AMA AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing Research
2022, Vol. 59(4) 755-774

DOI: 10.1177/00222437211059088
journals.sagepub.com/home/mrj
SAGE

Bayesian Consumer Profiling: How to Estimate Consumer Characteristics from Aggregate Data

Arnaud De Bruyn and Thomas Otter

Abstract

Firms use aggregate data from data brokers (e.g., Acxiom, Experian) and external data sources (e.g., Census) to infer the likely characteristics of consumers in a target list and thus better predict consumers' profiles and needs unobtrusively. The authors demonstrate that the simple count method most commonly used in this effort relies implicitly on an assumption of conditional independence that fails to hold in many settings of managerial interest. They develop a Bayesian profiling introducing different conditional independence assumptions. They also show how to introduce additional observed covariates into this model. They use simulations to demonstrate that in managerially relevant settings, the Bayesian method will outperform the simple count method, often by an order of magnitude. The authors then compare different conditional independence assumptions in two case studies. The first example estimates customers' age on the basis of their first names; prediction errors decrease substantially. In the second example, the authors infer the income, occupation, and education of online visitors of a marketing analytic software company based exclusively on their IP addresses. The face validity of the predictions improves dramatically and reveals an interesting (and more complex) endogenous list-selection mechanism than the one suggested by the simple count method.

Keywords

consumer profiling, data augmentation, data brokerage, Bayesian profiling, sociodemographic profiling, geolocation

Online supplement: https://doi.org/10.1177/00222437211059088

Consider the following three scenarios:

Scenario 1. Yahoo! "Smart Billboards" (Chen and Strimaitis 2016) will deliver targeted ads, for instance, by identifying the make and model of vehicles passing by on a highway and using this information to profile customers. From the patent application: "Demographic data (e.g., as obtained from a marketing or user database) for the audience can thus be determined for the purpose of, for example, determining whether and/or the degree to which the demographic profile of the audience corresponds to a target demographic." Because Honda drivers are known to be significantly more educated than Chevy owners (O'Malley Greenburg 2009), if several Honda cars are spotted on the highway, ads tailored to highly educated individuals would be more effective.

Scenario 2. In a real-time bidding application, where advertising inventory is bought and sold on a per-impression basis via programmatic auctions (see Geraghty et al. 2017), a company wanted to target online visitors with superior health coverage. Although that particular information was hard to obtain at the individual level, online visitors were assigned to one of the 500 online Merkle's DataSource segments (see Table 1), and their likelihood of health coverage was inferred from the average health coverage of the segment to which they were assigned.

Scenario 3. A charity wished to target donors between 50 and 65 years old to inform them about their "donation by will" program. Prior research found that donors were most likely to plan their succession during that period of their life. Despite the charity's extensive information about the donation history of its donors, their actual age was unknown. The charity acquired a reference table with the age pyramids of all first names in the country and partly targeted its donors based on that information.

While different, these three applications rely on consumer profiling, which we define as the process of inferring the

Arnaud De Bruyn is Professor of Marketing, ESSEC Business School, France (email: debruyn@essec.edu). Thomas Otter is Professor of Marketing, Goethe University, Germany (email: Otter@marketing.uni-frankfurt.de).

Table 1. Selected Managerial Illustrations Where Firms Infer Individual Variables from Aggregate Data (i.e., Various Reference Tables).

Census data	In the United States, zip codes can be matched with the U.S. Census Bureau data to qualify customers' profiles in terms of their likely age, race, family relationships, household types, educational attainment, marital status, employment status, or income. Such data are integrated into all the major data brokers' databases.
AcquireWeb	AcquireWeb (http://acquireweb.com) is managing a list of 190+ million IP addresses pinpointing to zip + 4 data to target people who live in neighborhoods with specific demographics.
Google	Google infers[a] the most likely age, gender, and parental status of online visitors based on the websites they visit, as long as these websites belong to the Google Display Network. Although Google does not share these data outside the company, it uses them to customize online analytics reports at the aggregate level (such as reporting conversion rates by age or gender) or to fine-tune ad targeting.
Yahoo!	Yahoo! filed a patent application (Chen and Strimaitis 2016) describing "Smart Billboards" that would deliver targeted ads, for instance by identifying the make and model of vehicles passing by on a highway and using this information to profile customers.
Experian	Experian's Mosaic USA "is a household-based consumer lifestyle segmentation system that classifies all US households and neighborhoods into 71 [segments]" (Experian 2014), describing each in terms of their sociodemographics, habits, lifestyles, behaviors, and culture.
Acxiom	Acxiom divides the U.S. population into 70 Personicx segments (55 in the United Kingdom,[b] 32 in France). The company provides an online tool[c] to predict to which cluster a U.S. citizen belongs based on their demographics. Once a customer is assigned to a cluster, other characteristics are inferred from the 600+ variables available at the segment level, such as propensity to own a business credit card, to travel often, to be well educated, or to read the business press.
Merkle	Merkle's DataSource classifies online visitors into 500 digital segments, and segment membership is then used to profile customers over 425 variables regarding demographics, wealth, and lifestyle, as well as indicators from the Population Census.
Conexance	Conexance (http://www.conexancemd.com) classifies mailing addresses at the city-block levels into 25 behavioral typologies (e.g., "fragile elderly," "dynamic suburbs") that translate into typical family composition, taxable income, and purchase indices on various product universes.
CACI	CACI, a data broker in the United Kingdom, classifies 49 million adults in the country into 50 distinct PeopleUK consumer segments. Segment-level data originate from over 15 million lifestyle surveys, but also from aggregate data such as the Electoral Roll and Census data.
Weborama	Weborama deploys smart displays in airports. These smart displays scan the environment and adapt ads based on the inferred sociodemographic of passing-by travelers.

[a]"Sarah's favorite hobby is gardening. Many of the gardening sites and blogs on the Display Network that she visits have a majority of female readers. Because of this, Sarah's browser could be added to the 'female' demographic category. As a result, Google may show Sarah ads from advertisers who have chosen to show their ads to women." See https://support.google.com/adwords/answer/2580383?hl=en.
[b]For an interesting presentation of Acxiom's U.K. Personicx segmentation, visit http://www.personicx.co.uk/personicx.html.
[c]See https://isapps.acxiom.com/personicx/personicx.aspx.

average profile and individual characteristics of a target list from aggregate data. In the examples, people's level of education (respectively, health coverage, age) is inferred from an aggregate reference table, using their car model (respectively, segment membership, first name) as key.

Numerous companies, referred to as "data aggregators" or "data brokers" (e.g., Acxiom, Corelogic, Datalogix, eBureau), have flourished recently to provide companies with individual-level data about their customers, prospects, and online visitors (Federal Trade Commission 2014). These data are, in turn, used either to *profile* a group of customers (e.g., for marketing communication or positioning purposes) or to *target* specific individuals.

However, looking past the media hype and press releases claiming that "Big Data knows everything about you" (Weisbaum 2014), the truth is that accurate data at the individual level are scarce and not as easily accessible as one may believe. Data augmentation through inferences from aggregate data remains widely used in practice.[1]

In this article, we show how to improve on traditional consumer profiling, which relies on simple counts and ratios and may lead to misleading inferences about the distribution of interest. We decipher the underlying *implicit* selection assumption behind this simple count method and show that other selection assumptions are both possible and potentially more appropriate. We propose Bayesian profiling as an alternative inferential framework and significantly outperform the simple count method in many managerially relevant settings.

We organize the remainder of this article as follows: First, we discuss the prevalence of consumer profiling using aggregate data and why it remains of high managerial relevance, even in a world of highly granular information where individual consumer data are widely available. We cast our developments in the larger context of endogenous selection, data augmentation, and data fusion. We then theoretically develop the various possible model alternatives and contrast them with the benchmark approach. To illustrate the models, we estimate consumers' income using information from the car model they drive. Then, we run simulations and show that under managerially relevant circumstances, the Bayesian approach will outperform the simple count method, sometimes by an order of

[1] We use the term "data augmentation" to refer to all actions that result in information about variables that are missing from a data set and not just to the specific numerical technique commonly employed in Bayesian inference.

magnitude. Next, we estimate the most likely age of a list of customers identified only by their first names; the Bayesian approach leads to much-improved estimates. Subsequently, we demonstrate how a software company unobtrusively profiled its online visitors from their IP addresses and compare various models. We report that the common simple count method is inferior to numerous alternatives. We conclude with policy implications and managerial recommendations.

Managerial Relevance

Suppose a company is interested in learning more about some individuals—be they existing customers, prospective customers, or online visitors. This learning effort may be warranted because the company has little information about these individuals in the first place (e.g., anonymous visitors on a website). It may be equally justified when information such as customers' past purchases or browsing behavior is already available at the individual level but is insufficient or not relevant for the problem at hand.

If this information is not readily available, and collecting data directly from customers is not an option (because it would be too slow, expensive, or intrusive), the firm could achieve its goal through data augmentation, enriching its customer list with personal information provided by data brokers such as Acxiom, Merkle, Epsilon, or Experian. Such data augmentation can be obtained from either individual or aggregate data.

Data Augmentation from Individual Data

Enriching a target list with individual data is impeded by three challenges:

1. *Data may not be available from data brokers.* Merkle Inc., one of the largest data aggregators in the United States, claims that it "captures information on 129MM households and 275MM individuals over 2,500 detailed attributes" (Merkle Inc. 2019) about demographics, wealth, lifestyle, vehicles, and consumer habits. But the number of indicators available varies greatly, with an average of 15.5 records per household. During an interview, a former director at a major U.S. data broker firm invited the authors to "think of [data brokers' databases] as a huge Swiss cheese … with a lot more holes than cheese."
2. *Data may not be matched with data brokers*. Even when data are available at the individual level in the data broker's database, they might not be matched easily[2] if the contact's information is inaccurate, incomplete, or outdated on either end of the transaction.
3. *Privacy laws*. In light of recent privacy laws (e.g., General Data Protection Regulation [GDPR], California Consumer Privacy Act) and, for example, Google's decision to phase out third-party cookies (Google 2021), it will be increasingly difficult for marketers to collect, keep, and exploit individual data.

[2] As anecdotal evidence, the authors interviewed the digital manager of a major European bank that tried to match its customer database with Facebook profiles. Despite Facebook's promise of an 80% match rate, and a large sample of 400,000 customers submitted as a test, only 10% of the bank's customers could be matched with existing Facebook profiles.

Data Augmentation from Aggregate Data

When individual data are not available, marketers can infer the information of interest from aggregate data. The Federal Trade Commission (2014) reports that "data brokers infer consumer interests from the data that they collect. They use those interests, along with other information, to place consumers in categories. Some categories may seem innocuous…. Potentially sensitive categories include those that primarily focus on ethnicity and income levels, … consumer's age, … and health-related topics or conditions." (p. 5)

However, a recent study comparing programmatic segment description (i.e., how individuals are first classified into segments and then described based on the segments to which they belong) and actual individuals' characteristics reports that "audience segments vary greatly in quality and are often inaccurate across leading data brokers" (Neumann, Tucker, and Whitfield 2019, p. 918). This lack of precision stems from two distinct issues: correct *segment assignment* (how to assign an individual to a specific segment based on limited, observed data) and accurate *consumer profiling* (how to describe an individual consumer or a collection of consumers based on data obtained from aggregate, segment-level data). This research focuses on the latter issue.

The Shortcomings of Data Augmentation from Aggregate Data

The process of estimating individual-level data from aggregate reference tables is quite common, as evidenced by the illustrations listed in Table 1. We argue that in many managerially relevant situations, customers being profiled are not a random selection of their category or segment. For instance, the owners of Lexus cars spotted by Yahoo!'s Smart Billboards on U.S. Route 101 (the highway that connects Silicon Valley with San Francisco) on a Monday at 7:00 A.M. are unlikely to be a random draw from the more general population of Lexus car owners. Similarly, an inhabitant of Aberdeen, Massachusetts, connecting to the Steam online game network, is unlikely to match the demographic profile of the average Aberdeen resident. Often, consumers are observed in "target lists" that are *not* random selections from their category-level populations. We purposely use the term "target list" to draw attention to the nonrandomness of the selection process. In their modeling efforts, analysts should select appropriate conditioning arguments to account for that nonrandomness. We

subsequently demonstrate that different list-selection specifications lead to dramatically different model accuracies.

