Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 875–895
https://doi.org/10.1093/jrsssa/qnae082
Advance access publication 2 September 2024
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

kpop: a kernel balancing approach for reducing specification assumptions in survey weighting

Erin Hartman[1], Chad Hazlett[2,3] and Ciara Sterbenz[3]

[1]Department of Political Science & Department of Statistics, University of California, Berkeley, USA
[2]Department of Statistics and Data Science, University of California Los Angeles, USA
[3]Department of Political Science, University of California Los Angeles, USA

Address for correspondence: Erin Hartman, Department of Political Science & Department of Statistics, University of California, 210 Social Sciences Building, Berkeley, CA 94720, USA. Email: ekhartman@berkeley.edu

Abstract

With the precipitous decline in response rates, researchers and pollsters have been left with highly nonrepresentative samples, relying on constructed weights to make these samples representative of the desired target population. Though practitioners employ valuable expert knowledge to choose what variables X must be adjusted for, they rarely defend particular functional forms relating these variables to the response process or the outcome. Unfortunately, commonly used calibration weights—which make the weighted mean of X in the sample equal that of the population—only ensure correct adjustment when the portion of the outcome and the response process left unexplained by linear functions of X are independent. To alleviate this functional form dependency, we describe kernel balancing for population weighting (*kpop*). This approach replaces the design matrix $\mathbf{X}$ with a kernel matrix, $\mathbf{K}$ encoding high-order information about $\mathbf{X}$. Weights are then found to make the weighted average row of $\mathbf{K}$ among sampled units approximately equal to that of the target population. This produces good calibration on a wide range of smooth functions of X, without relying on the user to decide which X or what functions of them to include. We describe the method and illustrate it by application to polling data from the 2016 US presidential election.

Keywords: balancing weights, calibration, nonresponse, survey weighting

1 Introduction

In an era of decreasing response rates, social scientists must rely on methods to adjust for the non-representative nature of survey samples. For example, Pew Research Center saw response rates to live-caller phone surveys decline from nearly one-third of respondents in the late 1990s, to only 6% in 2018 (Kennedy & Hartig, 2019). The nonrandom nature of this 'unit nonresponse' poses serious challenges for survey researchers and has led to greater use of nonprobability sampling methods, such as panel, quota, or river sampling for online surveys (Mercer et al., 2017). The concern, whether due to nonresponse or nonprobability sampling, is that the resulting survey respondents are not representative of the target population about which a researcher aims to draw an inference, leaving the potential for significant bias in estimates of target outcomes.

Researchers are, therefore, often obligated to construct survey weights to address this bias. Constructing these weights requires researchers to choose (1) what variables to account for in the weighting procedure, and (2) how to incorporate these variables in the construction of survey weights. For example, researchers have determined pollsters' failure to account for educational attainment in survey weighting resulted in inaccurate predictions leading up to the 2016 US Presidential election. Even those that did account for educational attainment often failed to

Received: July 8, 2021. **Revised:** July 22, 2024. **Accepted:** July 23, 2024
© The Royal Statistical Society 2024. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

account for low levels of Midwestern, white voters with lower levels of educational attainment, i.e. the interaction of region, race, and educational attainment (Kennedy et al., 2018). We return to this issue in our application, demonstrating how our proposed method can address these concerns.

We begin with the observation that in practice, researchers seek to make the sample and target population identical only on some summary of the characteristics, represented by the matrix $\mathbf{X}$. The variables taking the columns in $\mathbf{X}$ may include indicators for membership in intersectional strata, and/or the values of other variables. In practice, $\mathbf{X}$ is typically (i) low-dimensional and (ii) chosen or constructed by hand. Weights are then chosen to make the sample similar to the target population in terms of the means of these $\mathbf{X}$, thereby neglecting other moments of $p(\mathbf{X})$ that, unnoticed, can remain dissimilar between the weighted sample and target population.

Unfortunately, except in the case of full saturation (i.e. every combination of $\mathbf{X}$ values can be represented by an indicator), investigators are not generally in a position to argue that the outcome or response process is linear in such $\mathbf{X}$, which is needed to achieve unbiasedness through this adjustment (see e.g. Kott & Chang, 2010; Särndal & Lundström, 2005, with analogous results in the causal inference setting, see e.g. Zhao & Percival, 2016). Note that 'response', in this context, means that a participant was both sampled and responded and thus appears in the observed data. Further, leaving the choice of which variables and which higher-order terms to include in the hands of investigators allows almost unlimited researcher degrees of freedom. As we show, even across a set of seemingly reasonable choices, the resulting estimates can vary widely.

The question is then how researchers can choose what functions of covariates, $\phi(\mathbf{X})$, should be used for constructing weights. We provide one reasoned answer to this question, aiming to require the weakest workable assumptions and minimal user intervention. To do so first requires a clarification of how the nonparametric identification assumptions invoked to handle nonresponse become parametric assumptions once we are also constrained by estimation concerns. Specifically, we formulate the 'linear ignorability' assumption, which states that survey weights unbiasedly estimate the desired outcome among the target population only when the part of the outcome not explained by a linear combination of the $\phi(\mathbf{X})$ is independent of the part of the sampling process not explained linearly by $\phi(\mathbf{X})$ within a suitable link function. As we detail below, this refines and weakens existing results that call for both nonparametric ignorability of selection and linearity of the outcome (or selection model) in $\phi(\mathbf{X})$ as separate matters.

Our main contribution is to propose a specific kernel-based weighting procedure (*kpop*) as a practical estimation procedure that reduces bias by more nearly meeting the linear ignorability assumption. In short, this approach employs the kernel matrix, $\mathbf{K}$, whose linear span captures a wide range of smooth, nonlinear functions of $\mathbf{X}$. Weights are then chosen to make the weighted average row of $\mathbf{K}$ in the sample approximately equal to the average row of $\mathbf{K}$ in the target population. Weights chosen in this way approximately equate the (weighted) distribution of $\mathbf{X}$ in the survey with that of the target group, as would be estimated by a kernel density estimator (Hazlett, 2020).

In what follows, Section 2 establishes our setting and notation. Section 3 reviews calibration estimators and discusses identification, introducing a new minimal identification requirement. Section 4 describes our proposed kernel-based calibration estimator, while Section 5 demonstrates its behaviour in two simulated examples. We apply the technique to the 2016 US Presidential election in Section 6, showing how *kpop* can be used to mitigate concerns due to limited foreknowledge of what interactions or intersectional strata are important. Section 7 concludes.

2 Setting and notation

Our setting considers two primary objects. The first is a sample of the form $\{X_i, Z_i, Y_i, R_i\}_{i=1}^{N_s}$, where Y_i is the outcome of interest, and X_i is a collection of P auxiliary variables (covariates) that will be adjusted for. The auxiliary data initially encoded as X may be mapped to a richer feature expansion $\phi(X)$, with $X \mapsto \phi(X)$ from $\mathbb{R}^P \mapsto \mathbb{R}^{P'}$, potentially with $P' \gg P$. In the survey setting, typically many or all dimensions of X are categorical, as in education, party identification, etc. We will consider such cases here, though the methods described are equally natural for continuous variables. R_i is an indicator for selection into the sample with $R_i = 1$ for all units in the sample. Z_i is included in each tuple here to represent a potentially important unobserved factor, which

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

will become important when considering the conditions that will lead to biased estimates. Each tuple in this sample is presumed to be drawn independently from an unknown joint density.

The second object of interest is a larger but still finite target group or population. This is a collection $\{X_i, Z_i, Y_i, R_i\}_{i=1}^{N_{pop}}$, with $N_{pop} \gg N_s$. Critically, each tuple in this collection is drawn from a common joint density $p(X, Z, Y, R)$. Depending on field and custom, this common joint density is sometimes referred to as the data generating process, the true population of interest, or the super-population from which the target group or population was drawn but from which many others could have in theory been drawn. Note that X_i must be observed for all units in both groups to allow adjustment. However, Y is unobserved in this target group. The immediate target of inference is the mean of Y in this larger group. In some cases, this target group is identical with the ultimate target population of interest (e.g. when it is a census). In other cases, the mean of Y over the larger group is of interest principally as an estimate of the expectation of Y over $p()$. For example, the target may be a very large representative survey from a national population, as in our application. While this poses no problem in terms of bias (provided the survey is indeed representative of the referenced population), note that we do not consider here how one's uncertainty estimate would change when targeting the population mean rather than the mean over the observed target group (though see Opsomer & Erciulescu, 2021). In addition, context will determine whether the smaller survey sample is a subset of the larger, or if the smaller survey sample is independently or disjointly drawn. We proceed in the latter setting for congruence with our applied example and for notational simplicity. However, the former is easily accommodated and requires only small changes; see Appendix A, online supplementary material.

