Developing a Cutting Edge Control Group

Richer, More Flexible, More Powerful
INTRODUCTION

At Cuebiq, we continually evolve our technology to provide our customers with powerful solutions to understand how marketing dollars drive in-store visitation. We have made especially big changes during the last two years. Two core elements in particular have been developed and implemented in our new platform:

1. The first one is an innovative methodology to measure incremental visits. It relies on advances in causal machine learning to create a comparison of the exposed group with themselves had they not been exposed\(^1\). This new methodology powers our core product in Cuebiq’s measurement platform by comparing the exposed group with a synthetic version of itself with no exposure. We have the know-how to provide daily updates (almost real-time) to our clients showing the evolution of different metrics and KPIs related to campaign performance. This is an industry game-changer.

2. The second innovative element in our platform is the ability to perform these ongoing performance tests throughout a wide array of different media channels. We are now tracking complex campaigns across multiple channels (digital, CTV, linear TV, out-of-home, radio, print) and generating insights for our customers through a wide array of messaging outlets. From a technology stack point of view, this has required a constant evolution of the way we think about our product and our commercial relationships with our partners, and has forced us to adopt more sophisticated technologies and methods. This has been quite a journey for Cuebiq since the old days of analyzing campaigns only in the digital realm.

Our measurement platform is faster, more powerful, and more precise than it ever was.

There are also new learnings from this process. In a multitouch/multichannel environment we quickly realized that each media channel represents its own set of challenges from a measurement point of view. The way we conceptualize and measure attribution in a digital channel does not necessarily translate directly into a different channel—for example out-of-home. In the discussion below, we exemplify some of these differences in detail. Basically, each media channel has its own logic, its own issues, its own advantages and disadvantages. Our technology stack for multi-channel campaigns grew, and our methodology evolved.

But this big leap forward also revealed important areas of opportunity for one of our core building blocks: the control group. As we learned from more channels and more campaigns, it became clear that our methodology needed a more flexible and powerful algorithm to create a control group that could serve the more nuanced complexities of a multi-channel environment.

Today we have a flexible, evolved, and modern algorithm to create the control group that lives up to the task of providing a baseline for behavior in the exposed group. During our journey the past two years we found that multiple channels needed a different approach and the new control group is the best common denominator.

This document presents an overview of the new control group that we have designed and implemented. It starts with a definition of the problem, then Cuebiq’s solution, and finally a technical appendix for those interested in more methodological detail.

\(^1\) Called technically a counterfactual.
**WHAT IS A CONTROL GROUP?**

Marketers, agencies and their clients want to understand a campaign’s effect on a desired outcome—sales, visits, vaccination rates, etc. Rather than just observing whether the exposed group makes the desired behavior, the exposed devices need to be compared to a different set of unexposed devices. This second group serves as a behavioral baseline and is called the *control group*.

In ideal settings the control group will be statistically equivalent to the exposed group in all observed features—indistinguishable from the exposed group except for the treatment condition—in this case, ad exposure. It will look and behave like the exposed group *had it not been exposed*, so in theory, any differences in the behavior we are interested in may be attributed to exposure alone. It is a hypothetical construction, with random assignment at its core.

In practice, a properly built control group is said to function as a *counterfactual* to the exposed group. Once the control group is created, inferring causal effects from exposure is simply a matter of comparing averages across control and exposed groups over the desired outcome.

Unfortunately, exposure in advertising campaigns is not allocated randomly\(^2\). This opens the possibility for many potential biases in our assessment of campaign effectiveness—which usually result in overestimating campaign influence.

Therefore, alternative methods must be designed to generate control groups that fulfill as close as possible the original utility of random assignment: creating a baseline group that is fundamentally comparable to the exposed group.

A key driver in our product design and methodological approach is to take seriously the biases that arise when ad exposure is not the result of a proper A/B test. Since we know ad exposure is generally not randomly assigned, linking advertising to behavioral changes is rife with biases and confounding variables that likely generate misleading conclusions about campaign effectiveness.

We take these issues seriously and put them at the center of our product. With this mindset, we have reworked from the ground up our definition of *control group* (a core behavioral baseline to truly understand how ad exposure is moving the needle), as part of a broader effort to design measurement systems that mitigate as much as possible the biases and noisy measurements inherent to advertising and marketing.

Our goal is to continually improve an end-to-end measurement system that taps into state of the art research and development at all stages of the process (control group, incrementality estimation, actionability) to provide our clients with the best solutions in the market to understand how advertising dollars are changing offline behavior\(^3\).

---

\(^2\) Except in walled platforms such as Facebook or Twitter where they have full control over who gets assigned which ad and who doesn’t.

\(^3\) Please refer to Cuebiq’s white paper *Visit Incrementality: The Key To The Future Of Measurement* which you can find [here](#).
HOW DID WE UPGRADE OUR METHODOLOGY?

