed the simulation experiments. Section 4.3 presented the model's performance evaluation on real-world data about offline and online evaluation.

4.1 Evaluation Metric

In the simulation evaluation and offline evaluation, we use Qini coefficients, MAPE, and Click Over Predicted Click (COPC) to evaluate the uplift model. Since there are multiple intervention groups, we calculate the metrics for each intervention group and the control group, then take a weighted average based on the number of online users. Noted that we selected 15 groups in the offline evaluation, accounting for 92% of the total population. Therefore, the optimal COPC is expected to be approximately 0.92. In the online verification, economic benefit serves as a direct performance metric. We

calculate Gross Merchandise Value (GMV), Gross Profit (GP), and AA difference to estimate the overall business profit.

4.2 Simulation experiment

4.2.1 Data Generate Process. We use synthetic data for simulation analysis. To simulate the real scenario of online ride-hailing pricing, we need to follow three keys:

- The scale of observational data is much larger than RCT data.
- RCT data has certain heterogeneity.
- Features are relatively high-dimensional vectors compared to the data volume.

(1) **Covariates, interventions, and outcomes:** We set the feature covariates to 20 dimensions. In real scenarios, the covariates' distribution is complex. For this reason, the simulation data we generated includes binomial distribution, Poisson distribution, normal distribution, uniform distribution, bimodal distribution, and (non) linear combinations of the above distributions, thereby simulating complex real-world scenarios. Details are in the code repository.

According to the covariate X and the binary intervention discount T , we generate heterogeneous causal effects and obtain the binary label Y .

$$Y = \begin{cases} 1 & f_y(X, T) + \delta > \mu \\ 0 & f_y(X, T) + \delta < \mu \end{cases}$$

$$\delta \sim N(0, 1),$$

where

$$f_y(X, T) = a + b \cdot T + c \cdot X \cdot T + d \cdot X^2 \cdot T + e \cdot X + g \cdot X^2.$$

and a, b, c, d, e, g, μ are hyperparameters. Further, we adjusted the hyperparameters to ensure that the proportions of the four categories — Persuadables, Sleeping Dogs, Lost Causes, and Sure Things [26] — are reasonable and similar to the proportions observed in real-world scenarios. For example, when the amount of data is 1000, the ratio is shown in Table 1.

Table 1: The proportions of the four categories: Persuadables, Sleeping Dogs, Lost Causes, and Sure Things. OBS represents observational data.

Categories	$Y(T = 1)$	$Y(T = 0)$	#RCT	#OBS
Persuadables	1	0	237	23455
Sure Things	1	1	183	18683
Lost Causes	0	0	560	56083
Sleeping Dogs	0	1	20	1779

(2) **Generation of biased RCT and observational data:** We use selection bias as the factor that causes heterogeneity in RCT data. Specifically, for different covariates X , the probability of being subjected to different interventions is different, that is, $P(T|X) = f_b(X)$ and $\exists x_i, x_j, f_b(x_i) \neq f_b(x_j)$.

Similarly, we use selection bias as the key factor that causes heterogeneity in observational data. So $P(T|X^{obs}) = f_o(X^{obs})$ and $\exists x_i^{obs}, x_j^{obs}, f_o(x_i^{obs}) \neq f_o(x_j^{obs})$.

Note that the difference between the above two lies in the impact of selection bias on data heterogeneity. In observational data, selective bias has a greater impact, while this difference is relatively small in biased RCT data. We adjust hyperparameters to make the heterogeneity of the observational data larger than the biased RCT data, and make the amount of observational data 100 times that of the RCT data.

(3) **Generate the Ground-Truth:** To effectively assess the true performance of our model, we construct a large-scale, unbiased RCT, 100 times larger than the biased RCT, to serve as the ground truth and gold standard for evaluating the model's true capabilities. In real-world scenarios, it is common to encounter discrepancies where offline evaluation metrics and online performance diverge. This inconsistency may stem from biases present in the RCT data. Unfortunately, in practice, obtaining a truly unbiased RCT for model evaluation is often infeasible. However, with synthetic data, we can simulate an unbiased RCT by making assumptions about the data distribution and the functional relationships among Y , X , and T . By comparing test results using both unbiased and biased RCTs, we can evaluate how imperfections in RCT data quality impact model performance.