The key problem to address is that, due to the nature of data collection—e.g. the possibility of selective response to the survey—the finite sample is drawn from a different distribution than the $p(X, Z, Y, R)$ describing the target group, and the user has at best incomplete knowledge as to how these distributions differ. Our goal is to estimate weights such that the weighted mean of Y in the sample is the best possible estimate for the mean of Y in the target group.

3 Estimation and identification

3.1 Calibration estimators

Suppose researchers have only a few variables in X, and each is discrete with a small number of categories. In such settings, it is straightforward to adjust for X without any functional form or specification commitments: one can take the sample data, average the Y within each stratum of X, then reaverage these strata-wise averages together according to how often each stratum appeared in the target group. This is the *poststratification* estimator, and it can provide an unbiased estimate of the mean of Y in the target group under the nonparametric identification assumption that conditional on X, Y is independent of R (conditional ignorability; see below).

Unfortunately, such an approach is often infeasible because X contains one or more continuous variables, and/or some strata that may be nonempty in the target group are empty in the sample. In such cases, investigators most often turn to calibration estimators, which construct weights on the sampled units such that the weighted mean of $\phi(X)$ among the sample equals the mean of $\phi(X)$ among the target population, where $\phi(X)$ represents some chosen transformation of the original X. In general form, calibration weights w are estimated according to:

$$\min_{w} \; D(w, q) \quad \text{s.t.} \sum_{i:R_i=1} w_i \phi(X_i) = T, \quad \sum_{i:R_i=1} w_i = 1, \;\; \text{and } 0 \le w_i \le 1 \tag{1}$$

where q_i refers to a reference or base weight and $D(\cdot, \cdot)$ corresponds to a distance or divergence metric, acting as a measure of how extremely the weights diverge from q_i. In principle q_i may be the design weights for the sampling strategy employed. However, these are often unknown or unavailable, or far less influential after conditioning on the auxiliary variables. One may then let $q_i = 1/N_s$ be the uniform base weight for units in the respondent sample. The vector T describes target population moment constraints based on the mapping $\phi(X)$. Typically, and in our case, this is an average of $\phi(X)$ in the target population, which we treat as known, but which may be estimated, in which case that additional uncertainty should be propagated (Opsomer &

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Erciulescu, 2021). In other words, the constraint $\sum_{i:R_i=1} w_i\phi(X_i) = T$ in (1) is the 'balance condition' to be satisfied,

$$\sum_{i:R=1} w_i\phi(X_i) = \frac{1}{N_{pop}} \sum_{j=1}^{N_{pop}} \phi(X_j) \tag{2}$$

We note that meeting the conditions above, particularly the balance condition, is not always feasible, and becomes less feasible as the dimensionality of $\phi(\cdot)$ grows. Addressing this will require some combination of relaxing the balance constraints (i.e. 'approximate balance'), or reducing the richness of $\phi(\cdot)$. These trade-offs must be managed by any practical proposal.

Common types of survey weighting correspond to different distance metrics $D(\cdot, \cdot)$, and are closely related to generalized regression estimation (Särndal, 2007). We use $D(w, q) = \sum_{i:R=1} w_i log(w_i/q_i)$, commonly employed in 'raking' methods and variably known as exponential tilting (Wu & Lu, 2016), maximum-entropy weighting, or entropy balancing (Hainmueller, 2012). Other distance metrics correspond to other common weighting estimators, although the choice of distance metric matters far less than the choice of moment constraints (Deville & Särndal, 1992).

For broader reviews of calibration, see Särndal (2007), Caughey et al. (2020), or Wu and Lu (2016). The constraints in (1) ensure the weights are nonnegative and sum to one, ensuring they have a probability-like interpretation. Relaxing this constraint, e.g. allowing negative weights, would be to allow for 'extrapolation' beyond the support of the respondent sample. This increases the possibility of severe model dependency, but is employed in some techniques such as generalized regression estimators (Deville & Särndal, 1992).

Finally, the average outcome among the target population, $\mu = \frac{1}{N_{pop}} \sum_i Y_i$, is then estimated as

Estimator 1 (Calibration estimator for target population mean of Y).

$$\hat{\mu} = \sum_{i:R_i=1} w_i Y_i$$

with weights chosen by equation (1)

In principle, calibration is a general and powerful tool given the flexibility of the choice of $\phi(X)$. In practice, however, most applications of calibration simply seek to match the means of X in the sample to that of the population, i.e. $\phi(X) = X$. We refer to this as *mean calibration*, understanding that 'mean' refers to the original X. Such an approach holds intuitive appeal since, at minimum, pollsters and investigators seek to adjust a sample to closely match a target population on the margins, particularly on variables such as the proportion falling in some demographic or descriptive group. The risk, however, is that there is little reason to expect key identification assumption to hold under mean calibration—a problem we turn to now.

3.2 Identification: from nonparametric ignorability to 'linear ignorability'

Under what assumptions regarding the data-generating process is it possible for the calibration estimator to unbiasedly estimate the average Y in the target group? Typically this question is answered by appealing to the ignorability of the response conditionally on the observed covariates used for adjustment (Little & Rubin, 2019),

Assumption 1 (Nonparametric ignorability of response).

$$Y \perp\!\!\!\perp R \mid X$$

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

A key shortcoming of this identification strategy is simply that investigators cannot typically invoke nonparametric conditioning in practice, and alternatives such as calibration end up calling for different or additional assumptions. That is, supposing the target group and sample are disjoint, under nonparametric ignorability (Assumption 1) the target is given by

$$\mathbb{E}[Y \mid R=0] = \sum_{x \in \mathcal{X}} \mathbb{E}[Y \mid X=x] p(X=x \mid R=0) = \sum_{x \in \mathcal{X}} \mathbb{E}[Y \mid X=x, R=1] p(X=x \mid R=0) \quad (3)$$

Note that the poststratification estimator is simply the empirical analogue to the right-hand side of (3). Unfortunately, as noted above, these nonparametric assumptions and procedures are insufficient in many practical cases because values of X can appear in the target group that don't appear in the sample. This is certain to happen with continuous valued X, but also likely to happen even with discrete variables with more than a few dimensions and categorical values of discrete X. Turning to calibration estimators, we must import additional or different assumptions to achieve identification. We, therefore, rely on an assumption we term 'linear ignorability',

Assumption 2 (Linear ignorability in $\phi(X)$). Without loss of generality, let $Y_i = \phi(X_i)^\top \beta + \epsilon_i$, and the probability of unit i being sampled generated by $Pr(R_i = 1) = g(\phi(X_i)^\top \theta + \eta_i)$, where $g(\cdot) : \mathcal{R} \mapsto [0, 1]$. Linear ignorability holds when $\epsilon_i \perp\!\!\!\perp \eta_i$.

In words, this requires the part of Y not linearly explainable by (i.e. orthogonal to) $\phi(X)$ to be independent of the part of the response process not linearly explained by $\phi(X)$ via a suitable link function. Under linear ignorability [with a given choice of $\phi(\cdot)$], a feasible calibration estimator using that choice of $\phi(\cdot)$ will be unbiased:

Proposition 1 (Unbiasedness of calibration under linear ignorability). Under linear ignorability in $\phi(X)$ (Assumption 2) the calibration estimator using weights chosen by equation (1) will be unbiased for the target population mean of Y.

Proof of Proposition 1 can be found in Appendix A, online supplementary material. Linear ignorability's connection to existing thinking can be found in the two well-known special cases that it covers, each of which is sufficient but neither of which is necessary to satisfy linear ignorability. At one extreme would be the assumption that Y is truly linear in X without unobserved confounders of X, meaning ϵ is fully independent of any other variable in the system, including the usual 'conditional independence assumption', $\mathbb{E}[\epsilon \mid X] = 0$. It is important to note that to write '$Y = \phi(X)^\top \beta + \epsilon$' in Assumption 2 is not to require this, but only to invoke the decomposition of Y into a component in the span of $\phi(X)$ and a residual piece ϵ. Linear ignorability is slightly weaker as it requires only that the ϵ formed by removing what is linearly explainable by $\phi(X)$ is independent of η. That is, there can be unobserved confounders of X and Y here, which would appear as ϵ values that are not independent of X, but these are not problematic unless they are correlated with the unmodelled influences on selection, found in η.