In this article, we argue that the standard inference approach used in the industry *implicitly* assumes a very specific, nonrandom list-selection mechanism that is unlikely to hold in practice. This methodological approach leads to a form of ecological fallacy that may generate substantial biases. We then demonstrate that *explicitly* modeling different list-selection mechanisms in the likelihood function can significantly improve prediction accuracy.

Literature on Endogenous Selection

This research falls under the general class of selection based on unobserved variables, sometimes referred to as self-selection or endogenous selection. Selection based on unobservables—tracing all the way back to Heckman's (1979) seminal contribution—is a well-established phenomenon. A famous example is the self-selection of employees into training programs. Employees perform well in after-training evaluations not because the training helped, but because high-performing employees were more likely to enroll in the training program in the first place. The advantage from participating in the training does not generalize to other employees. In most applications, the goal is to overcome the biasing influence of correlated unobservables on structural parameters of interest, which in turn inform counterfactual queries (e.g., Wachtel and Otter 2013). However, the specific nature of the unobservables causing the problem is beyond the scope of this literature, which instead aims to achieve data-based identification of target parameters (e.g., the causal effect of participating in a training program) based on minimal assumptions about unobservables while plausibly correcting for their biasing influence (e.g., with instrumental variables). Broadly speaking, the goal of this literature is to leverage information in endogenously selected data to learn about (relationships in) the general population.

Our applications and methodology, instead, focus on inference on unobserved variables on the selection path. We detail how to harness a model of selection based on variables that are missing entirely from the target list to learn about the distribution of these very same unobserved variables in the target list under investigation. As such, our applications and methodology relate to the literature on missing data (e.g., Little and Rubin 2019) and data fusion (e.g., Feit and Bradlow 2018; Gilula, McCulloch, and Rossi 2006; Kamakura and Wedel 1997, 2000). What distinguishes our contribution from the classical missing data problem is that our target list is missing an entire variable of interest—instead of missing individual observations—necessitating the use of external data. The distinction from the extant literature on data fusion is that we (1) lack prior or data-based knowledge about the joint distribution of observed and entirely unobserved variables in the target list (Kamakura and Wedel 1997, 2000) and (2) lack an observable conditioning argument in the target list to link to external data (Gilula, McCulloch, and Rossi 2006). Thus, we address the nonignorable missingness of an entire variable in the target list that prevails even after conditioning on what is observed in the target list. We accomplish this through a model of selection based on the missing variable. In a nutshell, we replace the assumption of conditional independence based on *observed* variables (e.g., Gilula, McCulloch, and Rossi 2006) with an assumption about conditional independence based on *unobserved* variables. As we show, this change in perspective corresponds to different models of generating the target list from the population.

Finally, a recent application leveraging information across different data sources by McCarthy and Oblander (2021) addresses improved information about a general population from including information available from more detailed but endogenously selected data. The goal of McCarthy and Oblander's research is to debias how information that is indeed available from the detailed data contributes to more efficient inference about a population. In contrast, our goal is to learn about information that is missing entirely from our target list by leveraging known information in the population and a model of endogenous selection.

Model Development

Yahoo! Smart Billboards

We illustrate the model developments by casting it into the context of Yahoo!'s Smart Billboards (Chen and Strimaitis 2016). We imagine that, as described in the firm's patent, the company scans the makes and models of cars passing by on a highway, links that information to a data broker's database to profile the car owners, and attempts to customize the advertising display a couple of miles down the road based on the drivers' inferred characteristics. Car owners' demographics vary significantly across brands and models. For example, Jeep fans are more likely to be white, conservative, and work in construction or the military (AutoInsurance Center 2015); 70% of Honda drivers have a college degree versus 35% for Chevy owners (Greenburg 2009). These variations can be used to infer the characteristics of a target population, such as income distribution. It is no surprise that data brokers allow their customers to "target consumers based on the make, model, or style of vehicle they drive" (Merkle Inc. 2017).

Next, we imagine that Yahoo! is interested in identifying the most likely income of the drivers on a highway (i.e., the "target list") by observing their vehicle brands and models.

Model Notations

In the context of this research, reference tables (e.g., Census data, Acxiom database) are categorical along both the categories being described (e.g., zip codes, segments) and the variables to be profiled (e.g., income, age) (Federal Trade Commission 2014; Neumann, Tucker, and Whitfield 2019). If there exist reference tables that describe variables of interest using continuous distributions (e.g., quantiles, mean, variance), they are the exception rather than the rule. We discuss this limitation in greater detail subsequently and explicitly incorporate

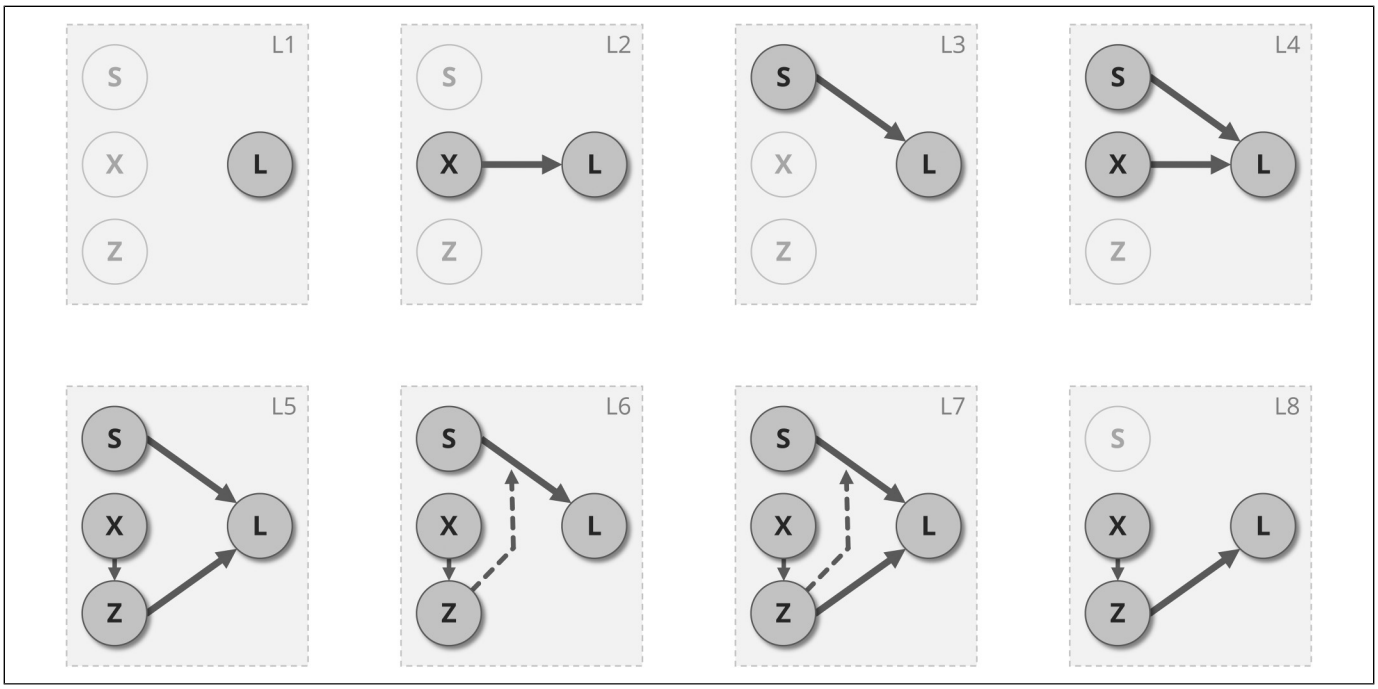

Figure 1. Different list-selection mechanisms, labeled L1 to L8 (the list is nonexhaustive), can be hypothesized by the analyst and lead to different likelihood functions and thus different estimated customer profiles.
Notes: L refers to the (possibly nonrandom) selection of individuals into the target list, S to the (unobserved) variable of interest (e.g., income, education), X to observable characteristics (e.g., zip code, car model), and Z to a continuous indicator stemming directly from X (e.g., distance to nearest store, car price). L4 cannot be estimated meaningfully.

that data-format constraint in our developments. We use these notations throughout:

$1 \dots i \dots I$	Individuals in the target list to be profiled.
$1 \dots j \dots J$	Data categories of interest to be inferred, which are known for the reference table (e.g., Merkle database) but missing and need to be inferred for the target list. In our example, J refers to income categories.
S_i	Unobserved category of individual i, such that $S_i = j$ means she belongs to income category j. We denote the vector of probabilities of belonging to unobserved categories S as p(S) and that of belonging to a specific category $p(S = j)$.
$1 \dots k \dots K$	Key used to link the target list to the reference table. In our illustration, the car brand and model provides the primary key.
X_i	Key of individual i such that $X_i = k$ means that she drives the kth car entry in the reference table.
$Y_{j,k}$	Number of individuals in the reference table who belong to the jth category and have the kth key. For simplification, a dot means ∀j or ∀k, respectively (e.g., $Y_{j,\cdot}$ is the number of individuals in the reference table who fall in the jth income bracket, $Y_{\cdot,\cdot}$ is the total number of individuals in the reference table).
N_k	Number of individuals in the target list with the kth key. $\sum_{k=1}^{K} N_k = I$.
$p(S_i = j\|X_i)$	Probability of individual i to belong to category j, conditional on their key being X_i.
$p(S_i = j\|X_i, L)$	Probability of individual i to belong to category j, conditional on their key being X_i and conditional on knowing that this customer is a member of the target list. These are the individual estimates of interest in most targeting applications.
$p(S = j\|L)$	Proportion of individuals in the target list estimated in category j. These are the focal quantities in profiling applications.
Z_i	An indicator of far lower dimensionality than X that may summarize its information contribution to the list-selection process.

If one tries to infer the unobserved characteristics (S: income) of a target list (L: drivers on Route 101) based on aggregate data available in a reference table through the observation of a readily available indicator (X: car model), there are multiple list-selection mechanisms that the analyst can hypothesize. Although not exhaustive, this article will cover eight of them (summarized in Figure 1) most likely to be relevant in a marketing context.

Before we discuss the eight selection mechanisms in detail, we note that (1) in some contexts, the "behavioral story" behind some selection mechanisms is unlikely and can be excluded a

priori based on managerial knowledge; (2) some—potentially valid—selection mechanisms cannot be estimated due to data limitations; and (3) if competing selection mechanisms are plausible and can be estimated with the aggregate data at hand, the analyst can evaluate them and identify the best model a posteriori using Bayes factors.

L1: Simple Random Sampling

In this scenario, every member of the reference table is hypothesized to be equally likely to be selected into the target list L. Profiling a specific target list becomes a pointless managerial exercise because each list is identical (in expectations) to the general population. Although this mechanism is of no managerial interest, we show that some of the more elaborate selection mechanisms discussed next include random selection as a special case.

L2: List Selection Based on X

The typical household income range of a Subaru Outback's owner is \$75,000–\$99,000. If Yahoo!'s Smart Billboard spots a Subaru Outback on Route 101, it would be natural to assume that the most likely income of that particular car driver is in the \$75,000–\$99,000 bracket as well (i.e., matches the distribution in the reference table). We label this approach the "simple count method." Although not immediately apparent, this approach suffers from a form of ecological fallacy, which we demonstrate next. Formally, the simple count method states that

$$p(S_i = j|X_i = k, L) = \frac{p(S = j, X = k)}{p(X = k)} = \frac{Y_{j,k}}{Y_{\cdot,k}} \quad (1)$$

In terms of customer profiling, the simple count method takes an average across all customers, (i.e., integrates over the observed distribution of X in the list L):

$$P(S = j|L) = \frac{1}{I}\sum_{i=1}^{I} p(S_i = j|X_i) = \frac{1}{I}\sum_{i=1}^{I} \frac{Y_{j,X_i}}{Y_{\cdot,X_i}} = \frac{1}{I}\sum_{k=1}^{K} N_k \frac{Y_{j,k}}{Y_{\cdot,k}}. \quad (2)$$

If many target consumers drive cars that tend to be owned by affluent customers, the method infers that the consumers in the target group are more likely to be wealthy. This intuitive and deceptively simple approach is used widely in academic research (see Cole, Dingle, and Bhayani 2005; Dias et al. 2019; Greene and Milne 2005; Van Dijk and Paap 2008[3]). Per our interviews, it is also the standard method employed in the industry, "trivial and very easy to implement" (Dr. Amelia Waddington, Director of Product, Data Science, LiveRamp). Although not considered state-of-the-art, it is believed to be "the only solution available lacking more granular data at the individual level" (F. Grellier, Chief Data Officer, Weborama).