4.2.2 Experimental Setting. We use TARNet (TAR) [30] and Causal Forests(CF) [34], which are widely used in industry scenarios as the model for causal effect estimation. We split the generated biased RCT into a training set and a test set in an 8:2 ratio. The data volume of biased RCT is 1000, 5000, and 10000. Correspondingly, the observational data is also 100 times larger than the biased RCT. In addition, we set the fusion ratio to 1, 3, and 5. We use the KD tree to reduce the time complexity to $O(\log N)$. We evaluate ten times and report the mean and standard deviation. The 1:3 fusion ratio is considered the main experiment, and three different fusion ratios (1:1, 1:3, and 1:5) are used to explore the impact of this hyperparameter.

4.2.3 Experimental Results. (1) **Visualization of data heterogeneity.**

We select two features with a data volume of 1000 and a fusion ratio of 3 as an example, shown in Figure 3.

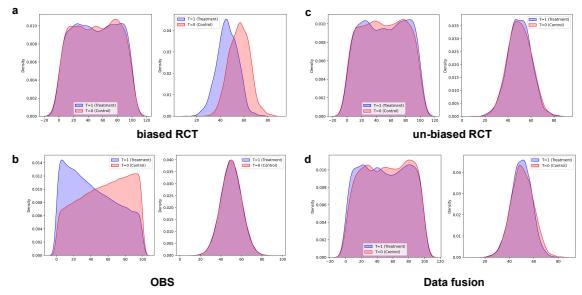


Figure 3: Illustration of data feature heterogeneity. We selected two features. (a) represents the biased RCT data, (b) represents the observational data, (c) is the generated unbiased RCT data at the same scale, and (d) represents the fused data after a 1:3 fusion.

From the analysis chart, we can observe that: (i) When the biased RCT data has no heterogeneity in a certain dimension and large heterogeneity in the observational data, our method can ensure

Table 2: Our method (_f) is compared with the baseline and a control group (_r) where added data is randomly selected in equal amounts from observational data.

Setting	n=1000			n=5000			n=10000		
	Qini	MAPE	COPC	Qini	MAPE	COPC	Qini	MAPE	COPC
TAR	0.222±0.019	0.273±0.089	0.969±0.147	0.259±0.007	0.091±0.037	1.047±0.090	0.299±0.024	0.170±0.101	1.003±0.210
TAR_r	0.222±0.024	0.259±0.125	0.973±0.226	0.295±0.008	0.101±0.069	0.950±0.923	0.327±0.011	0.108±0.089	0.992±0.125
TAR_f(ours)	0.250±0.027	0.219±0.160	0.927±0.150	0.323±0.005	0.078±0.074	0.981±0.101	0.341±0.004	0.057±0.046	1.021±0.073
CF	0.273±0.001	0.140±0.001	1.032±0.032	0.304±0.001	0.115±0.008	1.131±0.009	0.318±0.000	0.062±0.003	1.056±0.002
CF_r	0.296±0.005	0.120±0.001	1.109±0.000	0.325±0.001	0.070±0.000	1.056±0.003	0.328±0.000	0.053±0.000	1.053±0.000
CF_f(ours)	0.316±0.003	0.102±0.016	1.117±0.034	0.330±0.000	0.047±0.003	1.050±0.004	0.334±0.000	0.052±0.004	1.057±0.004

that the fused data will not introduce large deviations due to the differences in the observational data. (ii) When the biased RCT has large heterogeneity in a certain dimension, while the observational data has no heterogeneity in this dimension, our method can effectively correct the bias of the biased RCT data and ensure that the fused data does not have large heterogeneity.

(2) **Results and analysis.** We use the model with a 1:3 fusion ratio as our main experimental setting, with the baseline trained on the original biased RCT data. The control group consists of data randomly selected in a 1:3 ratio from the observational data. Then tested on a large-scale unbiased RCT as the ground truth. The results are shown in Table 2.

Moreover, we tested the model trained under different fusion ratios and explored the impact when the RCT data used in industry is biased. The test results are detailed in Figure 4.

(i) Data fusion can significantly improve the model’s effect on the ground truth. As shown in Table 2, the evaluation metrics outperform the baseline across different data scales.

(ii) Biased and unbiased RCTs yield contradictory trends. As shown in Figure 4, increasing the fusion ratio leads to improved Qini scores under the ground truth, while performance based on biased RCTs fluctuates and deteriorates, exhibiting high variance. MAPE and COPC show the same trend. This divergence underscores the risks of relying on low-quality RCTs, which may produce misleading or even opposite conclusions, emphasizing the critical need for high-quality RCT data in industrial applications.

(iii) Fusion ratio positively influences performance. In the synthetic setting, higher fusion ratios consistently enhance Qini and reduce MAPE.