At the other extreme, one could assume *link-linearity of the response model*, such that η_i is independent of any variable in the system. Such assumptions are standard in prior work on calibration such as Särndal and Lundström (2005), Kott and Chang (2010), and Zhao and Percival (2016), though the choice of link functions in these is sometimes more restrictive than the general case we show here (see Appendix A.1, online supplementary material). Linear ignorability is weaker than such an assumption, similarly, in that it requires only the part of $g^{-1}(Pr(R=1))$ not in the span of $\phi(X)$ to be independent of ϵ. That is, there may be systematic influences on selection that go unmodelled and appear in η, so long as these are unrelated to the unmodelled influences of Y, found in ϵ.

3.3 Bias due to violating linear ignorability

To clarify the commitments one makes by subscribing to the linear ignorability assumption, we illustrate how it might be violated. Consider the decomposition of Y as $\phi(X)^\top \beta + (Z + v)$. The ϵ

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

invoked in Assumption 2 is composed here of $Z + v$. Here, v is entirely exogenous random noise; Z is unobserved and, without loss of generality, orthogonal to $\phi(X)$ because it could equivalently be replaced by the residual from projecting Z onto $\phi(X)$. Whether linear ignorability holds is determined by Z's role in the selection process. If Z was purely exogenous random noise (like v) then $\epsilon = Z + v$ will be independent of η in the equation for R, satisfying Assumption 2. By contrast, if this Z is associated with R (and thus η, since it is independent of X), then Z would cause a violation of Assumption 2.

Problematic variables Z could take on two forms. First, there could be important omitted variables, which would also violate nonparametric ignorability (Assumption 1). Unobserved factors outside of $\phi(X)$ could be relevant to both R and to Y, thus entering into both ϵ and η, causing them to be correlated. For example, an individual's general level of interest in politics is predictive of many policy positions, and the strength of those preferences, in American politics. It is also highly predictive of response probability to political surveys, with those interested in politics overrepresented in respondent samples. Because political interest is not measured in many datasets used to define target populations, such as those defined by administrative records, it is an example of an unmeasured confounder Z that could violate both nonparametric ignorability and linear ignorability. No adjustment technique would eliminate bias in this scenario, but sensitivity analyses provide a natural approach to addressing potential remaining bias from such confounders (e.g. Hartman & Huang, 2023). We note that bias is generated only by the part of political interest orthogonal to the linear relationship with the included auxiliary variables in $\phi(X)$.

The second form of problematic Z would be one that generates a 'specification failure'. Suppose we did not omit any variable 'important to' R and Y, but Z is a nonlinear function of $\phi(X)$ [orthogonal to what is in $\phi(X)$], that is relevant to both Y and R. Z would then appear in both ϵ and η, driving their association. This is of particular concern for the commonly used mean calibration in which $\phi(X) = X$. This form of Z is difficult to rule out: investigators may suspect the outcome to 'involve' an X corresponding to some concept, but can rarely make strong arguments for the functional relationship to R and Y, or justify a particular link function for R. Such a problematic form of Z is the main motivation for our approach. Examples of how such a Z emerges, the bias it generates, and the *kpop* solution to it, are illustrated in first a simple reductive simulation and then a more complex one in Section 5.

4 Proposal: kernel-based weighting (*kpop*)

In this section, we present the technical details of our proposal, *kpop*. Readers who wish to work from a more concrete example may prefer to look first to Section 5.1, then return to this section for more formal details.

Many reasonable proposals are possible for how to choose $\phi(X)$ so as to mitigate violations of linear ignorability and the consequent bias. In plain terms, we want $\phi(X)$ to capture any (potentially nonlinear) systematic relationship between Y and X and/or R and X. This would expunge problematic 'Z' variables from ϵ and/or η, so that such a Z can no longer drive an association of ϵ with η, thereby achieving linear ignorability. *kpop* is designed to reduce the risk and magnitude of violating linear ignorability, with minimal user intervention, by replacing the design matrix $\mathbf{X}$ with a kernel matrix $\mathbf{K}$ that represents a rich choice of $\phi(\cdot)$. In this section, we present the technical details of our proposal, *kpop*.

4.1 Motivation for kernels through models

One way to motivate the use of kernels is through considering how they determine the choice of $\phi(\cdot)$ in the context of linear models. Consider linear functions of $\phi(X)$ that explain either the outcome or the response probability [transformed as $g^{-1}(Pr(R = 1))$] to be linear in $\phi(X)$. For simplicity, we consider the outcome. Consider the regularized regression problem

$$\arg\min_{\theta \in \mathbb{R}^{p'}} \sum_i (Y_i - \phi(X_i)^\top \theta)^2 + \lambda \theta^\top \theta \tag{4}$$

To be clear, this model will not actually be estimated in our setting. Rather, it describes an assumption about the space in which the conditional expectation function for the outcome falls. In doing

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

so it calls for a set of basis functions, $\phi(X)$, in which the outcome is assumed to be linear. The very same $\phi(X)$ are those on which mean balance will need to be achieved, as we describe below.

Ideally our choice of $\phi(X)$ would be one that includes very general, high-dimensional, nonlinear expansions of X. Fortunately, certain choices of $\phi(X)$ can be high- or infinite-dimensional, yet admit an N_s-dimensional representation of the data that can then be employed in calibration. Intuitively kernel functions 'compare' two observations, X_i and X_j by computing $k(X_i, X_j) : \mathbb{R}^P \times \mathbb{R}^P \mapsto \mathbb{R}$. A kernel function $k(\cdot, \cdot)$ is positive semidefinite if the kernel matrix it creates, $\mathbf{K}$, satisfies $a^\top \mathbf{K} a \geq 0$ for all real vectors a. For such positive semidefinite kernels, the value of $k(X_i, X_j)$ corresponds to choices of $\phi(X_i)$ through the relationships $k(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle$. As is well known and can readily be shown from first principles (see e.g. Hainmueller & Hazlett, 2014), the solution to equation (4) admits to the form $\theta = \sum_i c_i \phi(X_i)$, and consequently the predictions for Y_i are then given by $\phi(X_i)^\top \theta = \sum_j c_j k(X_i, X_j)$. Forming the kernel matrix $\mathbf{K}$ with entries $\mathbf{K}_{i,j} = k(X_i, X_j)$, this can be rewritten as $K_i^\top c$ where K_i is the ith row of $\mathbf{K}$, or the vector of predictions $\hat{\mathbf{Y}}$ is simply $\mathbf{K}c$.

The vital feature of this result is simply that *the functions linear in $\phi(X_i)$ have been replaced with those linear in $\mathbf{K}_i$*, or equivalently, the linear span of $\phi(X)$ is covered by the linear span of $\mathbf{K}$. This holds regardless of the dimensionality of $\phi(\cdot)$. Thus, to gain all the benefits of $\phi(\mathbf{X})$—whether for modelling or calibration purposes—one need only work with $\mathbf{K}$.

Here, we employ the Gaussian kernel,

$$k(X_i, X_j) = \exp(-\|X_i - X_j\|^2 / b)$$

where $\|X_i - X_j\|$ is the Euclidean distance. While no single choice of kernel can rigorously be established as optimal across settings or even in a particular application, the Gaussian kernel typically serves as the 'workhorse' kernel for a wide variety of kernel-based procedures. One reason for this is that the implicit $\phi(\cdot)$ for the Gaussian kernel is infinite-dimensional and has the 'universal representation property', such that as the number of sample points goes to infinity, every continuous function will be linear in these features (Micchelli et al., 2006). The values of $k(X_i, X_j)$ can readily be interpreted as a distance or similarity measure between X_i and X_j, with $k(X_i, X_j) = 1$ only when $X_i = X_j$, i.e. that the covariate profiles match exactly. The rate at which $k(X_i, X_j)$ approaches zero when X_i and X_j differ is dictated by the choice of b, which we discuss below. A linear combination of the elements of K_i is thus a weighted sum of unit i's similarity to every other unit j in the sample, where similarity is measured by centreing a Gaussian kernel over each X_j and measuring its height at X_i. Hainmueller and Hazlett (2014) provides further description and illustration of this function space.