One of the limitations of the simple count method, however, is that predictions are independent of context. The predicted income of a Subaru Outback driver will be identical whether the car is spotted near the Silicon Valley, on a Nebraska highway, or in Central Manhattan. We next explain the reason behind that surprising prediction invariance.

Formally, the analyst is interested either in the distribution of income in the list (i.e., estimate P(S|L) for profiling) or in the most likely income of a particular individual in that list (i.e., estimate $p(S_i = j|X_i = k, L)$ for targeting). The only pieces of information at the analyst's disposal are (1) a reference table of the general population describing income distribution for each car model, denoted P(X, S), and (2) the observed distribution of car models in the target list, denoted P(X|L). There are exactly K observations, one observed count for each X category (many are zeros). If the analyst assumes that list selection is only conditional on X (we indicate by the sign ≡ the key model assumption):

$$P(L|X, S) \equiv P(L|X) \quad (3)$$

Then, the simple count method (Equations 1 and 2) follows as the correct approach to profile the list and to target list members. Equation 3 clarifies the important structural assumption implicit to the simple count method that any dependence between L and S is perfectly mediated by X. In other words, it is assumed that conditional on X, the missing variable of interest S does not at all contribute to the selection into the list.

The simple count method has appealing properties. Its likelihood function is equal to

$$\begin{aligned} L &= \prod_{k=1}^{K} [p(X = k|L)]^{N_k} = \prod_{k=1}^{K} \left[\frac{p(L|X = k) \times p(X = k)}{p(L)}\right]^{N_k} \\ &= \prod_{k=1}^{K} \left[\frac{p(L|X = k) \times p(X = k)}{\sum_{m=1}^{K} p(L|X = m) \times p(X = m)}\right]^{N_k} \\ &= \prod_{k=1}^{K} \left[\frac{v_k \times p(X = k)}{\sum_{m=1}^{K} V_m \times p(X = m)}\right]^{N_k}, \end{aligned} \quad (4)$$

where v_k denotes the probability of selection into the list conditional on X = k. This parameter can be inferred from the data up to a multiplicative constant. However, there is no need to do so, because the elements p(X = k|L) on the left-hand side can be directly computed from the target list (i.e., without numerical estimation). Because the simple count method requires only p(X|L) and p(S|X) as inputs, where the latter is from the reference table, the analyst can proceed without ever thinking about the corresponding likelihood function. The method is, therefore, easy to use, which explains its wide use in practice.

[3] Van Dijk and Paap (2008) infer sociodemographics at the aggregate level using the simple count method as a prior for missing individual-level covariates in a regression model. Notwithstanding their development of efficient inference based on data augmentation in this situation, their approach will benefit from the Bayesian profiling developed herein if the sample to inform the regression equation is selected from the population based on the unobserved covariate.

The likelihood in Equation 4 clarifies that the simple count method implicitly uses K observations to calibrate a model of K parameters. As such, the simple count method is a saturated model for the observed distribution of X in the list, and no other model can fit the data better. In terms of likelihood, selection based on X perfectly rationalizes any differences between p(X|L) and p(X), the distribution of X in the list and that in the population.

The conditional independence assumption behind the simple count method, summarized in Equation 3, is valid only under two circumstances. First, if the customer list L is a random draw of the general population (i.e., $\forall k,\ v_k = c$), the simple count method reverts to case (L1), but the profiling exercise is moot and of no managerial interest. Second, and more generally, the simple count method is correct if X contains all the relevant information, such as when X is at the source of the list-selection mechanism itself. In other words, the simple count method assumes that, for example, the car make is directly related to why a particular driver is spotted on Route 101, rather than other unobserved variables correlated with the car make in the population. When this assumption is violated, the simple count method leads to biased estimates.

The simple count method assumes a list-selection mechanism that is unlikely to hold in many managerially relevant contexts. To the best of our knowledge, however, discussions about the underlying list-selection mechanism never occur in practice. Instead, researchers and data brokers routinely employ the simple count method, which implicitly assumes list-selection mechanism L2 (based on X), without realizing it. We show that explicitly hypothesizing different list-selection mechanisms, while computationally more complex, may lead to much-improved predictions.

L3: List Selection Based on S

An alternative approach assumes that the conditioning of list selection on X can be ignored when S is considered—that is, $P(L|X, S) \equiv P(L|S)$. Starting from the decomposition of the joint distribution $P(X, S|L)$ into $P(S|L) \times P(X|S)$ implied by this list-selection mechanism, we obtain the following likelihood function for an observed X in the target list:

$$p(X = k|L) = \sum_{j=1}^{J} \frac{p(L|S = j) \times p(S = j)}{p(L)} p(X = k|S = j). \quad (5)$$

Equation 5 defines a likelihood for estimating the unobserved list-selection mechanism P(L|S) on the right-hand side based on the observed distribution of X in the list on the left-hand side. Contrary to the simple count method, p(L|S) is not a direct function of the data and needs to be estimated numerically. Defining the *unobserved* conditional list inclusion probability $p(L|S = j)$ as weight parameter w_j, and making explicit that P(L) in Equation 5 equals $\sum_{j=1}^{J} p(L|S = j) \times p(S = j)$, we obtain the following likelihood function for observing a particular set of X values in a list L:

$$L = \prod_{k=1}^{K} \left[\sum_{j=1}^{J} \frac{w_j \times p(S = j)}{\sum_{n=1}^{J} w_n \times p(S = n)} \times p(X = k|S = j) \right]^{N_k} \quad (6)$$

The only parameters to be estimated in Equation 6 are the weights $w_1 \ldots w_J$, which are identified up to a multiplicative constant. We defer a formal analysis of identification to Web Appendix A but note that, in general, statistical information about the weights will improve as (1) the X variable has more categories, (2) the differences between the conditional distribution P(X|S) in the reference table and the corresponding marginal distribution P(X) increase, and (3) the size of the target list increases.

We note that, in the case of equal weights, $\forall j,\ w_j = c$, Equation 6 reverts to an unconditional list-selection mechanism where $p(X = k|L) = p(X = k)$. Thus, both mechanisms L2 and L3 contain mechanism L1 as a special case, but neither is L2 a special case of L3 or vice versa.

After estimating $w_1 \ldots w_J$, the prediction at the individual level becomes

$$p(S_i = j|X_i = k, L) = \frac{w_j \times p(S = j) \times p(X = k|S = j)}{\sum_{n=1}^{J} w_n \times p(S = n) \times p(X = k|S = n)}$$

$$= \frac{w_j \times p(X = k, S = j)}{\sum_{n=1}^{J} w_n \times p(X = k, S = n)} \quad (7)$$

Profiling follows as a margin over all X, with $p(S = j|L) \propto w_j \times p(S = j)$.

L4: List Selection Based on X and S

It would be tempting to calibrate a model that assumes that list selection is conditional on all available data, namely X and S. This assumption is theoretically superior to any other more restrictive assumption and should always be "true." Unfortunately, it cannot be implemented in practice due to a fundamental identification problem, which is the crux of this research.

Assuming that the contributions of X and S on list selection are additively separable and that parameter η captures the relative influence of X on list selection, whereas $(1 - \eta)$ captures the relative influence of S, the likelihood function becomes

$$L = \prod_{k=1}^{K} \left\{ \eta \left[\frac{v_k \times p(X = k)}{\sum_{m=1}^{K} v_m \times p(X = m)} \right] + (1 - \eta) \left[\sum_{j=1}^{J} \frac{w_j \times p(S = j)}{\sum_{n=1}^{J} w_n \times p(S = n)} \times p(X = k|S = j) \right] \right\}^{N_k} \quad (8)$$

The model has $K + J + 1$ parameters to be estimated numerically based on K observations. While a model with more parameters than observations can be estimated within a Bayesian

framework, the issue is that a list-selection based on X *perfectly rationalizes any differences between* P(X|L) *and* P(X) *already*. Once X is included as a conditioning argument in the list-selection mechanism, there is no information to learn about the potential role of S in the selection process.

Going back to mechanism L2, the model P(L|X) has as many parameters as there are categories in X: it fits the dependent variable perfectly, but the model has zero degrees of freedom. If a second regressor S is introduced on top of X, as in mechanism L4, the fit would not be improved, but the solution would become unstable. For any vector $w_1 \ldots w_J$ chosen randomly, there exists a vector $v_1 \ldots v_K$ that perfectly rationalizes the likelihood. The specific values of the parameters would be meaningless and would only reflect the influence of subjective priors.

List-selection mechanism L4 is overparametrized. It has K + J + 1 parameters, whereas mechanism L2 has only K. P(L|X, S) will offer an identical fit to P(L|X) while requiring J + 1 additional parameters. Therefore, mechanism L4 will systematically be rejected in favor of the more parsimonious model L2. We illustrate this issue in our last empirical application.

Lacking more disaggregated data, the analyst is forced to assume either mechanism L2 (based on X) or mechanism L3 (based on S). When common sense cannot point to one in particular, the analyst can estimate both and compare their Bayes factors, as we discuss subsequently.

This limitation is a challenging problem for the analyst. Some contexts command a list-selection mechanism based on both X and S. For instance, a hard-discount chain that wants to profile its customers' income (S) based on their zip codes (X) would be hard-pressed to choose between list-selection mechanisms L2 and L3. While it is undeniable that its patrons are selected based on income (low-income customers are more likely to visit discount stores), it is equally undeniable that the geographical proximity of its stores plays a similarly important role in the list-selection mechanism (individuals who live near a store are more likely to shop there).

Note that the inherent difficulty of including X as a conditioning argument resides in its dimensionality. S categories are often in the dozens, whereas X categories are usually in the thousands. If a firm wants to profile customers' occupations from their addresses, the data published by the Census Bureau include 13 occupation categories (S) for 41,692 zip codes (X). If one wants to infer customers' income from the cars they drive, there are 16 income categories for 9,840 car models. In a subsequent application where we estimate age from first names, our reference table includes 21 age categories for 12,834 first names.

In practical applications, reference tables are always categorical along both X and S. This real-life constraint creates a high-dimensional optimization problem that prevents the estimation of the "true" model P(L|X, S). The question becomes, Can this dimensionality be reduced?

Reducing the dimensionality of S. In practical applications, variables S may include income, education attainment, age, ethnicity, number of children, marital status, or occupation. While some S categories are nominal (ethnicity, occupation, marital status), others are ordinal (education, number of children). They may even represent an arbitrary slicing of an underlying continuous construct (age, income).

In the two latter cases, selection probabilities of adjacent S categories might be related, and a weight w_j can be considered informative of the most likely values of weights w_{j-1} and w_{j+1}. We could obtain such property by constraining the weights $w_1 \ldots w_J$ to follow, for example, a beta density function. This constraint reduces the dimensionality of the estimation problem along S from J parameters to the two parameters a and b in the beta distribution.

Reducing the dimensionality of X. In typical reference tables, S categories are in the dozens, whereas X categories are often in the thousands (e.g., Steenburgh, Ainslie, and Engebretson 2003). Reducing the dimensionality of S is useful, and often desirable, but the dimensionality of X is the real culprit that prevents estimating the "true" model P(L|X, S).

In marketing applications, typical X categories include (see Table 1):

- Zip codes (profiling from Census data),
- City blocks (Conexance behavioral typologies),
- IP addresses (AcquireWeb demographic profiling from geolocation),
- Websites visited (Google Display Network),
- Car models (Yahoo! Smart Billboard; see Scenario 1 in the introduction),
- Segment membership (programmatic segmentation deployed by Experian, Acxiom, or Merkle; see also Geraghty et al. [2017] and Scenario 2), and
- First names (age pyramid estimated from first names, e.g., CACI 2002; Scenario 3).