4.3 Real World Data Verification

To test the effect of our method in real-world scenarios, we conducted the offline and online verification. Specifically, in the online ride-hailing pricing scenario, the RCT experiment accounts for 0.65% of the total online traffic. To train the uplift model, we use RCT data from more than 200 main cities in China collected over one year. And we use the RCT data closest to the current week as the test set. We selected the 1:3 model, which performed well in offline experiments, for online deployment. It was used for one week across 62 major cities in China, during which economic gains were observed.

4.3.1 Offline Evaluation. We conducted data fusion with multiple ratios using the (biased) RCT data points obtained from experiments in the aforementioned 200 cities. The fusion ratios were set to 1, 2, 3,

5, 7, and 9, respectively. During the matching process, we imposed a constraint that required the distance to RCT samples to reach at least a certain threshold. The experimental results are shown in the Table 3.

Table 3: The baseline represents the model trained using the original biased RCT. The ratios from 1:1 to 1:9 indicate different fusion ratios. The (w_) represents the weighted average of the 15 treatments mentioned above based on their population proportions (thus, the theoretical optimal value of w_COPC should be close to 0.92). The gray row is what we chose in the online experiment.

Setting	w_Qini	w_MAPE	w_COPC
baseline	<u>0.0907</u>	0.0441	0.9509
1:1	0.0697	0.0402	0.8385
1:2	0.0967	0.0627	<u>0.9496</u>
1:3	0.0796	0.0308	0.9343
1:5	0.0865	<u>0.0360</u>	0.8718
1:7	0.0891	0.1064	0.7588
1:9	0.0788	0.1396	1.2216

Upon examining the result, we observe that data fusion has produced a negative effect on the Qini coefficient across almost all experimental groups, while it has shown a positive effect on MAPE and COPC at certain fusion ratios. Since our test set is derived from biased RCT data, the results may differ from or even contradict the model’s true capabilities, as indicated by the simulation experiment results in Section 4.2, but it still serves as an important reference for mitigating online risks. The 1:3 setting requires fewer fused data while simultaneously achieving better COPC and MAPE scores. Therefore, through further consideration of the model and cost estimation, we have selected the 1:3 model as the deployment model and conducted online experiments in the next section. In addition, we selected two key features and visualized them, with the results presented in Figure 5. The eta and ecr are crucial for the supply and demand relationship, which plays a significant role in our scenario. As shown in the visualization, the fused data helps mitigate the bias and heterogeneity in the RCT data, resulting in an improved representation of the underlying relationships.

4.3.2 Real World Online Evaluation. Offline evaluation fails to reflect the actual online economic benefits. Therefore, we conducted

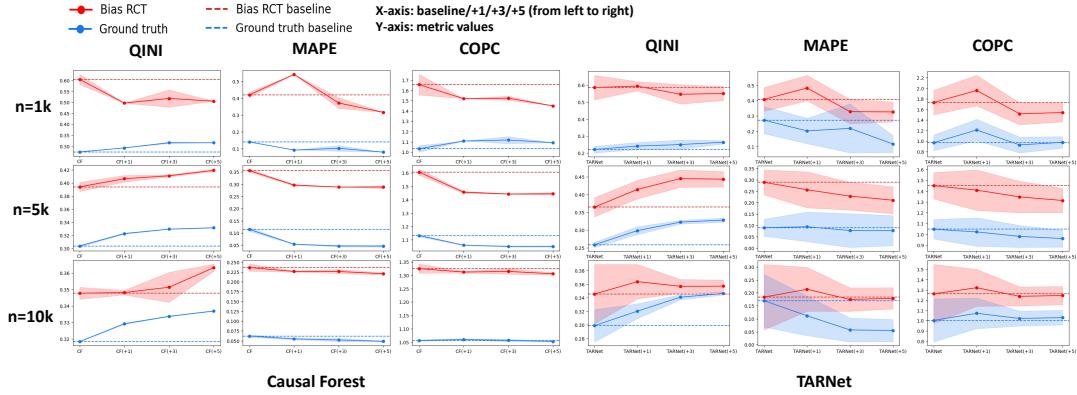


Figure 4: The variation of the Qini, Mape, and COPC under different fusion ratios and using different test sets, where n represents the data size and (+ratio) denotes the fusion ratio. In each subplot, the fusion ratio increases sequentially from left to right, with red representing the biased RCT used as the test set and blue representing the ground truth.