4.2 The ideal *kpop* estimator

In this section, we describe the 'ideal' *kpop* estimator, which will be revised below to an approximate version. Replacing $\phi(X_i)$ with K_i, we seek to satisfy equation (2) by choosing weights that achieve,

$$\sum_{i:R=1} w_i K_i = \frac{1}{N_{pop}} \sum_{j=1}^{N_{pop}} K_j, \quad \text{s.t.} \ \sum_i w_i = 1, \ w_i \geq 0, \ \forall i \tag{5}$$

Note that every K_i here is a transformation of X_i that compares unit i to each of the units in the survey sample. The matrix $\mathbf{K}$ has a row for every unit in the sample and in the target population, yielding dimensions $N_s + N_{pop}$ by N_s. The term on the right gives an (unweighted) average row of K_j for units in the target population. Note that each K_j is an N_s-vector, with the ith element indicating how similar unit j in the target population is to unit i in the survey sample, i.e. $k(X_j, X_i)$. The term on the left is a weighted average of K_i over the survey sample. Here too each K_i is an N_s-vector, with the lth element indicating how similar unit i in the survey sample is to unit l in the survey sample, i.e. $k(X_i, X_l)$.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

In cases where (known) weights $w^{(pop)}$ are used to adjust the target population itself—as in our application below—then *kpop* would instead seek weights that bring the weighted means of K_i among the sampled units to approximately equal the $w^{(pop)}$-weighted means of K_i in the target population. Thus the weighting condition in (5) becomes

$$\sum_{i:R=1} w_i K_i = \sum_{j=1}^{N_{pop}} w_j^{(pop)} K_j, \quad \text{s.t. } \sum_i w_i = 1,\ w_i \geq 0,\ \forall i \tag{6}$$

This formulation is thus more general. We also include w^{pop} in describing the bias bound and approximation routine below to accommodate cases where it is required.

Calibrating through this kernel transformation achieves balance on a wide range of nonlinear functions of X, without requiring the researcher to prespecify them. For example, as we will show in Section 6, *kpop* achieves balance on the interaction of educational attainment, region, and race/ethnicity in a 2016 US Presidential survey without requiring the researcher to have foreknowledge of its importance, much less requiring specific knowledge that Midwestern, white voters with lower levels of educational attainment must be accounted for in the survey weights to yield accurate national predictions.

Another view of what these weights achieved, discussed in Hazlett (2020), is that approximate balance on a kernel transformation approximately equates the multivariate distribution of X in the two groups, *as it would be estimated by a corresponding kernel density estimator*. We also note closely related work on kernel-based balancing and imbalance metrics including Wong and Chan (2018), Yeying et al. (2018), and Kallus (2020).

4.3 The necessity of approximate balance

Weights that achieve equal means on every column of $\mathbf{K}$ are often infeasible. Even where this can be achieved within numerical tolerance, such calibration could lead to extreme weights. Instead, we use approximate calibration weights designed to minimize the worst-case bias due to remaining miscalibration. While numerous approximation approaches are possible, we use a spectral approximation. Specifically, we use singular value decomposition to decompose $\mathbf{K}$ into the matrix product $\mathbf{VAU}^\top$. Singular value decomposition is similar to eigendecomposition, but works for nonsquare matrices. In this arrangement, each column $\mathbf{V}$ is orthogonal to all others and is a linear combination of the original columns of $\mathbf{K}$, similar to the role eigenvectors. The columns of $\mathbf{V}$ are also closely related to the principal components in principal component analysis. $\mathbf{A}$ is a diagonal matrix whose entries (the 'singular values') indicate the 'importance' of each singular vector, akin to the eigenvalues.

Even granting that the linear ignorability assumption holds, approximate balance means the calibration step is not complete, which can introduce *additional* bias, referred to here as the approximation bias. The worst-case bound on this approximation bias is given by Hazlett (2020)

$$\sqrt{\gamma}\|(w^{(pop)\top}\mathbf{V}_{pop} - w_s^\top \mathbf{V}_s)\mathbf{A}^{1/2}\|_2 \tag{7}$$

where $\mathbf{V}_{pop}$ is the matrix containing the rows of $\mathbf{V}$ corresponding to target population units, $\mathbf{V}_s$ contains the rows of $\mathbf{V}$ corresponding to sampled units, and $\mathbf{A}$ is the diagonal matrix of singular values. In this bias bound, $w^{(pop)}$ denotes the (optional) known weights for adjusting the target population itself. The scalar γ is the (reproducing kernel Hilbert) norm on the function, equal to $c^\top \mathbf{K} c$ effectively describing how complicated or 'wiggly' the chosen function is. This is an unknown constant that need not be estimated during the optimization we describe below.

We make three remarks on the form of this worst-case approximation bias in equation (7). First, the L_2 norm of the regression function ($\sqrt{\gamma}$) controls the overall scale of potential bias. Second, the imbalance on the left-singular vectors of $\mathbf{K}$ after weighting, $(w^{(pop)\top}\mathbf{V}_{pop} - w_s^\top \mathbf{V}_s)$, enters directly. Third, the impact of imbalance on each singular vector is scaled by the square root of the corresponding singular value.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

The third point, in particular, suggests the approximate balancing approach we use: calibrate to obtain nearly exact balance on the first r singular vectors (columns of $\mathbf{V}$), leaving the remaining $(r+1$ to $N_s)$ columns uncalibrated. The choice of r is then chosen to minimize the bias bound [equation (7)]. In practice, the singular values of a typical matrix $\mathbf{K}$ decrease very rapidly (see Appendix B.1, online supplementary material for an illustration from the application below). Thus, balance on relatively few singular vectors achieves much of the goal, though the procedure continues beyond this to minimize the worst-case bias bound in equation (7) directly.

'Mean first' *kpop* A 'no-worse' solution. Achieving approximate balance on $\mathbf{K}$ will typically yield good, but not perfect balance on the means of the original variables, $\mathbf{X}$. In practice a visible difference in means or proportions on a given variable can be unsettling: researchers and pollsters may reasonably hope for nearly exact mean calibration on variables of known importance to the outcome of interest, even if the means are in fact no more important to balance than unseen higher-order moments. Further, it may be useful to know that a given estimator achieves balance on the same moments as conventional raking or mean calibration, in addition to possibly calibrating higher-order moments.

To this end, we advocate for using a 'mean first' procedure, in which the weights are constrained to obtain equal means (within a set tolerance) on a chosen set of variables $\mathbf{X}$, in addition to calibrating on r singular vectors of $\mathbf{K}$ chosen so as to minimize the bias bound described above. The cost of enforcing mean balance is that there may be fewer dimensions of $\mathbf{K}$ that can additionally be balanced on within feasibility constraints. Nevertheless, the virtue of this approach—at least as a transitional methodology—is that in terms of balance and anticipated bias, it is arguably 'no worse' than the conventional approach of calibrating on the means of $\mathbf{X}$ alone. For improved stability and performance in practice, we recommend an approximate balancing approach that appends the left-singular vectors of the chosen set of $\mathbf{X}$, choosing the number to balance on by minimizing the worst-case bias bound in equation (7). We refer to this as *kpop + mf* below.

Inference. Following the calibration weighting literature, we use a linearized variance estimator (Fuller, 1975; Kott, 2016). Due to the often large number of dimensions of $\mathbf{K}$ chosen by the method described in Section 4.3 for the *kpop* calibration constraints, we use a ridge regularized regression of the outcome on the r selected columns of $\mathbf{K}$, leaving any columns corresponding to 'mean first' constraints unregularized if they are included. In Appendix D.3.3, online supplementary material, we show the performance of these standard errors, which accurately estimate the empirical standard error and achieve near-nominal coverage rates in our simulations.

4.4 Choice of kernel, data scaling, and b

One obstacle to adopting kernel-based methods is that while they can greatly reduce researcher degrees of freedom in terms of selecting $\phi(\cdot)$, they do still require choosing the kernel function and the value of any of its hyperparameters. In this work, we employ the Gaussian kernel, which we regard as reasonable on account of its universal representation property (Micchelli et al., 2006). There are important considerations regarding how $\mathbf{X}$ is scaled, and relatedly, the choice of the Gaussian kernel bandwidth, b.

Data scaling. Prior to constructing the kernel, continuous valued X are scaled to have a variance of one. Such a choice is convenient as it ensures no 'unit of measure' choice will affect the results. Under this standardization, a one-standard-deviation difference on a given continuous covariate will add one to the squared Euclidean distance that forms the numerator of the exponents in the kernel function. For categorical variables, there is no added distance between two observations that have the same value of a given variable. However, if two units have different values on a categorical variable, we scale the data such that it adds a distance of one to the numerator of the exponent in the kernel function. This is one choice that keeps categorical and continuous variables on reasonable relative scales in terms of their influence in the kernel function. We can achieve this scaling numerically simply by (i) one-hot encoding all binary and categorical variables (without dropping a level) and (ii) rescaling those one-hot encoded indicators by $1/\sqrt{2}$.