While this observation may not hold in other fields, the X categories in the reference tables typically used in marketing are *genuinely* nominal. IP addresses or car models are not an arbitrary slicing of an underlying continuous construct. And unlike S categories such as income or age, X categories such as first names or city blocks cannot be classified from high to low.

However, we suggest that the analyst may summarize X's information contribution by an indicator Z of far lower dimensionality. For instance, in the hard-discount store example, the analyst might engineer an indicator Z that captures the distance of a zip code to the nearest store. They could also experiment with the log-distance, or with a weighted average of the distances to the three nearest stores, or with the total number of stores within a 20-mile radius. Such an indicator Z does not stem from the arbitrary categorization of a continuous dimension into X-categories (unlike some S variables, such as age or income). Therefore, its construction would require domain knowledge, a solid dose of feature engineering, and most likely, extensive testing.

While different approaches can be envisioned, in the interest of space, this research focuses on the cases where selection

based on X (e.g., zip codes) can be modeled as selection based on a continuous variable Z (e.g., distance). We use the notation $Z(X = k)$ to refer to the deterministic transformation of X into continuous Z.

L5: List Selection Based on S and Z (Additive)

If we assume $P(L|X, S) \equiv P(L|Z, S)$ and separability of the respective influence of Z and S on list selection, the likelihood function P(X|L) becomes

$$L = \prod_{k=1}^{K} \left\{ \eta \left[\frac{p[L|Z(X=k)] \times p(X=k)}{\sum_{m=1}^{K} p[L|Z(X=m)] \times p(X=m)} \right] + (1-\eta) \left[\sum_{j=1}^{J} \frac{w_j \times p(S=j)}{\sum_{n=1}^{J} w_n \times p(S=n)} \times p(X=k|S=j) \right] \right\}^{N_k}. \quad (9)$$

Because we only consider the case where $Z(X = k)$ is continuous, $p[L|Z(X = k)]$ can be advantageously replaced by a continuous function $f[Z(X = k), \beta]$ such as a sigmoid transformation, where only the parameter vector β needs to be estimated:

$$f[Z(X=k), \beta] = \{1 + e^{-[\beta_0 + \beta_1 \times Z(X=k)]}\}^{-1}. \quad (10)$$

The quantity of interest for profiling $p(S = j|L)$ becomes

$$p(S=j|L) = \eta \sum_{k=1}^{K|L} p(S=j|X=k) \frac{f(Z(X=k), \beta) \times p(X=k)}{\sum_{m=1}^{K|L} f[Z(X=m), \beta] \times p(X=m)} + (1-\eta) \frac{w_j \times p(S=j)}{\sum_{n=1}^{J} w_n \times p(S=n)}. \quad (11)$$

Here, the notation K|L indicates that, for profiling, the contribution of "direct" selection into the list based on $f[Z(X = k), \beta]$ only matters for profiling through the set $\{Z(X)|L\}$ (i.e., through the set of X values that ended up in the list). This is a consequence of conditional independence between S and L in this part of the mixture model in Equation 9.

L6: List Selection Based on S and Z (Moderator)

An interesting case arises when, instead of assuming that the influence of S and Z are additively separable in P(L|Z, S), we investigate the case where Z *moderates* the influence of S on list selection. In the previous example of the hard-discount stores, for instance, the analyst may want to investigate the "behavioral story" where individuals become patrons of the chain *because* of their income (S), while geographical proximity to stores (Z) moderates list selection. For example, selection based on income may be less influential for consumers who live nearby.

Mechanism L6 is similar to mechanism L3, with the exception that the selection weights $w_j = p(L|S = j)$ are now moderated by J functions $f[Z(X = k), \beta_j]$, such as the one defined in Equation 10. Equation 6, therefore, becomes

$$L = \prod_{k=1}^{K} \left\{ \sum_{j=1}^{J} \frac{w_j \times f[Z(X=k), \beta_j] \times p(S=j)}{\sum_{n=1}^{J} w_n \times f[Z(X=k), \beta_n] \times p(S=n)} \times p(X=k|S=j) \right\}^{N_k}. \quad (12)$$

Intuitively, the weights $w_1 \ldots w_J$ represent the list-selection mechanism based on income for individuals in close proximity to hard-discount stores, while the moderating function f accounts for changes in the income-based selection pattern as the distance to the store increases (e.g., income-based selection may be more pronounced as distance increases).

Once $w_1 \ldots w_J$ and $\beta_1 \ldots \beta_J$ are estimated based on the likelihood in Equation 12, the quantity of interest is obtained by replacing w_j by $w_j \times f[Z(X = k), \beta_j]$ in Equation 7.

L7: List Selection Based on S and Z (Additive + Moderator)

This list-selection mechanism combines models L5 and L6 and assumes that the indicator Z has both a direct role in the list selection and a moderating role through S. This model requires $(3J) + 2$ parameters and could account for the effect that prospective customers living at greater distances from the target stores are generally less likely to frequent them.

L8: List Selection Based on Z

A list-selection mechanism based exclusively on Z is also possible. Its likelihood is obtained by replacing $v_k = f[Z(X = k), \beta]$ in Equation 4. Additional model specifications are also possible but would require highly detailed reference tables rarely available in typical marketing applications.[4]

[4] For instance, one could envision several advanced conditioning arguments, such as models conditioning on multiple S's (i.e., P[L|S1,S2,S3]). Interestingly, some models have well-defined likelihood functions, but the estimated weights cannot be translated into meaningful consumer profiles due to the data format and data scarcity in typical reference tables. For instance, while $p(X, S_1)$, $p(X, S_2)$, and $p(X, S_3)$ are known, stem directly from the available reference tables, and can be combined in a likelihood function to explain list selection (assuming an additively separable mixture), such a model does not map back to the joint profile $p(S_1, S_2, S_3|L)$. Identification of the joint profile $p(S_1, S_2, S_3|L)$ would require access to a separate hypercube for every X category in the population, cross-referencing, say, all combinations of income, age, and occupation category per car model. Such complex reference tables would require very detailed information about a substantial sample of the population. To the best of our knowledge, data brokers such as LiveRamp or Merkle do not have enough granular data to create and maintain said hypercubes. Likewise, even though the U.S. Census Bureau has access to more complete and detailed cross-tabulated data, it is forbidden to publicize fine-grained information (for privacy and legal reasons, because cross-tabulated data at such a high level of detail may allow for partial inference of individual information). Note that in the rare situation where multidimensional, cross-tabulated data are available, the analyst can create as many S categories as there are $S1 \times S2 \times S3$ combinations and revert to the mechanism L3. Similarly, multiple X variables become fruitful when their joint distribution with S variables is available for the population.

Contribution Summary

This research makes two distinct contributions. First, we demonstrate that the simple count method relies on the *implicit* assumption that list selection is entirely mediated by X. We subsequently demonstrate that, when this assumption is violated, the simple count method leads to biased inference. Given the ubiquitous use of the simple count method in the industry and academic research (e.g., Cole, Dingle, and Bhayani 2005; Dias et al. 2019; Greene and Milne 2005; Van Dijk and Paap 2008), this finding is not trivial.

Second, we show that when the analyst considers list-selection mechanisms *explicitly*, several models—each corresponding to a different "behavioral story"—are possible. While our list of models is not exhaustive, we develop novel models likely to be useful in marketing applications.

Model Selection and Bayes Factors

The remainder of this article is divided into two parts. First, we cast our analyses in contexts where we can rule out a priori list selection based on X, and careful deliberation failed to produce additional Z variables to consider. These situations are practically relevant. For instance, we can likely ignore models built on the assumption that the Honda Civic an individual drives renders other variables (e.g., income, education, the driver's appearance on Route 101) conditionally independent.

When no Z indicator is available, only two mechanisms can be tested: L2's simple count method, which implicitly assumes $P(L|X, S) \equiv P(L|X)$, and L3's simple Bayesian profiling method, which explicitly assumes $P(L|X, S) \equiv P(L|S)$. We compare both, first in a simulation, then in a real-life application, and show when the first mechanism results in misleading inference.

The last part of this research is dedicated to cases where an observed variable X cannot be ruled out a priori and the analyst can identify a suitable low-dimensional variable Z to capture its influence. Our second empirical application covers that more complex situation. We compare all eight mechanisms on several S variables and multiple possible operationalizations of the Z indicator and report the results.

In the presence of multiple plausible and estimable mechanisms, the analyst can compare marginal likelihoods of the observed data in the list under various list-selection hypotheses; Bayes factors will identify the most appropriate model to use. In the interest of space, we defer that discussion to the Web Appendix B, which presents the underlying theoretical considerations and provides a numerical example. We report the Bayes factors for all subsequent simulations and empirical illustrations discussed in this article.

Bayesian Estimation

We use Markov chain Monte Carlo for Bayesian inference that couples the likelihoods with subjective priors for the weights $w_1 \ldots w_J$ and $v_1 \ldots v_K$ and the parameters β_0, $\beta_1 \ldots \beta_J$, and η.

When directly estimating weights for each of the J (respectively K) categories of the unobserved variable S (respectively X), we use independent weakly informative log-normal distributions as subjective priors. The log-normal distribution ensures that estimated weights are positive. Although we do not resort to it in this research, the analyst could also impose more informative priors for some categories based on managerial knowledge. For instance, a discount store could impose higher subjective priors on low-income categories, thus capturing the belief that low-income customers are a priori more likely to self-select into the target list. Likewise, a bank could exclude certain age categories based on legal considerations (e.g., if customers must be adults to open an account, informative priors expressing minimal or even no prior support for age categories of minors can further improve posterior inference).

When adjacent weights $w_1 \ldots w_J$ are expected to be related, we constrain the weights to follow a beta density function (e.g., see application #1), with diffuse priors on beta parameters a and b.

Simulations

Simulation Setting

To illustrate the conditions under which the Bayesian method will outperform the simple count method, we run the following simulation: Suppose Yahoo! wants to estimate the income distribution of the drivers on U.S. Route 101 by augmenting the car brands scanned by its Smart Billboard on the highway with information from the Merkle database. We assume that Merkle's database lists 500,000 car owners in California (reference table) along with their income (S), spread across ten car models (X). We also assume that Yahoo! observes 25,000 cars within a given time frame (customer list L).

For the sake of this simulation, we generate five distinct income categories from an underlying normal distribution such that, after centering, the lowest category groups customers with an income that falls between $-\infty$ and -2.5σ, the second-lowest income category groups customers with income between -2.5σ and $-.83\sigma$, the middle (and most populated) income category groups those with income between $-.83\sigma$ and $.83\sigma$, and so on. We design the data-generating mechanisms such that car brands have no direct influence on the likelihood of being spotted on that highway but are correlated with a (potentially) causal variable, income.

We assume that the analyst does not have an indicator Z that would be both managerially relevant and readily available. Therefore, they can only test and compare mechanisms L2 and L3 (our second empirical application relaxes this assumption).

We manipulate two factors. First, we vary the probability that a Californian driver will be spotted driving on U.S. Route 101 in the observational time frame, conditional on income (i.e., we vary the extent to which income influences the list inclusion mechanism, manipulation #1). For the sake of the simulations, we assume that a proportion p of the drivers in California will drive on U.S. Route 101 in the observational time frame. We write $p(L|S) = [p, p, p, p, p]$. We vary this vector of probabilities such that income becomes increasingly predictive of list inclusion, with multiplicative factors of [.6, .8, 1, 1.2, 1.4], [.2, .6, 1, 1.4, 1.8], [0, .4, 1, 1.6, 2], [0, .2,

Table 2. RMSE of Income Distribution, True List Versus Simple Count Method Estimates.

		Dependence Between List Inclusion and Income				
		← Low				High →
Dependence between income and car brand	↑ Low	.024	.048	.072	.096	.121
		.022	.043	.066	.087	.110
		.018	.036	.054	.072	.091
	↓ High	.012	.025	.038	.051	.065
		.005	.010	.015	.020	.025

Notes: The less representative of the general population the list is, the more biased the results.