In order to understand how a campaign is moving the needle, we need to compare the behavior of the exposed units with a proper baseline, which should be statistically equivalent to the exposed group — aka the control group. This means that, except for the fact that they have not been exposed, the control and exposed groups should be indistinguishable from one another. While this is achieved through randomization, we must resort to building a control group that satisfies the same properties when randomization is not possible. A standard approach is to match each user in the exposed group with a similar user from a pool of control candidates. The matching procedure is the key part of this process; the goal is to reproduce the effect of randomization by balancing out anything but the exposure.

The key then becomes defining the concept of similarity between exposed and control units. Notice that since we are interested in whether a user visits a point of interest during the campaign (i.e., if the user converts), the only thing that we need to balance is the probability (before being exposed) of converting. Hence, we say two users are similar if they have the same probability of visiting such stores prior to the exposure. This is a key concept:

*The ideal is to match an exposed user to a control one if they have the same probability of conversion before the exposure.*

However, it is impossible to exactly compute the probability of conversion of a user. Instead, we use a proxy: We select all features that affect the probability of conversions (for instance, how far from the store the user lives or the frequency of past visits) and we balance these. The rationale is, if two users (one exposed and one control) are similar in all factors affecting the probability of visiting, then they are a perfect match as they are balanced in all relevant variables. Adopting this strategy for building a control group guarantees that we reach the desired statistical balance.

At this point it should be clear that, from a modeling standpoint, the choice of the features is key to a good control group. We use a trove of numeric and categorical variables to describe the devices in our panel, all following two guiding criteria.

1. First, they have to be related to the probability of visiting the stores. For example, including hair color as a feature when evaluating Burger King’s campaign would not be a relevant choice.
2. Second, and more importantly, we must try to capture the most important factors influencing such probability. Imagine if we did not include the geo-location of a user when evaluating a campaign in Chicago; the control group might have users located all over the U.S. Clearly, users in a different state have a much lower probability of visiting stores in Chicago, and therefore the campaign impact would be overestimated.

There are two keys to building a good control group:

A. Choose features that describe the likelihood of a user visiting the target store, regardless of the campaign.

B. For each exposed user, find a control user that is as similar as possible (as defined by the selected features).
Our approach leverages our expertise in location data and focuses on the behavior of the users rather than on their socio-demographic attributes. This is because the matching of users should not be based on whether they have the same gender, age, or income, when what really matters is their behavior. The question we are asking is: “Are these devices as likely to visit the targeted store?” We answer this question by looking at the history and behavior of the user rather than socio-demographic characteristics.

Moreover, socio-demographic data is error-prone: It is very easy to misclassify a user’s gender, age, or education. In contrast, geo-location data is more factual and requires substantially less modeling because it is generated directly from the device.

To reach its goal, our control-group algorithm uses several attributes that quantify:

- The range of mobility of a user (how much people are moving around)
- The average number of points we received in the past seven days
- How much we see the user throughout the day, i.e., the portion of the user’s day that is covered by our data
- The home-store (and office-store) distance from the measured location
- The county of the home location
- Whether or not the user visited the campaign’s brand stores previously
- Whether or not we saw visits from the user in the recent past
Markets evolve. And we evolve with them. These new attributes and methodologies describe a device with a richer depth than ever before, and they do so in the space that Cuebiq knows best: device mobility. We are leveraging more features to select which users in our panel will be matched to the users in the exposed group, and we do this across all channels to maximize the flexibility of our methodology.

The control group lies at the heart of our effort to empower marketers to understand what is moving the needle. From a methodological standpoint, a control group is ground zero to a host of services and methodologies downstream that use it as input: multichannel/multi-touch-point attribution, incrementality, etc. A richer, more robust and flexible definition of a control group is part of the overhaul and evolution of our platform, the market, and our client’s expectations and needs. It fits neatly within a new market stage in which more sophisticated agencies and clients require more sophisticated and flexible solutions, all to make sense of a marketing world with ever-increasing complexities in terms of channels, touch points, mixes, and strategies.

It will also serve as a new framework to face an evolving technical, social, and political landscape in which privacy concerns, legal requirements, and technical advancements interact to bring new conditions that affect marketers and agencies alike. New challenges in the form of reduced granularity in a new cookie-less world will force a transformation of our worldview and our methodologies and technologies to deliver timely attribution assessments to our clients. Like this, other—yet unforeseen—challenges will arise in the future. Such is the nature of our market. Having a flexible methodology to build the core element of a wide menu of downstream marketing solutions will only benefit our clients—and their clients.
In the old control group, all features had only a limited number of possible values, whereas the new features are continuous. For example, the number of active days in the previous week can only have values from 0 to 7. So, in the old version, we could do exact matching on the features (e.g., match users having the same value for the number of active days). Oppositely, it is not possible to do exact matching on, e.g., home-store distance, because likely there is no pair of users with exactly the same value. Hence a new algorithm is needed.

The idea is simple: First, divide all users (exposed and candidate controls) into buckets based on the most important features—home location and whether or not the user has visited the geo-set recently. An exposed user can only be matched to a candidate control within the same bucket; therefore, we only compare users within the same bucket. To ensure a close match, we need to mathematically define the users’ similarity by defining a “distance” between users. We define such a distance as the sum of all features’ differences. This distance can be thought of as geographical distance, where the closer the users are, the more similar they will be. Therefore, to pick the best match to an exposed user, we simply choose the closest.