Kernel sampling and feasibility. In the present setting, we rely only on the observations in the sample to formulate the columns of a kernel matrix. This is because (i) if there are millions of

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

observations in the target population, constructing a matrix with that many rows would be infeasible, and (ii) the representation of each unit based on its similarly to other units in the sample is most relevant to how we reweight members of the sample; if there were members of the population that are very different from members of the sample, then no weighting of the sample will account for this.

Kernel bandwidth. The choice of b in the kernel definition scales the similarity measure and is thus effectively a feature extraction choice, constrained by feasibility. Too small a choice of b makes each observation appear 'too unique,' pushing the kernel distance, $\exp(-\|X_i - X_j\|^2/b)$, to zero for any given pair of units; on the other hand, too large a choice of b makes each observation seem 'too similar', producing a kernel distance approaching one for all pairs. A choice of b is, therefore, desirable when it yields a $\mathbf{K}$ with meaningful variability in the similarity measure among different pairs of units. We use the variance of $\mathbf{K}$ as a measure of the useful information available at a given choice of b and turn to this metric to motivate our choice of bandwidth, selecting the value which produces maximal variance in $\mathbf{K}$. We make no claim as to the optimality of this result, but it offers a reasonable choice that can be established without looking at the result. In our simulations and applications, this choice produces consistently good performance, though the results are shown to be stable across a wide range of b regardless (see Appendix C.5, online supplementary material). Further discussion of the kernel bandwidth, as well as details on data preprocessing and scaling decisions appropriate for categorical, continuous, or mixed variable settings can be found in Appendix B, online supplementary material.

4.5 Practice and diagnostics

We recommend several diagnostics that can be used to better understand the resulting weights and what they achieve or fail to achieve. First, the number of dimensions of $\mathbf{K}$ optimally selected for calibration (r) should be checked. If this is very small (e.g. 1 or 2), the user should be aware that balance improvements were minimal. Second, researchers should compare the weighted sample and target population margins on the original $\mathbf{X}$ and explicitly chosen functions of these variables that may be of concern, such as interactions. We illustrate this below. Third, we suggest two summary statistics to assess the degree to which multivariate balance has been improved. The first is an L_1 measure of the distance between the distribution of X for the survey and the population, summed over the units of the survey. This can be obtained both before and after weights are applied to assess the reduction in multivariate imbalance (Hazlett, 2020). The second is the ratio of the bias bound [equation (7)], calculated with and without the weights, to determine the proportional improvement in the degree of potential bias due to remaining imbalances on $\mathbf{K}$. Both serve to indicate to the user whether substantial improvements in multivariate balance were achieved by the weights.

Finally, it is often valuable to understand how extreme the weights are and thus how heavily the solution depends on a small number of observations. This can be done by the investigator's preferred means, such as inspecting the distribution of weights visually, or constructing statistics such as the effective sample size or the number of observations (working from the most heavily weighted towards the least) that one needs to sum to achieve 90% of the total sum of weights. We present these diagnostics for our application in Appendix C.4, online supplementary material.

5 Illustration and simulation

5.1 An illustration

We provide here an example to illustrate how easily bias can emerge in a simple case with just two variables and with mean balance holding perfectly by construction. Suppose a target population of interest consists of four groups in equal shares: college-educated females, college-educated nonfemales, noncollege-educated females, and noncollege-educated nonfemales. A given policy happens to be supported by 80% of college-educated females and only 20% of those in the other three groups. Thus, the mean level of support in the target population would be $\frac{1}{4}(0.8) + \frac{3}{4}(0.2) = 35\%$. Further, suppose the sample is designed to carefully quota on gender and education, obtaining 50% female and 50% college-educated respondents. We use quota sampling in this example for simplicity as it allows us to have a sample already matched to the target

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Table 1. Illustration of sample weighting and ideal weights

Characteristics		Proportions		Outcome	Weights (times N_s)		
Female	College	Target population	Sample	Pr(support)	Unweighted	Mean cal.	Ideal
1	1	1/4	3/8	0.80	1	1	2/3
1	0	1/4	1/8	0.20	1	1	2
0	0	1/4	3/8	0.20	1	1	2/3
0	1	1/4	1/8	0.20	1	1	2
		Target population mean:		0.35			
		Weighted mean:			0.425	0.425	0.35

Note: Quota sampling ensured the sample was representative on the means of college and female. Mean calibration weights will thus be uniform. College-educated female respondents are overrepresented in the sample, however, as are noncollege-educated nonfemales. Because the outcome also varies based on this interaction, these mean calibration weights fail to balance on all important strata of X, producing bias. The 'ideal' weights represent the choice that would bring the sample proportion of each stratum to match that in the target population.

population on the margins. The same considerations would apply, however, in a convenience sample or more generally if weighting were required to achieve mean calibration.

While this sampling procedure seems reasonable and includes the right variables in principle, it neglects intersectional strata. Suppose that, among females, the sample drew a higher proportion of college-educated respondents (three-quarters, as opposed to half in the target population). Conversely among nonfemales, suppose that fewer respondents were college-educated (one-quarter, instead of half). The average level of support for this policy in the unweighted sample would then be $\frac{3}{8}(0.8) + \frac{5}{8}(.2) = 42.5\%$, rather than the 35% in the target population. In other words, a key interaction term (female × college), or the indicator for being in that intersectional stratum, has a different mean in the sample and the target population, and it influences the outcome. It is thus an example of an omitted variable, Z, that drives an association between ϵ and η, violating Assumption 2 unless Z was included.

Table 1 summarizes this situation, and describes a set of 'ideal weights' that would correct the proportions of each intersectional stratum, thus producing the correct answer. The weights would downweight the strata that the sample overemphasized and upweight the strata that the sample underemphasizes. The correction using these weights can be verified by multiplying the sample proportions ($\frac{3}{8}$ or $\frac{1}{8}$) by the proposed ideal weights ($\frac{2}{3}$ or 2, respectively), always producing $\frac{1}{4}$ as the effective postweighting sample proportion. Note that in this simple setting with just two binary variables, this would be feasibly and perfectly achieved by using poststratification. However, this is an artefact of working with a simple illustration; the challenge will be in more complex examples, where poststratification is infeasible, as shown in our next example and application.

Table 2 is similar, but reveals what *kpop* does to achieve the same weights. The middle portion of the table shows the first four columns of the kernel matrix, **K**, corresponding to the same four types of observations. For any two units i and j with the same values on the covariates, $k(X_i, X_j) = e^0 = 1$. Thus, the diagonal of the kernel matrix will always be one. Choosing the exponential denominator b conveniently as 1 for illustrative purposes, individuals that differ on one trait but not the other will have $k(X_i, X_j) = e^{-((1-0)^2+(0-0)^2)} = e^{-1} \approx 0.37$. Individuals who differ on both characteristics will have $k(X_i, X_j) = e^{-((1-0)^2+(1-0)^2)} = e^{-2} \approx 0.14$. All values in **K** will be one of these three values in this simple example.

Weights must then be found that will multiply each row of the data, including the four shown here. In short, because there are 'too many' female college graduates relative to the target group, the average 'similarity' of observations in the sample to the female college graduate group will be too high. Specifically, the sample mean of column one ($k(, 1)$) will be 0.52, whereas in the target group, it would have been only 0.47. Likewise, the average similarity of observations to the nonfemale noncollege graduate group will be too high, as they too are overrepresented. On the other hand, the average similarities to the other two groups will be too low because they are underrepresented in the sample drawn. *kpop* exploits the idea that the weights that would reproduce the right

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Table 2. Illustration: kernel balancing to adjust sample to target

Characteristics (X)			Kernel matrix (K)						Outcome	Weights
Female	College	Sample %		$k(,1)$	$k(,2)$	$k(,3)$	$k(,4)$	(Repeats)	Pr(support)	(kpop)
1	1	3/8	$k(1,)$	1	0.37	0.14	0.37	…	0.80	2/3
1	0	1/8	$k(2,)$	0.37	1	0.37	0.14	…	0.20	2
0	0	3/8	$k(3,)$	0.14	0.37	1	0.37	…	0.20	2/3
0	1	1/8	$k(4,)$	0.37	0.14	0.37	1	…	0.20	2
			⋮	⋮	⋮	⋮	⋮	⋮		
Target mean:				0.47	0.47	0.47	0.47		0.35	
Sample mean:				0.52	0.42	0.52	0.42		0.425	
kpop-weighted mean:				0.47	0.47	0.47	0.47		0.35	

Note. Kernel matrix representing each of four unique types of individuals in the sample. Each element $k(X_i, X_j)$ is equal to $\exp(-\|X_i - X_j\|^2/1)$, where the numerator in the exponent will be equal to two times the number of features on which i and j differ and the denominator b is chosen as 1 for convenience. The columns provide new bases for representing the data.

proportion of each type of observation would also reproduce the average similarity to each of these types that we see in the target group. The explicit process of choosing these weights involves an optimization step, simply minimizing the difference between the weighted mean of each column of **K** in the sample and the (unweighted) column averages of **K** in the target group.