1, 1.8, 2], and [0, 0, 1, 2, 2].[5] In the final condition, drivers in the two lowest income brackets will never drive through U.S. Route 101, whereas drivers in the two highest income brackets will be twice as likely to commute through that route than drivers in the middle-income category. Drivers falling in the middle, most widely populated income category have a fixed probability p (.5 in our setting) of being spotted on the highway, regardless of the manipulation level.

Second, we vary the extent to which the unobserved characteristic of interest (income) is correlated with the observed characteristic (car brand) in the reference table. We set correlations[6] to .2, .4, .6, .8 and .99 (manipulation #2). With a correlation of .2, income distributions weakly correlate with car brands (Pearson's $R^2 = .04$), and the latter conveys very little information about the former. At .99, a car brand is a solid indicator of its driver's income. We exclude the case of zero correlation *in the reference table* from our simulations. If income is orthogonal to car brand in the reference table, a selection mechanism based on income cannot induce dependence between income and car brand. No firm would attempt to infer the former from the latter in this situation.

We simulate 5 × 5 experimental conditions, replicated 100 times, for a total of 2,500 simulations. For each simulation, we generate a reference table of 500,000 drivers in California with a specified joint distribution of income and car brand (manipulation #2) and then determine if they may be spotted on U.S. Route 101 according to the conditional probabilities (manipulation #1). We then randomly draw a sample of 25,000 drivers on that highway, whose car brands we use to infer their income distribution, using either the simple count or the Bayesian profiling method conditioning on S (for completeness, computational considerations such as convergence, speed, and mixing properties are reported in Web Appendix C).

Profiling Results

For each simulation, the simple count method provides an estimated distribution of income groups among the 25,000 drivers in the sample. we report in Table 2 the root mean squared error (RMSE) between this distribution and the true distribution, averaged over 100 replications for each cell. RMSEs are, therefore, on the scale of probabilities of belonging to a particular income group. When list inclusion plays little role in the list-selection mechanism (leftmost column), the list is a near-random draw from the population, and the simple count method handles that case relatively well. As list inclusion more strongly relates to income, however, errors become substantial.

Table 3 reports the RMSE obtained from the Bayesian method conditioning on S. As we increase dependence between income and car brand, errors dramatically decrease (lower rows).

Table 4 reports the relative difference between Tables 2 and 3 (simple count method RMSE minus Bayesian method RMSE divided by the Bayesian RMSE); a positive value indicates the Bayesian method leads to fewer errors and dominates the simple count method.

It is apparent from Table 4 that the Bayesian profiling method leads to better estimates in all but one case, when two conditions are simultaneously met: (1) when the observed characteristic (car brand) and the unobserved characteristic of interest (income) are barely correlated in the reference table and (2) when the list inclusion mechanism (being spotted driving one's car on U.S. Route 101) is almost independent of the unobserved characteristic (income), and the list L is a near-random draw of the reference table. As soon as either of these two conditions is violated—as would be the case in most managerially relevant situations—the Bayesian profiling method outperforms the simple count method. Furthermore, we show how Bayes factors correctly identify the data-generating mechanism (selection based on S vs. X or random selection) across the simulation conditions in Web Appendix B.

[5] Note that both the Bayesian profiling method and the simple count method would cover the special case where income does not influence the list-selection mechanism at all (i.e., [1, 1, 1, 1, 1]). This special case would be of no managerial interest, though, because predictions would coincide with income distribution observed at the population level.

[6] Car brands are generated by discretizing a normally distributed variable using -2.5σ, -1.875σ, -1.25σ, $-.625\sigma$, 0, $.625\sigma$, and so on as truncation points. The correlations in manipulation #2 refer to correlations between the normal variates used to generate discrete income groups and car brands. In estimation, we do not exploit the implied ordering of categories. We simply use the underlying bivariate normal distribution to generate a bivariate categorical distribution where larger correlations between underlying continuous normal variates translate into more dependence between categorical variables as measured by, for example, Cramér's V. Thus, we use the following weakly informative subjective prior for the weights in the likelihood defined in Equation 6: $\log \mathbf{w} \sim N(0, \mathbf{I}\ .1)$, where I is a diagonal matrix of dimensionality equal to the number of S-categories—five income categories in our simulation. We update all weights in a single Metropolis–Hastings step with a random walk proposal for $\log \mathbf{w}$.

Table 3. RMSE of Income Distribution, True List Versus Bayesian Method Estimates.

		Dependence Between List Inclusion and Income				
		← Low				High →
Dependence between income and car brand	↑ Low	.099	.076	.054	.027	.033
		.026	.029	.026	.019	.018
		.011	.011	.011	.010	.010
	↓ High	.004	.004	.003	.004	.003
		.001	.001	.001	.001	.001

Notes: The higher the dependence between the observed (car brand) and unobserved (income) characteristics, the smaller the errors.

Table 4. Relative RMSEs, Simple Count Method Versus Bayesian Method Conditioning on S.

		Dependence Between List Inclusion and Income				
		← Low				High →
Dependence between income and car brand	↑ Low	−66%	−10%*	79%	429%	443%
		71%	172%	324%	531%	628%
		132%	409%	627%	901%	1,359%
	↓ High	311%	839%	1,503%	2,046%	3,594%
		414%	1,032%	1,438%	3,016%	5,370%

*Not statistically different from 0 at $p = .05$.
Notes: A positive value indicates that the Bayesian method leads to more precise estimates, whereas a negative value indicates the simple count method outperforms the Bayesian method.

Targeting Results

In practice, consumer profiling is used to infer not only the general profile of a list (e.g., income distribution) but also the characteristics of specific individuals in that list for targeting purposes. While targeting specific individuals on a highway would be of little managerial interest in the context of Yahoo! Smart Billboards—the managerial goal is "grouplization," not individual-level customization of the ads displayed on the billboard (Chen and Strimaitis 2016)—we can still illustrate the relative merits of the Bayesian profiling method in targeting individuals numerically using our simulation. We find that the Bayesian profiling method improves targeting (hit rates) by up to 9.8% and achieves hits with higher certainty at the individual level. This would be useful if a company only wanted to target modal predictions that achieve a minimum level of certainty (for detailed results, see Web Appendix D).

Next, we illustrate Bayesian profiling with two empirical examples. The first illustrates the situation where X is unlikely to contribute to list selection based on prior reasoning and additional covariates to direct selection (Z[X]) are unavailable. The second, more general application illustrates the possibility of joint selection based on unobserved S and observed Z.

Empirical Illustration #1: Age Estimation

Estimating Age from First Names

Numerous readily available pieces of information can be matched (i.e., serve as keys) to external sources of information and offer insights about customers (e.g., physical addresses, IP addresses, car brands, visited websites). For example, first names can help predict age, which is an important element for consumer profiling and appears to relate to concepts such as brand loyalty, repurchase behaviors, customer profitability (Lambert-Pandraud, Laurent, and Lapersonne 2005), and charitable giving (Clotfelter 1980). In practice, age often serves as a predictor for direct marketing. Crié and Micheaux (2006) report that age remains one of the best discriminant variables in targeting and scoring models in the insurance industry. Richard Webber, one of the developers of the Mosaic segmentation at Experian (2019), has shown that first names can segment customers into cultural, ethnic, and religious groups (Webber 2007).

People's first names experience trends and fads over time. In the United Kingdom, for example, Ernest was a very popular first name in the 1930s; Kelly was very popular in the 1980s; and, as an extreme example, Adolph as a given name almost disappeared after World War II.

According to one marketing research firm (CACI 2002), "Age is one of the most important factors for understanding more about your customers—and communicating with them more successfully as a result. [First name] classification can help you to identify the age of the people on your database by looking at the likely age profile of their first names. It can also help you to target new data sources which match your distinct customer age profile."

Data Set

The data comprise the first names and dates of birth of 14,075 people. The target list comes from a French database of highly educated, wealthy, and older customers of a private bank. The

sponsor remains anonymous for confidentiality reasons. The reference table reports the age pyramids of all first names in France at the time of the analysis. The age is reported in 21 bins. Using the notation we introduced previously, $I = 14{,}075$ (size of the target list), $J = 21$ (number of age categories), and $K = 12{,}834$ (first names in the reference table).

We use names to estimate the age distribution of the target list with both the simple count and Bayesian consumer profiling models. Obviously, this particular exercise is of little interest to the sponsor organization, which is required by law to collect the age of its customers. But the actual dates of birth will serve as the benchmark to measure the models' accuracy in the more common situation when we do not observe the variable of interest in the target list. For legal reasons, none of the customers on the list were younger than 18 years. We use this information as a face validity test in our subsequent comparisons.

The sponsor organization reported that most of its customers come from affluent and highly educated families. These families tend to adopt naming patterns that differ from those of the general population (e.g., Lieberson and Bell 1992). In particular, they tend to be more innovative in the first names they choose for their children. After a first name becomes popular among the upper classes, it often is adopted by less affluent families, enters the mainstream, and declines in popularity among the wealthiest families. This cycle generally takes approximately four to five years (Levitt and Dubner 2005).

For the sake of demonstrating the impact of using a more appropriate reference table (i.e., a table conditioned on affluence and high education), and following Levitt and Dubner's finding (2005), we shift naming patterns by one age category (five years) in the reference table and calibrate all models using this shifted table. In other words, if the first name Adrien became popular again among the general French population in 1980, we assume that it became popular among upper-class families as early as 1975.[7]

Modeling Considerations

Finding a meaningful indicator Z that maps 12,834 first names into a lower dimension is a theoretical challenge. Similar to the previous simulation, this application falls in the category of problems where the analyst is limited between mechanisms L1 (random selection), L2 (simple count method conditional on X), and L3 (Bayesian profiling conditional on S).

[7] In a previous version of this article, we reported the results of both the simple count method and Bayesian profiling with and without the shift of one age category in the reference table. Results are directionally consistent, though both methods suffer when relying on a less appropriate reference table. This result nicely illustrates the usefulness of an appropriately conditioned reference table. For instance, if customers in the list come from a specific geographic area (e.g., New England), a reference table focusing on that specific area (rather than an enlarged geographic region; e.g., the United States as a whole) would be more appropriate. More generally, as we move away from Census data, if the reference table comes from a third-party firm (e.g., Experian), the aggregate data collected and provided by the firm might not be fully representative of the true population from which the list emanates, in which case predictions will suffer. However, this problem will affect all methods.

For the Bayesian profiling method, because the underlying construct S (age) is continuous, and we expect selection probabilities of adjacent age categories to be related, we constrain the weights $w_1 \ldots w_J$ using a beta density function with diffuse subjective prior distributions on its parameters a and b. In addition to being more parsimonious than estimating 21 independent weights, this prior effectively borrows information from all other age categories when determining the weight for a particular age category.

Profiling Results

We use names to estimate the age distribution of the target list, testing the three distinct list-selection mechanisms L1, L2, and L3. The simple count method is built on the assumption that age (S) can be ignored as a selection criterion. Because this common method implicitly assumes that list selection is conditional on first names, which is unlikely (an assumption that is confirmed by model comparisons; see reported Bayes factors in Web Appendix B), the method fails to capture the selection mechanism. The method infers an age distribution that is different from but heavily biased toward the reference table, as evidenced in Figure 2.

In Table 5, we report the correlation and goodness-of-fit measures. In terms of Pearson's R^2, the simple count estimates correlate at .790 with the true age distribution of the target list. This correlation improves to .978 in the Bayesian model. Pearson's χ^2 must be rejected in both cases (as is often the case with large samples), but it decreases from 4,234 for the simple count model to 679 for the Bayesian one. The RMSE also improves and is reduced by 70.4%, from 2.81% to .83%. Importantly, from a managerial point of view, the proportion of critical errors (i.e., the proportion of the target list estimated to be 18 years or younger) dramatically decreases from 6.30% to .01%. From a performance and goodness-of-fit perspective, the Bayesian profiling approach outperforms the simple count model on all counts.

Targeting Results

In targeting applications, a firm will be interested in estimating the age of a specific customer or prospective customer, for instance, concerning charitable giving (see Scenario 3 in the introduction), or for direct marketing in the health care, automotive, or educational service industries. In these contexts, estimating the age distribution of a target list as a whole is of less immediate interest, but estimating the most likely age of each individual in a list (as well as the degree of certainty of these estimates) is crucial.