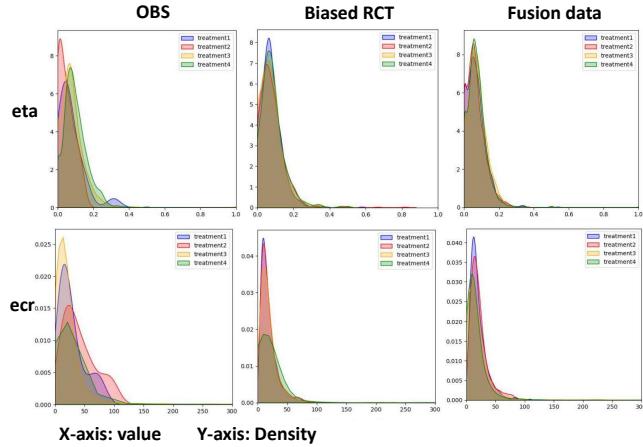


Figure 5: Visualization of heterogeneity in real high-dimensional data. We selected two features. From left to right are observation data, biased RCT data, and our fusion data.

one week¹ of online experiments in 62 major cities in China, with the experimental traffic accounting for 20% of the total traffic. The model trained using the 1:3 fused data was tested to verify whether it could have positive results in an online setting. The model is used to predict users' real-time sensitivity to discounts when hailing a ride online, and the predicted values serve as the basis for the coupon issuance decisions made by the budgeting system. We compared the performance of the fusion model and the baseline model regarding GMV and GP, and calculated the AA difference for the two experimental groups, as shown in Table 4. The model achieved a 0.87% improvement in GMV. Because in large-scale online tests, it is difficult to fully control all variables. Therefore, we rely on the computed adjusted profit as the primary evaluation criterion, considering the GP loss and AA difference, with the baseline as the reference. The profit gain brought by the model is +0.41% on the overall platform, demonstrating the effectiveness of the method.

¹In the ride-hailing scenario, online performance exhibits weekly periodicity, and the large volume of online data ensures statistical significance;

Table 4: The online performance of the model trained on the original data and the 1:3 fused data. We compute and care more about the profit, which corresponds to the gray area.

Setting	GMV	GP	AA	profit
fused data(ours)	+0.87%	-0.19	+0.15%	+0.41%
baseline	0.00%	+0.01	+0.09%	0.00%

5 Conclusion

To address the issues of potential complex heterogeneity and significant generalization bias due to insufficient sample sizes in industrial RCTs, we propose a method for combining RCTs and observational data based on pseudo-sample matching. Our method leverages flexible bucketing and fusion rates to adapt to complex business scenarios. We validate its effectiveness through simulations, offline evaluation, and online experiments. In the simulation experiments, we revealed the harm caused by biased RCTs to model training and evaluation, as well as their impact on economic benefits. Our method performed well on the simulation dataset. Through offline validation with large-scale real-world data and online deployment experiments of the model in multiple major cities in China, our method achieved a 0.41% improvement in economic profit. We introduce a new setting compared to previous studies, aiming to solve real-world industrial problems, and hope that this work provides valuable insights for handling data in industrial applications and emphasizes the critical importance of RCT quality for the uplift model performance.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China (2024YFE0203700), National Natural Science Foundation of China (62376243), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010). All opinions in this paper are those of the authors and don't necessarily reflect the views of the funding agencies.

GenAI Usage Disclosure

We used generative AI, ChatGPT, to check for syntactic and grammatical errors in the manuscript. We carefully verified the correctness of the revised content.