The weights that achieve this are shown in the final column. The final row in Table 2 verifies that these weights achieve the desired weighted average similarities (columns of **K**). For example, after weighting, the average similarity of observation to the female college graduate type (first column of **K**) will be $\frac{3}{8}(\frac{2}{3})(1) + \frac{1}{8}(2)(.37) + \frac{3}{8}(\frac{2}{3})(.14) + \frac{1}{8}(2)(.37) = 0.47$, as it is in the target group. The same holds true for the average similarity of the weighted sample to each of the other three intersectional strata. As expected, the solution down-weights units in the oversampled groups (female × college and nonfemale × noncollege) and upweights those in the remaining two, undersampled groups. Further, these weights are numerically equal to the 'ideal' weights in Table 1, and, when multiplied by the sample proportions of each group, all produce the intended value of $\frac{1}{4}$, matching the target group proportions.

Consequently, with only the matrix **X** for the sample and target population and no further information about the importance of the interaction, the *kpop* weighted estimate of the mean level of support matches that in the target population (35%).

5.2 Realistic simulation setting

The above example was made as simple as possible for purposes of clarifying the method. In practice, however, methods such as poststratification would have also worked in that setting. We next consider a more complicated setting to demonstrate, first, the approach in a context where other methods will encounter difficulties, and second, to more fully illustrate performance on both bias and variability. We design our semisynthetic simulation to closely match the application below, while allowing us to specify models governing the selection and outcome processes. This type of calibrated simulation helps to assess the performance of *kpop* in a realistic setting, with a data structure, selection process, and outcome more similar to those encountered by applied researchers (e.g. Dorie et al., 2019). We construct the selection and outcome models based on observed relationships in the original data, so that the mean in the target population is known (see Hill, 2011; Kern et al., 2016 for related examples). We emphasize that while this allows us to assess the performance of *kpop* under known misspecification, we do not consider a range of conditions for the data structure, size, or severity of the selection process.

We first construct a sample model through which respondents in the smaller sample are drawn from the larger target population. In our case, the target population is given by the postelection

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

wave of the 2016 Congressional Cooperative Election Study (CCES) (Ansolabehere & Schaffner, 2017). We specify our selection model $p(S = 1) = logit^{-1}(\mathbf{X}\beta)$ and construct new samples by taking Bernoulli draws from the CCES population according to this model. Our outcome model is linear in the same set of covariates, $p(Vote = Dem) = \mathbf{X}\gamma$, allowing us to directly control the mechanism of bias by specifying the correlation of β and γ. In the following simulation, this correlation is about −0.75, producing negative bias of about −3.5 p.p. in the unweighted sample.

Both models are (link) linear in the same, fairly simplistic set of covariates **X**: party identification, age (four-way), gender, educational attainment (three-way), race/ethnicity (four-way), born-again Christian status, and a subset of two-way interactions between party identification and age as well as born-again status and age. Coefficients in the selection model are chosen to produce roughly realistic samples comparable to the observed Pew survey and scaled to produce a sample size of roughly 500. For the outcome model, coefficients are adjusted through an automated procedure to produce values in probability scale. Below, we present results across 1,000 iterations. Further details can be found in Appendix D, online supplementary material.

We compare several *kpop* and *kpop + mf* specifications to a range of approaches that we anticipate thoughtful investigators might attempt, as well as three approaches that exploit the true specification or selection probability for comparison. These include raking on just the basic demographic variables [*mean calibration (demos)*], on demographics and education [*mean calibration (demos + edu)*], and on all available variables [*mean calibration (all)*]. See Appendix C.1, online supplementary material for additional information on these variables. We motivate these specifications and why investigators might choose them in the context of the application (Section 6).

Finally, for benchmarking purposes, we compare *kpop* as it would be implemented (without access to true information about the model or variables to include) with three methods that do have access to this information: (i) poststratification on the correct (minimal) set of intersectional variables described [*poststratification (true)*]; (ii) the Horvitz–Thompson estimator employing the true (unknown) sampling probability [*Horvitz–Thompson (true)*]; and (iii) mean calibration on just the variables required for linear ignorability [*mean calibration (true)*].

Table 3 shows the resulting estimates in terms of bias, mean squared error, and absolute bias reduction. We find that the four *kpop* specifications all significantly reduce the bias (by 92%–99.9%) and have MSEs roughly an order of magnitude smaller than those of *mean calibration (demos)* and *mean calibration (demos + edu)*, both of which reduce bias only by about half. By contrast *mean calibration (all)* happens to perform very well, highlighting the specification sensitivity of this approach as compared to the stability of *kpop*. Further, *kpop* performs well even compared to estimators that are given access to the true model or set of variables to include. Even with our fairly simplistic selection model, *Poststratification (true)* still struggles with the 'empty cell' problem, dropping on average 23% of population units and reducing bias only by 68%. Notably, the Horvitz–Thompson estimator is roughly unbiased as expected, but has an MSE almost as large as the unweighted estimator and over 30 times larger than any *kpop* estimator. *Mean calibration (true)* performs well here, but is still comparable to the more naive *kpop* estimators.

6 Application: 2016 US Presidential election

In the 2016 US Presidential election, state-level polls in key states were severely biased, with polling aggregators making overconfident predictions that Donald Trump would lose. National polls correctly predicted that Hillary Clinton would lead the national popular vote, while many overstated the margin. The challenges of correctly weighting a highly nonrandom sample to match the national electorate likely contributed to these errors. As Kennedy et al. (2018) note, existing polls were especially likely to overrepresent college-educated whites.

We test whether weighting with *kpop* would have improved on this, absent foreknowledge of what functions of covariates and intersectional strata are essential to address sources of bias. Because voters may have changed their mind between a given preelection survey and the day of their vote, simply checking whether weighting the outcome of a preelection survey produces an estimate close to the true election result would not provide a meaningful test of weighting techniques. We instead estimate what the average 'retrospective vote choice', measured postelection, would have been among voters in the 2016 election. This involves (1) training a model that predicts stated retrospective vote choice as a function of X using a large postelection survey which we

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Table 3. Simulation results

	Bias (p.p.)	MSE	Abs bias reduction (%)
Unweighted	−3.510	12.603	0.000
Mean calibration (demos)	−1.618	2.893	0.539
Mean calibration (demos + edu)	−1.296	1.961	0.631
Mean calibration (all)	−0.029	0.226	0.992
kpop	−0.272	0.357	0.923
kpop + mf (demos)	−0.165	0.297	0.953
kpop + mf (demos + edu)	−0.150	0.268	0.957
kpop + mf (all)	0.012	0.244	0.997
Horvitz–Thompson (true)	−0.160	10.229	0.954
Poststratification (true)	−1.120	1.571	0.681
Mean calibration (true)	−0.010	0.214	0.997

Note. Bias, mean squared error, and absolute bias reduction by weighting method across 1,000 simulations wherein the outcome and selection model are specified using the same set of variables to directly control the mechanism of bias. The models above the line represent specifications that investigators might realistically attempt without access to the unknowable, true selection model. For comparison, those below the line demonstrate the performance of estimators given 'true' information about the correct selection model that would be unknown to investigators.

define as the target population; (2) applying this model to predict the 'retrospective vote choice' of each individual in a preelection survey using their covariates X; (3) constructing weights to calibrate the preelection sample to the target population; then (4) comparing the weighted average of the predicted 'retrospective vote choice' in the preelection sample to the stated vote choice the target population. We emphasize that, were it not for the dynamic nature of vote intention in preelection surveys, using this modelling outcome would not be necessary in settings where we can directly estimate the outcome, Y, of interest.

6.1 Data and details

Survey sample. For the respondent sample, we use survey data from the final poll conducted by the Pew Research Center before the general election in November 2016. Pew is a high-quality, nonpartisan public opinion firm. The survey was conducted from 20–25 October 2016 using telephone interviews among a national sample of 2,583 adults. On landline phone numbers, the interviewer asked for the youngest adult currently home (647), and cell phone interviews (1,936) were done with the adult who answered the phone (Pew Research Center, 2016). Random-digit dialing was used, combined with a self-reported voter registration screen. We keep only the $N_s = 2,052$ respondents who report that they plan to vote or have already voted. The publicly available data do not include survey design weights, and we use $q_i = 1$ for all respondents, although researchers could let q be defined using design weights or previous calibration weights. The survey includes proprietary multistage calibration weights, where the first-stage accounts for differential sampling probabilities due to the random-digit-dialing procedure, and the second-stage conducts a calibration procedure to match the US population on many of the same auxiliary variables as we include. We do not include these weights as our base weights because we are weighting to a different target population, namely one defined by verified voters from a national survey, the CCES (discussed below), and under Assumption (2), our estimator is unbiased even starting from equal weights. Finally, we code vote choice as being for 'Republican Donald Trump', 'Democrat Hillary Clinton', or 'Other/Don't Know', and we include voters who 'lean' towards one of the two major party candidates.