With the simple count method, it is straightforward to estimate the most likely age (or age pyramid) of a specific individual; it is the age pyramid of the individuals in the reference table who share an identical first name. With the Bayesian method, we multiply each conditional probability from the reference table by the estimated weights to obtain the estimated age distribution (see Equation 7).

Going back to our target list in the financial industry, we take the example of a customer named Adrien (see Figure 3). This first name is particularly interesting because it has a bimodal distribution in the population. Adrien was a particularly common first

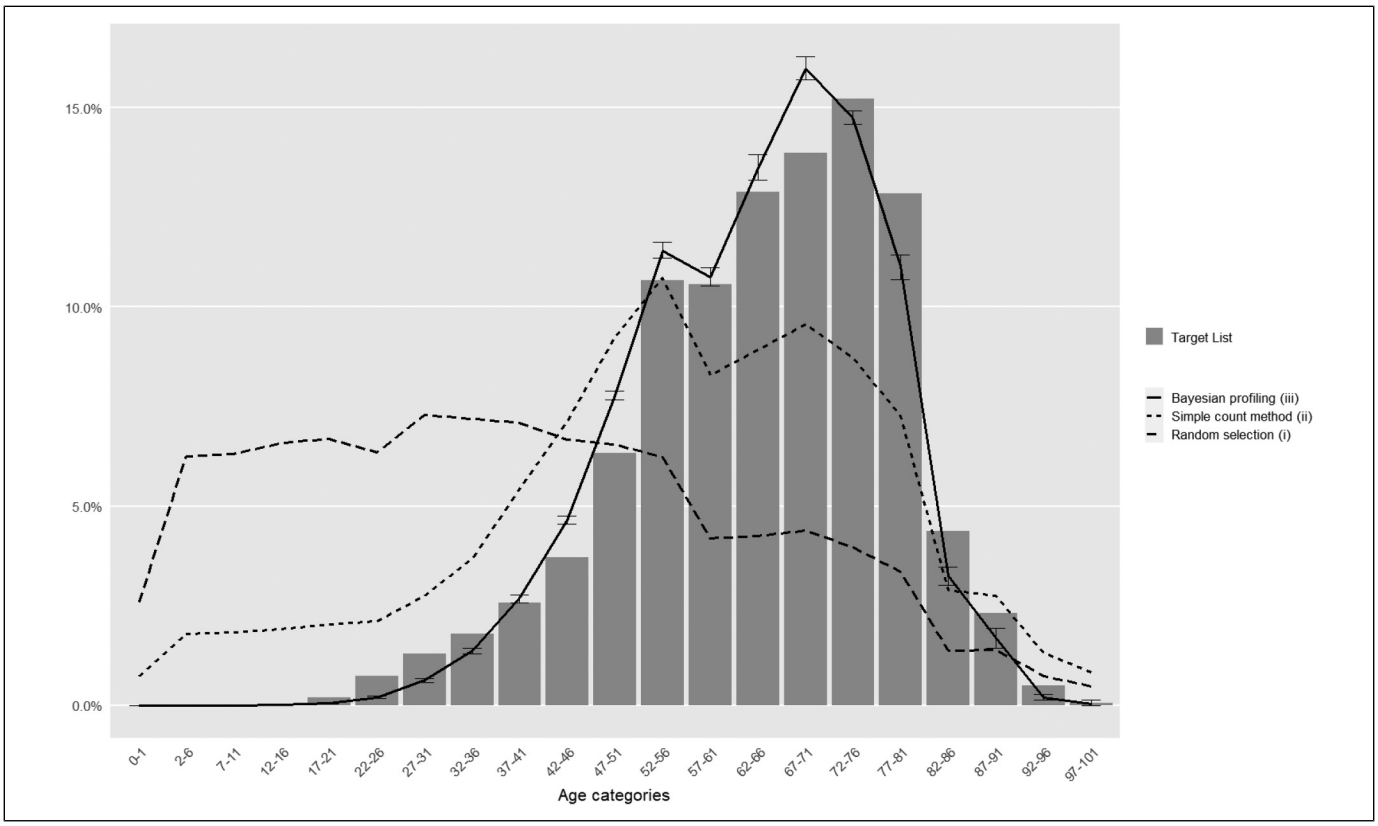

Figure 2. True age pyramids of the target list versus age pyramids of the same estimated assuming L1 random selection, L2 list selection conditional on X (simple count method), and L3 list selection conditional on S (Bayesian profiling).
Notes: The error bars display the 95% credible interval of the posterior.

Table 5. Correlations and Goodness-of-Fit Measures for the Two Profiling Models (Simple Count Method and Bayesian Profiling) After Age-Shifting the Reference Table to Account for the Name Leadership of Wealthy Families.

List-Selection Mechanism	Pearson's R	χ^2	RMSE	Log-Predictive-Likelihood	Critical Errors
Random sampling (L1)	−.092	22,971	5.97%	−44,505	21.80%
Simple count method (L2)	.889	4,234	2.81%	−36,262	6.30%
Bayesian profiling (L3)	.989	679	.83%	−33,901	.01%

name generations ago and became popular again recently but was quite unpopular for several decades in between.

The simple count method estimates that an Adrien in the target list is 18.8 years old on average and has an 82.8% probability of being 21 years old or younger (an improbable possibility given the context). The Bayesian profiling method estimates that he is 65.1 years old and has only a 2.9% probability of being 21 years old or younger. After applying the estimated corrective weights to the conditional age distribution in the population, the bimodal distribution becomes essentially unimodal, and predictions dramatically improve.

Because this customer belongs to a target list that has been estimated to be quite elderly, the estimated weights correct for the least likely parts of the age pyramid. Bayesian profiling correctly identifies that, among all the men named Adrien in the reference table, those who belong to the target list are likely to belong to the right-hand side of the age pyramid.

To further test the ability of the Bayesian profiling method to recoup age estimates at the individual level, we classified the 14,075 individuals into 1 of 21 age categories based on their first names, using both the Bayesian profiling and the simple count method.

The hit rate (prediction of the precise age category to which individuals belong based on the maximum predicted probability[8])

[8] Because Bayesian profiling delivers predicted probabilities that take all posterior uncertainty in estimated selection weights into account, a decision maker could use this information to target only individuals subject to some minimum level of certainty about their predicted age classification.

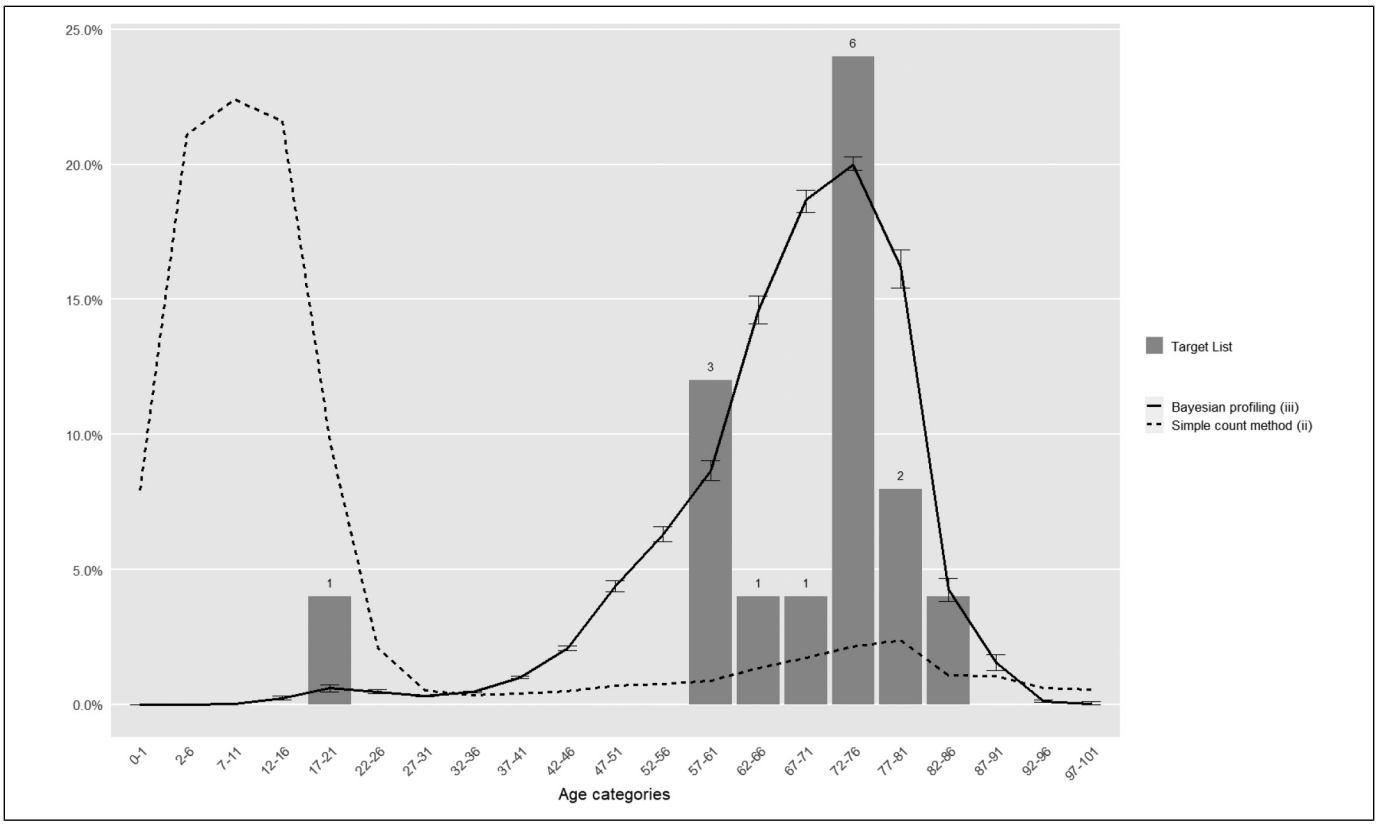

Figure 3. Age pyramids of a customer named Adrien, using the simple count method and Bayesian profiling.
Notes: The histogram corresponds to the actual age of the 15 Adriens found in the target list. The error bars display the 95% credible interval of the posterior.

increased from .1739 with the simple count method to .1977 with Bayesian profiling (a 13.7% improvement). The hit probability (i.e., the average predicted probability of the actual age), improved from .1111 to .1458 (a 31.2% improvement). The predictive log-likelihood improved from −35,113 to −30,987. If we expand predictions to ±1 age category, the hit rate increased from .4511 to .5271 (a 16.8% improvement). The Bayesian profiling method predicted the correct age category (±1) for 6,327 individuals, versus 5,505 for the simple count method. Moreover, the simple count method was also 840% more likely (1,175 vs. 140) to make an erroneous prediction by eight age categories or more than the Bayesian profiling method.

Empirical Illustration #2: Profiling from IP Addresses

In practice, it could be challenging to feature-engineer a Z indicator that captures the influence of X categories (e.g., 9,840 car models, 12,834 first names) into a lower-dimensional scale. In addition, based on domain knowledge and common sense, managers can often assume a priori that conditioning list selection on S (e.g., age, income) will likely lead to far superior results than conditioning list selection on X (e.g., first name, car model). The simulations and first empirical application addressed those common situations.

Our second empirical application covers the more complex case where conditioning list selection on a Z indicator is both feasible and plausible. In this example, as in many realistic business applications, the true distributions of the target list are unknown to both the research team and the sponsor organization. Thus, we cannot report out-of-sample fit measures. We still perform "in-sample" comparisons using Bayesian information criteria (BIC) and Bayes factors.

Data Set

The sponsor organization is a U.S.-based marketing analytic software company that primarily serves the needs of the educational and academic market. The company tracked the IP addresses of its registered users during a year. The original list consisted of 36,036 unique IP addresses worldwide, for a total of 2,512,796 pages visited. For our analyses, we only retained users who visited ten pages or more over the year.

We geolocalized all IP addresses and removed both bots (e.g., search engine spiders) and customers outside the United States for a final list of 10,771 individuals distributed across 2,995 U.S. zip codes. Because customers in the sample retained for analysis were spread out across the entire United States, we selected a reference table that covered the whole country, namely the Census bureau data (note that if customers were all located in California, a reference table limited to the Californian zip codes would have been more appropriate). We cross-referenced the geolocation of these 10,771

Table 6. Results of Various Models for the Second Empirical Application.