References

- [1] Ahmed M Alaa and Mihaela Van Der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems* 30 (2017).
- [2] Susan Athey, Raj Chetty, and Guido Imbens. 2020. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676* (2020).
- [3] Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (1975), 509–517.
- [4] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.
- [5] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adalora A Adimora, Joseph J Eron, and Michael J Mugavero. 2018. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society* 181, 4 (2018), 1193–1209.
- [6] Ryan Calo and Alex Rosenblat. 2017. The taking economy: Uber, information, and power. *Colum. L. Rev.* 117 (2017), 1623.
- [7] Nancy Cartwright. 2007. Are RCTs the gold standard? *BioSocieties* 2, 1 (2007), 11–20.
- [8] Zhengyu Chen, Teng Xiao, Kun Kuang, Zheqi Lv, Min Zhang, Jinluan Yang, Chengqiang Lu, Hongxia Yang, and Fei Wu. 2024. Learning to reweight for generalizable graph neural network. In Proceedings of the AAAI conference on artificial intelligence, Vol. 38. 8320–8328.
- [9] David Cheng and Tianxi Cai. 2021. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012* (2021).
- [10] Stephen R Cole and Elizabeth A Stuart. 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American journal of epidemiology* 172, 1 (2010), 107–115.
- [11] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. 2024. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science* 39, 1 (2024), 165–191.
- [12] Seth D. Contreras and Alexander Paz. 2018. The effects of ride-hailing companies on the taxicab industry in Las Vegas, Nevada. *Transportation Research Part A: Policy and Practice* 115 (2018), 63–70. doi:10.1016/j.tra.2017.11.008 Smart urban mobility.
- [13] Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. 2020. Extending inferences from a randomized trial to a new target population. *Statistics in medicine* 39, 14 (2020), 1999–2014.
- [14] Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. 2016. Mean Absolute Percentage Error for regression models. *Neurocomputing* 192 (June 2016), 38–48. doi:10.1016/j.neucom.2015.12.114
- [15] Floris Devriendt, Darie Moldovan, and Wouter Verbeke. 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big data* 6, 1 (2018), 13–41.
- [16] Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)* 3, 3 (1977), 209–226.
- [17] Kairui Fu, Zheqi Lv, Shengyu Zhang, Fan Wu, and Kun Kuang. 2025. Forward Once for All: Structural Parameterized Adaptation for Efficient Cloud-coordinated On-device Recommendation. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, 318–329.
- [18] Kairui Fu, Qiaowei Miao, Shengyu Zhang, Kun Kuang, and Fei Wu. 2023. End-to-End Optimization of Quantization-Based Structure Learning and Interventional Next-Item Recommendation. In *CAAI International Conference on Artificial Intelligence*. Springer, 415–429.
- [19] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3–7, 2003. Proceedings*. Springer, 986–996.
- [20] Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. 2022. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891* (2022).
- [21] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.
- [22] Jinping Hu. 2023. Customer feature selection from high-dimensional bank direct marketing data for uplift modeling. *Journal of Marketing Analytics* 11, 2 (2023), 160–171.
- [23] Melody Huang, Naoki Egami, Erin Hartman, and Luke Miratrix. 2021. Leveraging population outcomes to improve the generalization of experimental results. *arXiv preprint arXiv:2111.01357* (2021).
- [24] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.
- [25] Nathan Kallus and Angela Zhou. 2018. Confounding-robust policy improvement. *Advances in neural information processing systems* 31 (2018).
- [26] Victor SY Lo. 2002. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter* 4, 2 (2002), 78–86.
- [27] Diego Olaya, Kristof Coussement, and Wouter Verbeke. 2020. A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery* 34 (2020), 273–308.
- [28] Colm O'Muircheartaigh and Larry V Hedges. 2014. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics* 63, 2 (2014), 195–210.
- [29] Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment* (2010), 39–58.
- [30] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*. PMLR, 3076–3085.
- [31] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. 2004. Randomized controlled trials. *American Journal of Roentgenology* 183, 6 (2004), 1539–1544.
- [32] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society* 174, 2 (2011), 369–386.
- [33] Elizabeth Tipton. 2013. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics* 38, 3 (2013), 239–266.
- [34] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
- [35] Anpeng Wu, Kun Kuang, Mingqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319* (2024).
- [36] Yu Xie, Jennie E Brand, and Ben Jann. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological methodology* 42, 1 (2012), 314–347.
- [37] Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. 2020. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)* 67, 8 (2020), 705–724.
- [38] Jinluan Yang, Zhengyu Chen, Teng Xiao, Yong Lin, Wenqiao Zhang, and Kun Kuang. 2025. Leveraging invariant principle for heterophilic graph structure distribution shifts. In *Proceedings of the ACM on Web Conference 2025*, 1196–1204.
- [39] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.
- [40] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*.
- [41] Weijia Zhang, Jiayong Li, and Lin Liu. 2021. A unified survey of treatment effect heterogeneity modelling and uplift modelling. *ACM Computing Surveys (CSUR)* 54, 8 (2021), 1–36.
- [42] Yao Zhang, Alexis Bellot, and Mihaela Schaar. 2020. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1005–1014.
- [43] Ziyu Zhao, Yuqi Bai, Ruoxuan Xiong, Qingyu Cao, Chao Ma, Ning Jiang, Fei Wu, and Kun Kuang. 2024. Learning individual treatment effects under heterogeneous interference in networks. *ACM Transactions on Knowledge Discovery from Data* 18, 8 (2024), 1–21.
- [44] Ziyu Zhao, Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, Zhihua Wang, and Fei Wu. 2024. Networked Instrumental Variable for Treatment Effect Estimation With Unobserved Confounders. *IEEE Transactions on Knowledge and Data Engineering* (2024).