Defining the target population. Ideally we would define the target population using verified voter records from the Secretaries of State. However, we do not have access to such an administrative file. Instead, following Caughey et al. (2020), we define our target population using the common content from the postelection wave of the 2016 Congressional Cooperative Election Study (CCES)

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

(Ansolabehere & Schaffner, 2017). The CCES is a large survey that aims to be representative of all voters, and the survey weights for the postelection wave lead to an estimate of the popular vote margin between the two major parties (2.48 percentage points, D-R) that is very close to the truth (2.3 percentage points). Second, the CCES includes a number of demographic survey questions that overlap with those asked in the Pew study which we can use for calibration. We incorporate the weights provided by the CCES (*commonweight_vv_post*) into the definition of our target population. Limiting the data to voters for whom the outcome variable was not missing, and who stated that they 'definitely voted' leaves a total of $N_{pop} = 44{,}932$ units.

Our auxiliary data, $\mathbf{X}$, are defined using all of the overlapping variables in our data sets: age, reported gender, race/ethnicity, geographic region, educational attainment, party identification, income, born-again Christian status, church attendance, and religion. All variables are self-reported except for region. Appendix Table C.1, online supplementary material summarizes the distributions of these variables and how they differ in the target population (CCES) compared to the survey sample (Pew). For example, those with higher levels of educational attainment and higher income are overrepresented in the Pew sample, as are older voters and Independents. By contrast Black voters and women are underrepresented in the sample relative to the target population.

Modelled outcome. As noted above, the outcome variable to be weighted in this example is itself a modelled quantity representing the difference in probability of voting Democratic vs. Republican given one's covariates ($p(D_i - R_i \mid X_i)$). We use a multinomial logit model (see e.g. Long, 1997) to estimate the relationship between $\mathbf{X}$ and three-way 'retrospective vote choice' (Republican, Democrat, and Other) measured by asking respondents who they voted for in the postelection CCES survey. We use regularization in doing so, to mitigate overfitting concerns (see e.g. Hastie et al., 2009). This model includes gender, three-way party identification, race/ethnicity, six-way education, region, six-way income, five-way religion, four-way church attendance, born-again status, continuous age, age^2, gender × party identification, and age (continuous) × party identification.

Recall that our goal is to find weights for the Pew observations (sample) such that the weighted average value of predicted $p(D_i - R_i \mid X_i)$ matches that in the CCES data, here 2.48 percentage points. None of the subsequent weighting methods are aware this particular choice of $\phi(X)$ has been made. Using this specification, the outcome can be modelled quite effectively. For example, choosing the highest-probability outcome as an individual's final vote choice leads to an 85%–86% correct classification rate for nonindependents. This fitted postelection outcome model is then used to predict $p(D_i - R_i \mid X_i)$ using the $\mathbf{X}$ data from the Pew preelection survey. Additional details can be found in Appendix C.2, online supplementary material.

Weighting methods. We compare *kpop* estimators to the two common methods researchers use for constructing survey weights discussed above, mean calibration and poststratification. For mean calibration, we consider four specifications that represent a range of choices thoughtful researchers might attempt: (i) basic demographic variables only *mean calibration (demos)*, including: age (four-way), gender, race/ethnicity, geographic region, and party identification; (ii) those variables plus educational attainment [*mean calibration, (demos + edu)*]; all available data [*mean calibration, (all)*], adding income, religion, born-again Christian status, and church attendance (see Table C.1, online supplementary material). Finally, we include one model that is given retrospective benefit, based on the analysis of Kennedy et al. (2018): *mean calibration (retrospective)*, which includes the interaction of party identification with age (four-way) and, separately, with gender. It additionally addresses the importance of low-education white voters in the 2016 election, particularly in Midwestern states, by also including party identification × region × race/ethnicity for all voters. Among white respondents, this is expanded to also interact with educational attainment (six-way). We include this model to evaluate the question of whether our proposed *kpop* method can perform as well as a retrospective-informed model that serves as a best-case for what expert knowledge could hope to achieve.

Turning to poststratification, using all 10 available variables results in highly complex cross-sectional strata which, in turn, make missing cells a significant hurdle when reweighting the survey sample, with nearly 92% of population units representing strata not present in the sample. To bring the number of empty cells to a reasonable level, we coarsen age and income into three- and four-category variables, respectively, and do not include religion or church attendance.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

In order to give poststratification the best advantage possible, we omit full six-way education and instead stratify on a coarsened version of the above-mentioned important, retrospectively informed interaction among race/ethnicity and education, namely, coarsened three-way educational attainment among white with no stratification by education among nonwhite voters. The resulting estimator stratifies on gender, race/ethnicity, region, party identification, born-again status, three-way income, three-way age, and the three-way educational attainment and white interaction. Unfortunately, this still results in dropping around 30% of units due to empty strata.

We compare the models described above against *kpop*, applying our proposed kernel balancing method with all the available categorical variables available as described in Table C. 1, online supplementary material. For comparison with the preceding raking specifications, we also include three models, *kpop + mf (demos), kpop + mf (demos + edu)*, and *kpop + mf (all)* that first conduct mean calibration before proceeding to balance on the kernel matrix, as discussed in Section 4.3.

6.2 Results

Balance. We first consider the balance achieved by each method on the observed covariates. Table 4 presents the absolute error, weighted by the target population proportion for each level, for each auxiliary variable (rows 1–10) and a set of interactions (11–17). By construction, the mean calibration methods, as well as the *kpop + mf* methods, perfectly match the marginal distributions for any variables that are included in the model.

All methods greatly improve representativeness in the respondent sample as indicated by the reduction in error across variables and interactions relative to the unweighted sample. As expected, *kpop* (without 'mean first') achieves good but imperfect balance on the included covariates and interactions, despite not being directly constrained to achieve balance on them. Though we should expect poststratification to produce perfect balance on the included terms, we see that, even with significant variable coarsening, empty cells pose a significant problem. As a result, poststratification fails to get the correct margins, much less produce the nonparametric adjustment one hopes for under Assumption 1.

In the lower rows of Table 4, we investigate the balance on important interactions, including one that Kennedy et al. (2018) deemed important: midwest × educational attainment × race/ethnicity (bottom row). Without explicitly incorporating knowledge about the importance of these variables, *kpop* significantly improves balance on this interaction, reducing the mean absolute error from the initial value of 1.68 down to 0.1 percentage point—a much greater reduction than any of the non-*kpop* estimators. When incorporating the mean first requirements *kpop + mf* also effectively addresses this interaction, reducing absolute bias to between 0.04 and 0.07 percentage points. We see similar patterns of improvements in the performance of the *kpop* methods across a number of important interactions. Notably, in each case, the *kpop + mf* method outperforms its mean calibration counterpart, emphasizing the 'no worse' nature of the mean first approach. Additionally, balance is achieved regardless of the specified mean first constraints, highlighting the robustness of *kpop* to standard user-specified constraints.

Weight severity. The additional constraints solved by *kpop* weights can lead to reduced effective sample sizes compared to other approaches. To calculate the effective sample size, we use the Kish formulation of $\frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}$. Here, *poststratification* and *mean calibration (all)* have effective sample sizes of 983 and 1,101, respectively, while *kpop* and *kpop + mf (all)* have effective sample sizes of 789 and 749, respectively. Similarly, the (minimum) number of observations required to arrive at 90% of the total weight is 1,235 and 1,362 for *poststratification* and *mean calibration (all)*, respectively, but 1,193 and 1,223 for *kpop* and *kpop + mf (all)*. Thus, a price is paid for the *kpop*'s ability to balance on more general functions of **X**, but it is a fairly modest one here.