Model Alternatives	List-Selection Mechanism	# Param.	Z Indicator[a]	BIC[b]	Bayes Factor (Log)[c]
Simple Selection Models					
Random	P(L)	0	—	(90,464)	—
ZIP code	P(L|X)	33,120	—	(240,034)	(149,570)
Z indicator	P(L|Z)	2	log(Decay 2)	(85,417)	5,047
Conditional on Income					
Income	$P(L\|S_1)$	16	—	(88,592)	1,872
Income + ZIP code	$P(L\|X,S_1)$	33,137	—	(240,122)	(149,658)
Income + Z (additive)	$P(L\|Z,S_1)$	19	log(Decay 2)	(84,910)	5,554
Income + Z (moderator)	$P(L\|Z,S_1)$	48	Decay 5	(88,249)	2,215
Income + Z (both)	$P(L\|Z,S_1)$	51	Nearest	(85,100)	5,364
Conditional on Occupation					
Occupation	$P(L\|S_2)$	13	—	(88,706)	1,758
Occupation + zip code	$P(L\|X,S_2)$	33,134	—	(240,106)	(149,642)
Occupation + Z (additive)	$P(L\|Z,S_2)$	16	log(Decay 2)	(84,840)	5,624
Occupation + Z (moderator)	$P(L\|Z,S_2)$	39	Nearest	(88,168)	2,296
Occupation + Z (both)	$P(L\|Z,S_2)$	42	log(Decay 2)	(84,961)	5,503
Conditional on Education					
Education	$P(L\|S_3)$	4	—	(87,892)	2,572
Education + zip code	$P(L\|X,S_3)$	33,125	—	(240,060)	(149,596)
Education + Z (additive)	$P(L\|Z,S_3)$	7	log(Decay 2)	(84,601)	5,863
Education + Z (moderator)	$P(L\|Z,S_3)$	12	log(Decay 20)	(87,640)	2,824
Education + Z (both)	$P(L\|Z,S_3)$	15	log(Decay 2)	(84,606)	5,858

[a]Represents the operationalization of the Z indicator that provides the best fit.
[b]We rely on BIC because a better estimator for extremely high-dimensional models is out of reach, and BIC allows for comparable fit metrics across models.
[c]Represents the Bayes factor (log) of each model versus the random selection model.

individuals with the publicly available data for the 33,120 zip codes in the United States regarding three distinct S variables: income, occupation, and educational attainment.

Modeling Considerations

This article lists eight possible list-selection mechanisms, numbered L1 to L8. Out of these eight possible mechanisms, three are independent of S: namely L1 as P(L), L2 as P(L|X), and L8 as P(L|Z). The five other models are tested on each of the three candidate S variables. Thus, we report $3 + (3 \times 5) = 18$ competing models.

In terms of Z indicator, given the nature of the company (selling marketing analytics software targeted at the education market), and after discussing with the company's management, we found that the geographical proximity of universities and business schools with strong marketing departments seemed to be a promising indicator to capture the influence of zip codes in the list-selection process. We identified and geolocalized 409 U.S. universities and business schools with a marketing department. We then searched each institution's website to identify its active marketing faculty and found a total of 3,885 individuals. From that raw data, different operationalizations of the Z indicator were possible, such as the distance of each zip code to the nearest university, the number of marketing departments within a 2-, 5-, 10-, 20-, or 50-mile radius, or the number of marketing faculty within the same radius. We also tested an exponential decay function that considers both the number and proximity of marketing faculty. In this configuration, the influence of marketing faculty on the list-selection process has a "half-life" of M miles (i.e., for every additional M miles, the impact on the list selection is cut in half). For instance, if a particular zip code is 33 miles away from a strong marketing department of 23 faculty, and we set $M = 20$, the contribution of that department to the Z indicator is equal to $[23 \times (1/2)^{33/20}] = 7.33$. All 409 universities contribute to the Z indicator of each of the 33,120 zip codes, although their influence decreases exponentially as distance increases. We tested 32 different operationalizations of Z, for a total of 297 models. We report the best ones in Table 6.

Model Selection and Bayes Factors

Among the models that condition on a single variable (e.g., X, S, or Z), the model that conditions on the best available Z indicator achieves the highest log-Bayes factor (5,047), followed by |education (2,572), income, and occupation relative to the baseline of random list selection. The fact that conditioning on the Z indicator performs well indicates that (1) geographic proximity to business schools with large marketing departments indeed influences list selection and that (2) the particular engineered feature captures that influence well. The simple count method and its 33,120 parameters, which implicitly conditions on zip codes, achieves the worst performance of all.

As we expected, all the models that condition list membership on both X and S (L4) achieve a worse BIC than the

simple count method, which conditions on X only (L2). Because conditioning on X alone perfectly rationalizes the likelihood function already, adding S as an additional conditioning argument adds model parameters without improving model fit.

Regarding the models that include both S and Z as conditioning arguments of list selection, they all outperform models that condition on S or Z alone. While assuming the presence of a moderating effect improves model fit, this influence subdues once the model incorporates an additive influence. Mechanisms L5 (i.e., P(L|Z, S) additive) achieve the best BIC across all the sociodemographic variables under consideration.

Profiling Results

We report in Figure 4 the consumer profiling obtained from the simple count method (based on X), the simple Bayesian profiling approach (based on S), and the more complex additive model specification (based on S + Z). Note that all models are independent of one another, and the results should not be interpreted as a *joint* profile. While it would be possible to construct a unique likelihood function to estimate the joint influence of income, occupation, and education on list selection in an additive mixture model similar to model L5, these weights could not be used to reverse-engineer a joint customer profile. However, if the joint distribution of $P(X, S_1, S_2, S_3)$ in the population were publicly available from the Census Bureau, models of joint selection based on all S variables could be readily estimated and translated into joint profiles.

Comparing model assumptions L2 (based on X) and L3 (based on S), the first striking result is that the simple count method predictions closely mimic the distributions of the U.S. population as a whole, consequently providing limited or worse (if taken at face value) misleading insights.

In terms of occupations, the Bayesian profiling concentrates its predictions on three industries only: education or health sector (44.7%); information technology and information systems (38.9%); and professional services, scientific, or managerial roles (16.4%). The simple count method predicts that 56% of the customers in the target list work outside these industries, such as in retail (10.2%), entertainment (11.0%), or agriculture (1.1%). Given the focus of the software company (marketing analytics) and its primary target market (education), the Bayesian profiling results seem to provide higher face validity.

Regarding the education models, the simple count method estimates that 47.6% of the target list is college-educated (bachelor's degree or higher), versus 31.7% for the U.S. population. The simple count method estimates that the proportion of college-educated individuals is higher in the target list than in the U.S. population as a whole. Still, the Bayesian model estimates that this figure is largely underestimated and puts it at 95.5% instead.

The simple Bayesian approach predicts a highly skewed distribution of income as well. In the United States, 11.7% of the population has an annual income of $150,000 or more. The simple count method predicts that this portion of the population is overrepresented in the target list (17.2%). The Bayesian profiling method puts that figure at a striking 69.8%. Because consultants, data scientists, and college professors constitute a large portion of the company's target audience, the latter result seems to provide high face validity as well.

The more complex Bayesian profiling models that hypothesize a joint (additive) effect of S and Z on list selection tell interesting stories. The geographical proximity to a strong marketing department plays an important role in list selection, with η_Z varying between .4738 (occupation) and .6231 (income). The farther away a zip code is from an education hub, the less likely its inhabitants will be selected into the list. Once the contribution of Z is accounted for, however, the (partial) contribution of S becomes even more clear-cut. The education model predicts that selection based on S includes 99% of individuals with a bachelor's degree or more. The S component of the occupation model predicts that all users whose joining the list cannot be explained by their proximity to universities and business schools must have management or scientific roles or work in professional services. Given the specific target audience of this marketing analytic company (in education), and with the knowledge that these predictions have been achieved only by collecting anonymous IP addresses, this is a striking prediction.

Using the Bayesian profiling methodology, the company may aim to (1) capture and geolocalize the IP address of its online visitors and prospective customers, (2) predict visitors' profile by applying Equation 11 (which does not require intensive computations once the weights are estimated), and (3) adapt the online content of its website to its visitors' profile. The company can also use Bayesian profiling to assess potential changes in the composition of the website's audience over longer periods. Finally, it could also assess the effective (differential) positioning of subsites intended for different audiences, or addressing different research techniques and decision problems.

Discussion and Conclusions

In their quest to target and tailor marketing strategies to consumers' specific profiles and needs, many firms use external data sources to infer the most likely demographics, psychographics, spending propensities, and lifestyle characteristics of customers. A massive industry has emerged to provide such data, both online and offline, and to enable firms to portray and target their existing clients, prospective customers, and online visitors much more precisely.

Due to the data scarcity in brokerage firms' databases and various technical constraints (e.g., anonymity, obsolescence of existing data, privacy laws), individual-level data may not always be available, accessible, or retrievable. In such situations, data brokers and firms alike heavily rely on aggregate data to infer individual characteristics.

On the one hand, many question the reliability of programmatic segmentation and inferences from aggregate data (Neumann, Tucker, and Whitfield 2019). In light of our

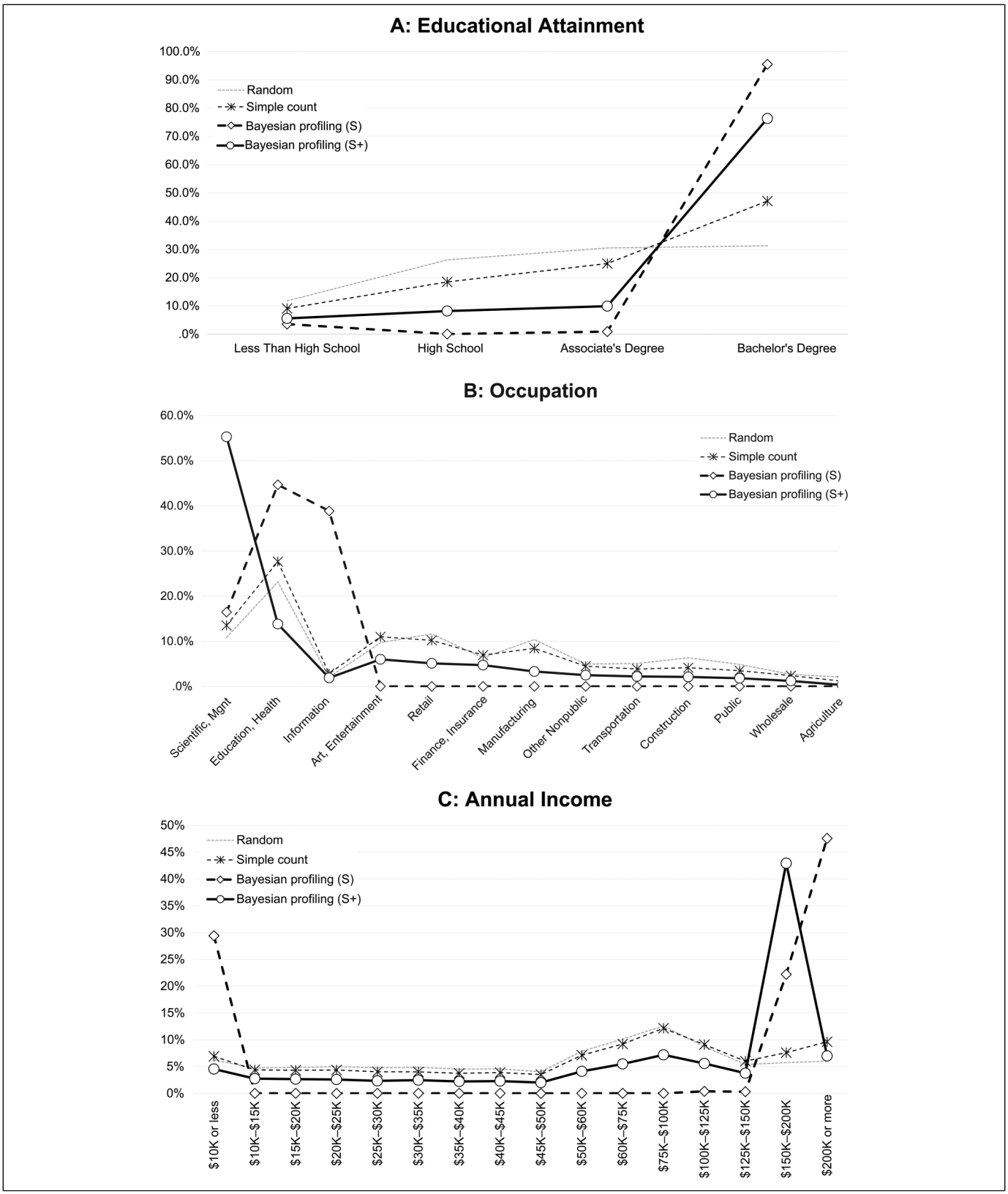

Figure 4. Comparison of the estimated distributions of educational attainment (top), occupation (middle), and annual income (bottom).
Notes: Distributions estimated based on the geolocation of IP addresses of 10,771 U.S.-based customers of a marketing analytic software company. Random selection corresponds to the distributions of the U.S. population as a whole (Census data), provided as benchmark. All profiles are estimated independently of one another.

results, and with the active participation of data brokers, it would be interesting to investigate the extent to which this low reliability is due to individuals being assigned to the wrong segments or to poor individual inferences from aggregate, segment-level data (the topic of this research). While it is not clear that firms that already have access to customers' purchase histories may see value in data augmentation efforts (Rossi, McCulloch, and Allenby 1996), others report that even firms with extensive internal data could benefit from external data sources (Trusov, Ma, and Jamal 2016).