Even if the idea is quite simple, there are a few things that are worth noting.

FEATURE TRANSFORMATION AND NORMALIZATION

Before matching, we apply different transformations to each of the features, as different features need to be transformed in different ways to adjust their scale. For example, we take the logarithm of the home-store distance.

The second important step is the normalization of the features. Normalization can be thought of as an elimination of the scale of a variable. The rationale is that, as different variables have different units of measure, and therefore different scales, in order to compare them, we want to eliminate the scale. Normalization is done within each bucket of users. This simply amounts to considering each bucket separately and dividing all user variables by their mean (computed over the users in the bucket).

Note, it is important that this is done separately over each bucket, as we want to eliminate the scale within a single bucket. For example, suppose we have two buckets, one associated with Manhattan and one with a rural county in South Dakota. Let’s also assume there is a store for our campaign in both areas. Now, consider the home-store distance feature. This will have different scales for these two buckets (small for the former and large for the latter). If we do not consider the two cases separately, the mean will be some number in the middle. Dividing by this would make the feature’s values, respectively, too small and too large.

---

4 This is a standard approach in many artificial-intelligence problems. We define an ad-hoc distance and therefore a new concept of space, where closer means more similar.

5 That is because small differences in low values are more meaningful than big differences in high values. In other words, it is a lot different to live 1 or 10 kms away from a store. But it does not make much difference whether you live 200 or 300 kms away.
Consider the two following problematic situations:

1. A bucket contains only a few users. If enough are exposed, there may not be available options for control users, or most of the control candidates may be very different and thus lead to bad associations.

2. Within a bucket, all exposed users have high home-store distances, while most of the candidate controls have low values for the same feature. In this case, once the better candidates are selected, we are stuck with low-quality candidates.

In order to solve these potential problems, we allow control users to be chosen more than once. If a control user is picked several times, the duplicates will be considered as different users when it comes to computing RR and uplift. As an example, a control user associated with 3 different exposed will count as 3 different controls, and if he converts will count as 3 different control conversions. This is fundamental in order to provide unbiased RR and uplift.6

However, we set a limit on the number of times a control user can be matched. If no limit is imposed, a control user may be picked too many times. This would make the RR too dependent on conversions of this user—a conversion will be counted as many times as the user was duplicated. Oppositely, no conversions means all replicas of the user did not convert. Therefore, one actual user would have a big impact on the control group’s visit rate, which significantly increases the variance.

---

6 This is because it is important to keep the ratio of likely-to-convert over unlikely-to-convert users approximately the same in control and exposed groups. If, for example, there are many more duplications in likely-to-convert control users than unlikely-to-convert ones, we need to consider duplicates as different users to keep the ratio the same.
Validating a control group is a hard task. One could be tempted to check the quality of the matching — that is, how different the features' values are for each matched pair of exposed and control. However, this only measures how well the matching procedure works, not how good the choice of features is.

One powerful test is to consider the exposed and control group and analyze the distribution of the RR\(^7\) over an independent set of stores. For example, suppose our campaign is run for a QSR brand. First, we use our algorithm to build the control group. Then, instead of using the campaign's stores, we use some stores taken from unrelated brands (for example: jewelry or clothing brands) to determine conversions. Since we can assume that being exposed to a fast-food campaign does not change the probability of visiting a jewelry or clothing store, the exposed and controlled users should have the same probability of visiting such stores. Consequently, the expected value for the RR is 1, validating that no bias was introduced while building the control group.

Going into a little more detail, the distribution of the RR variable over a set of independent stores should be a log-normal distribution with mean equal to 1. This is the same as saying that the distribution of the logarithmic of RR should be a bell curve. These concepts are visualized in the figures on the following page, where we plot (for two campaigns taken as example) the estimate of the mean of \(\log(\text{RR})\) (in black), the estimate of distribution of \(\log(\text{RR})\) (in red) and the desired distribution (in blue).

To summarize, the red line should be as close as possible to the blue line. However, the most important things to check are the following:

1. The black dotted line, representing the mean of \(\log(\text{RR})\), should be as close as possible to zero. This means the expected value of RR is 1.
2. The red line to be symmetrical with respect to zero.
3. The red line should decrease smoothly without abnormal bumps. This means that there is no hidden bias showing only for some sets of stores.

---

\(^7\) We recall that the RR measures the strength of association between exposure and conversion. When RR equals 1, exposure has no impact on conversion. An RR that is higher than 1 indicates that exposure positively influences conversion. Vice versa, an RR lower than 1 indicates that exposure negatively influences conversion.
Note that, obviously, there is a slight asymmetry in the red line, and the black line does not fall exactly on zero. However, this is not enough to denote a bias, especially considering that the average of log(RR) is very close to zero. Given that the asymmetry is very small and the black dotted line falls very close to zero, we can say that the results validate the control group. This test is also used by Facebook for evaluation of their methodology (Gordon et al., 2019).

Bibliography