On the other hand, automated consumer profiling raises ethical issues, such as algorithmic discrimination (Lambrecht and Tucker 2018), collection by largely unaudited commercial data brokers (e.g., ChoicePoint) of extremely sensitive personal data on behalf of law enforcement firms (Hoofnagle 2003), or the —sometimes incorrect—inferences of sensitive information (Federal Trade Commission 2014). In particular, is it appropriate to use statistical techniques to infer customers' information when these customers are not willing to share these data in the first place? This question is especially relevant, knowing that using personal information about customers in marketing communications may backfire (Tucker 2014). The notion of "grouplization" advocated by Yahoo! (Chen and Strimaitis 2016) hides the ethical issues more than it solves them (Bisson 2016). In addition, political views differ on how to treat personal data. The European Union has recently adopted the GDPR, and California adopted the Consumer Privacy Act, where the primary objectives are to give back control to individuals over their personal data; in contrast, internet service providers in the United States no longer require customer permission to collect, use, and sell information about their customers' online habits.

In this burgeoning and fast-moving field, we focus on the actual methodology used to infer personal information from aggregate data. We show that the methodology commonly used in the industry to profile a target list is built on the implicit—and often misunderstood and unwarranted—assumption that list membership L and the unobserved variable of interest S (e.g., income, age, occupation, educational attainment) are independent conditional on the observed variable X (e.g., car brand or model, first name, zip code, geolocation inferred from IP address). When this condition is not met because the unobserved variable of interest S influences list membership, biases from the simple count method are extensive. Instead, we develop and expand the profiling consequences of multiple common list-selection assumptions and develop inference for the possibility that selection into a list depends on multiple mechanisms, including lower-dimensional aspects of X that extend beyond S.

It is apparent from our simulation study that the improvements from the proposed Bayesian profiling method are contingent on the relevance of unobserved S as a selection criterion. Our two empirical illustrations indeed constitute examples of relevant selection based on, for example, age and education (as confirmed by Bayes factor analyses; see Web Appendix B).

We also demonstrate that both selections based on observed X and unobserved S contain random sampling from the reference table as a special case. However, neither is a special case of the other.

If it is not clear which list-selection mechanism is at play, we illustrate how to compare list-selection mechanisms statistically. Specifically, Bayes factors can produce positive evidence for the simpler model and allow for the comparison between nonnested models (Kass and Raftery 1995; see Web Appendix B).

The Bayesian profiling method tackles a specific data imputation problem where *completely* missing data (i.e., missing variables) are inferred. The inference extracts information from a reference table, hypothesizing and testing various list-selection models. An interesting situation arises when the analyst has access to a reference table but can also observe some values of S in the customer list. Data may be only partially missing, such as when a firm knows the age of some of its customers (application #1) or when a website knows the education of some of its online visitors (application #2). In those cases, the analyst may also invoke the Little and Rubin (2019) missing data framework to infer missing data from complete cases.

Given consumers' increasing reluctance to share personal information and the burgeoning data brokerage industry, the Bayesian profiling method developed in this article should be a welcome addition to the toolbox of methods for indirect and unobtrusive profiling and targeting. This is especially true in a world where privacy laws (e.g., GDPR) and anonymity software (e.g., private VPN) will make individual data increasingly challenging to collect and use, and where inferences from aggregate data might become even more strategic tomorrow than it is today. As F. Mariet (Director of Weborama) mentioned during an interview we conducted for this research, "In the press, everybody is talking about using individual-level data, but between matching problems, privacy issues, and the like, real improvements in the future may very well come from smarter use of aggregate data instead." We thus have great hopes for the application of the Bayesian profiling method in practice.

Associate Editor

Eric Bradlow

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Arnaud De Bruyn https://orcid.org/0000-0002-4413-6167

References

AutoInsurance Center (2015), "Average Fans of Brand Cars," (accessed December 3, 2021), http://www.autoinsurancecenter.com/the-average-fan-of-car-brands.htm.

Bisson, David (2016), "Yahoo Has a Creepy Plan for Advertising Billboards to Spy on You," Graham Clueley (October 10), https://www.grahamcluley.com/yahoo-creepy-plan-advertising-billboards-spy/.

CACI Information Solutions (2002), "MONICA: Classification of Age by Name for Marketers," marketing brochure, http://www.caci.co.uk.

Chen, Jian and Romualdas Strimaitis (2016), "Measuring User Engagement with Smart Billboards," United States Patent Application Publication, US 2016/0292713 A1 (October 6).

Clotfelter, Charles T. (1980), "Tax Incentives and Charitable Giving: Evidence from a Panel of Taxpayers," *Journal of Public Economics*, 13 (3), 319–40.

Cole, Karen, Rachel Dingle, and Rajesh Bhayani (2005), "Pledger Modelling: Help the Aged Case Study," *International Journal of Nonprofit and Voluntary Sector Marketing*, 10 (1), 43–52.

Crié, Dominique and Andrea Micheaux (2006), "From Customer Data to Value: What Is Lacking in the Information Chain?" *Database Marketing & Customer Strategy Management*, 13 (4), 282–99.

Dias, Felipe F., Patricia Lavieri, Taehooie Kim, and Chandra Bhat (2019), "Fusing Multiple Sources of Data to Understand Ride-Hailing Use," *Transportation Research Record: Journal of the Transportation Research Board*, 2673 (6), 214–24.

Experian (2014), "Mosaic® USA: The Consumer Classification Solution for Consistent Cross-Channel Marketing," commercial brochure (accessed December 3), http://www.experian.com/assets/marketing-services/brochures/mosaic-brochure-october-2014.pdf.

Experian (2019), "Mosaic® USA Consumer Lifestyle Segmentation," (accessed December 3, 2021), http://www.experian.com/marketing-services/consumer-segmentation.html.

Federal Trade Commission (2014), "Data Brokers: A Call for Transparency and Accountability," (May), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.

Feit, Elea M. and Eric T. Bradlow (2018), "Fusion Modeling," in *Handbook of Market Research*, Christian Homburg, Martin Klarmann, and Arnd Vomberg, eds. Cham, Switzerland: Springer.

Geraghty, Kevin, Eric Sonmezer, Matthew Maron, and Daniel Ruble (2017), "360i Generates Nearly $1 Billion in Revenue for Internet Paid-Search Clients," *Interfaces*, 47 (1), 24–37.

Gilula, Zvi, Robert E. McCulloch, and Peter E. Rossi (2006), "A Direct Approach to Data Fusion," *Journal of Marketing Research*, 43 (1), 73–83.

Google (2021), "An Updated Timeline for Privacy Sandbox Milestones," (June 24), https://blog.google/products/chrome/updated-timeline-privacy-sandbox-milestones/.

Greenburg, Zack O'Malley (2009), "What Your Car Says About You," *Forbes* (October 6), https://www.forbes.com/2009/10/06/car-personality-wealth-lifestyle-vehicles-gender-income.html.

Greene, Henry and George R. Milne (2005), "Alternative Data Sources in Targeted Marketing: The Value of Exographics," *Journal of Targeting, Measurement and Analysis for Marketing*, 14 (1), 33–46.

Heckman, James. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1), 153–62.

Hoofnagle, Chris Jay (2003), "Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect, Process, and Package Your Data for Law Enforcement," *North Carolina Journal of International Law*, 29 (4), 595.

Kamakura, Wagner A. and Michel Wedel (1997), "Statistical Data Fusion for Cross-Tabulation," *Journal of Marketing Research*, 34 (4), 485–98.

Kamakura, Wagner A. and Michel Wedel (2000), "Factor Analysis and Missing Data," *Journal of Marketing Research*, 37 (4), 490–98.

Kass, Robert E. and Adrian E. Raftery (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90 (430), 791–95.

Lambert-Pandraud, Raphaëlle, Gilles Laurent, and Eric Lapersonne (2005), "Repeat Purchasing of New Automobiles by Older Consumers: Empirical Evidence and Interpretations," *Journal of Marketing*, 69 (2), 97–113.

Lambrecht, Anja and Catherine Tucker (2018), "Algorithmic Discrimination? Apparent Algorithmic Bias in the Serving of STEM Ads," *Management Science*, 65 (7), 2966–81.

Levitt, Steven and Stephen J. Dubner (2005), *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York: William Morrow/HarperCollins.

Lieberson, Stanley and Eleanor O. Bell (1992), "Children's First Names: An Empirical Study of Social Taste," *American Journal of Sociology*, 98 (3), 511–54.

Little, Roderick J.A. and Donald B. Rubin (2019), *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ: John Wiley & Sons.

McCarthy, Daniel and Elliott Shin Oblander (2021), "Scalable Data Fusion with Selection Correction: An Application to Customer Base Analysis," *Marketing Science*, 40 (3), 459–80.

Merkle Inc. (2017), "Merkle's DataSource—Digital Segments," commercial brochure.

Merkle Inc. (2019), "Merkle Digital Segmentation," commercial brochure.

Neumann, Nico, Catherine E. Tucker, and Timothy Whitfield (2019), "Frontiers: How Effective Is Third-Party Consumer Profiling? Evidence from Field Studies," *Marketing Science*, 38 (6), 918–26.

Rossi, Peter E., Rober E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321–40.

Steenburgh, Thomas J., Andrew Ainslie, and Peder Hans Engebretson (2003), "Massively Categorical Variables: Revealing the Information in Zip Codes," *Marketing Science*, 22 (1), 40–57.

Trusov, Michael, Liye Ma, and Zainab Jamal (2016), "Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting," *Marketing Science*, 35 (3), 405–26.

Tucker, Catherine E. (2014), "Social Networks, Personalized Advertising, and Privacy Controls," *Journal of Marketing Research*, 51 (5), 546–62.

Van Dijk, Bram and Richard Paap (2008), "Explaining Individual Response Using Aggregated Data," *Journal of Econometrics*, 146 (1), 1–9.

Wachtel, Stephan and Thomas Otter (2013), "Successive Sample Selection and Its Relevance for Management Decisions," *Marketing Science*, 32 (1), 170–85.

Webber, Richard (2007), "Using Names to Segment Customers by Cultural, Ethnic or Religious Origin," *Journal of Direct, Data and Digital Marketing Practice*, 8 (3), 226–42.

Weisbaum, Herb (2014), "Here's the Real Score: Big Data Knows Everything About You!" NBC News (April 9), http://www.nbcnews.com/business/consumer/heres-real-score-big-data-knows-everything-about-you-n75741.