Weighting diagnostics. Improvements in balance on **K** can be assessed using the diagnostics described above in section 4.5. The L_1 gap between the kernel-based estimates of multivariate density fell from 0.0318 prior to weighting to 0.0011 or below under all *kpop* estimates, roughly a 29-fold improvement. Similarly, the bias bound showed fivefold to sixfold improvements under each set of weights as compared to the unweighted bias bound. Additional diagnostic and descriptive results can be found in Appendix C.4, online supplementary material.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Table 4. Weighted mean absolute error on auxiliary variables (percentage points)

	Pew orig	kpop	kpop + mf (demos)	Mean calib (demos)	kpop + mf (d + edu)	mean calib (d + edu)	kpop + mf (all)	mean calib (all)	mean calib (retro)	poststrat (reduc)
Female	3.65	0.11	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)
Pid (three-way)	2.53	0.28	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.45)
Age (four-way)	4.85	0.22	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	1.95
Race/ethnicity (four-way)	1.54	0.16	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	1.25	(3.50)
Region (four-way)	1.50	0.02	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	0.58	(2.51)
Educ (six-way)	8.64	0.32	0.12	8.68	(0.00)	(0.00)	(0.00)	(0.00)	1.67	2.09
Income (six-way)	4.35	0.35	0.21	4.04	0.15	3.35	(0.00)	(0.00)	3.12	1.78
Born-again (bin)	1.72	0.14	0.01	2.95	0.10	0.18	(0.00)	(0.00)	0.62	(4.95)
Religion (five-way)	6.42	0.48	0.34	5.24	0.19	6.66	(0.00)	(0.00)	7.77	6.17
Church attnd. (four-way)	12.88	0.44	0.17	12.13	0.16	11.92	(0.00)	(0.00)	12.61	12.11
Pid × race/ethnicity	0.71	0.50	0.34	0.86	0.15	0.68	0.63	1.15	0.42	1.20
Educ × pid	6.22	0.36	0.21	6.25	0.29	0.45	0.39	0.57	1.56	1.49
Educ × pid × race/ethnicity	3.81	0.45	0.48	3.87	0.34	0.44	0.67	0.58	0.46	0.90
Race/ethnicity × educ × reg	2.91	0.24	0.44	2.97	0.27	0.48	0.70	0.49	0.32	0.87
Educ × white	10.80	0.28	0.51	10.64	0.21	0.25	0.09	0.42	1.00	(2.69)
Midwest × white × educ	1.54	0.44	0.52	0.96	0.38	0.47	0.36	0.27	0.35	0.56
Midwest × edu × race/ethnicity	1.68	0.10	0.07	0.69	0.04	0.06	0.04	0.05	0.88	1.01

Note. Absolute error in the distribution of categorical variables, weighted by the target population proportion for each level. Numbers in parentheses indicate the variable was included as a calibration constraint, and so imbalances near zero are expected. Note that all interactions with educational attainment use a three-way education coding.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

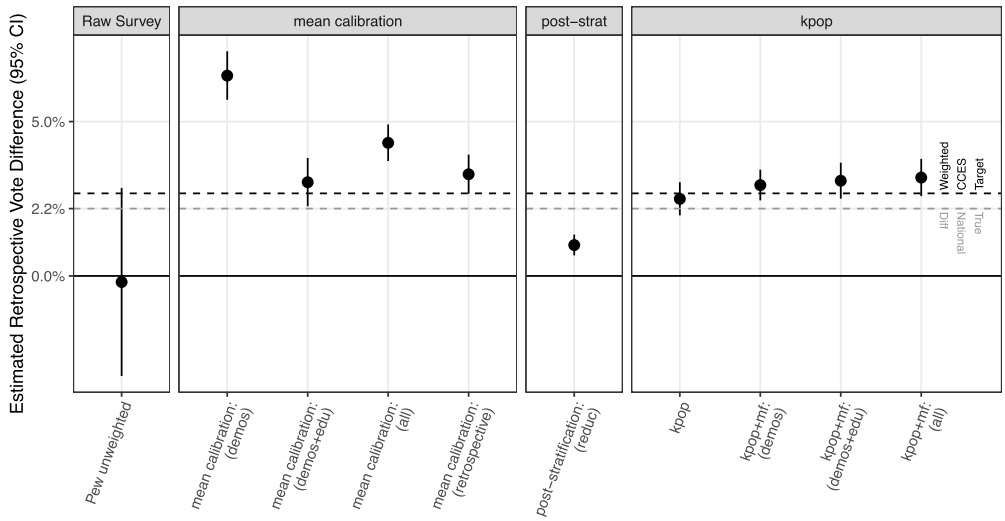

Figure 1. Comparison of approaches for weighting Pew survey data weighted to Congressional Cooperative Election Study (CCES) target. *Note:* Comparison of weighting methods on Pew survey data to weighted CCES target. Points represent the estimated vote margin from the predicted 'retrospective vote choice' using Pew survey data and corresponding weighting scheme. The dashed black line indicates the target, the reported two-way vote margin in the weighted CCES. The dashed grey line indicates the true values from national vote returns.

Estimates. Results are shown in Figure 1. Recall that our target is a two-way vote difference on reported vote choice (D-R) of 2.48 percentage points and that, for each weighting model, we evaluate the average of the predicted 'retrospective vote choice.' The Pew survey, without weights, is extremely nonrepresentative, with an unweighted average Clinton two-party vote margin of −0.194 [−3.24, 2.85] percentage points (95% confidence interval in brackets). Mean calibration on basic demographics, excluding educational attainment, flips the signs of the estimate, producing an estimated margin to 6.49 [5.71, 7.28] percentage points. Mean calibration including educational attainment performs well, with an estimate of 3.04 [2.26, 3.82]. This is consistent with the findings of Kennedy et al. (2018) that educational attainment was a significant driver of both nonresponse and voting for Donald Trump, especially among white voters. Moving to mean calibration on all auxiliary variables, the estimate moves farther from the truth to 4.31 [3.72, 4.91]. Finally, mean calibration on the retrospectively informed choice of variables and the potentially important interaction of region and education among white voters generated an improved point estimate of 3.30 [2.66, 3.93] substantially closer to the truth.

The *kpop* estimates are both stable and close to the truth across different specifications. Using *kpop* alone results in a weighted estimate of 2.50 [1.96, 3.04] percentage points, the closest to the truth of all nine methods tested. *kpop* + *mf* with additional mean first calibration produces point estimates of 2.95 [2.45, 3.44] (*demos*), 3.09 [2.51, 3.67] *(with education)*, and 3.19 [2.59, 3.79] percentage points for (*all*), all close to the target margin in the CCES. Appendix Table C.3, online supplementary material summarizes these results across all weighting methods.

We note that the standard errors (and thus confidence intervals) for *kpop* are not necessarily larger than, and in fact are often smaller than, those of other methods. On the one hand, *kpop* may lead to more variable weights than other approaches, which can contribute to larger variance estimates. Simultaneously, however, the linearization/residualization standard errors we employ (Fuller, 1975; Kott, 2016) remove from the outcome any signal that can be explained linearly by the $\phi(X)$ that were calibrated upon. Thus, when a significant component of the outcome variance can be explained by $\phi(X)$, this leads to a substantial reduction in the estimated standard error. This reduction may be more than sufficient to make up for potentially higher variance weights, resulting in overall shorter intervals. As demonstrated in Appendix D.3.3, online supplementary material, these estimated standard errors have nominal coverage under simulation and closely reproduce the empirical standard deviation of estimates under resampling.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

7 Conclusion

The challenges we seek to manage regarding common survey weighting techniques, particularly poststratification and mean calibration, are well known in the survey weighting literature (e.g. see Berinsky, 2006; Hartman & Levin, 2019; Kalton & Flores-Cervantes, 2003). Recent methods aim to address trade-offs between these approaches, as well as the relationship between mean calibration and inverse propensity score weighting (Ben-Michael et al., 2021; Linzer, 2011). Variable selection for weighting has addressed one aspect of feature selection (Chen et al., 2019; McConville et al., 2017). Here, we describe an approach that helps to reduce user discretion in the related problem of deciding what features and functions of observed covariates must be made to look similar in the sample and target population.

We note several limitations and areas for future work. First, the implementation described here makes no use of outcome data when constructing weights. This allows users to choose weights blind to the outcome to protect themselves against data 'snooping'. Further, these weights would be appropriate for estimating any outcome, as several are often of interest in a given survey. The downside, however, is that this leaves possible gains in efficiency and bias reduction on the table if there are functions of X that predict response (and are thus imbalanced) but that do not predict the outcome and so need not be calibrated to achieve linear ignorability. Still, recognizing such variables and choosing not to calibrate on them could lead to improved calibration on the remaining variables, possibly resulting in less extreme weights. Such an approach remains an option worth exploring in future work.

While we suggest the use of the Gaussian kernel with a variance maximizing choice of b, optimal selection of b remains an area of ongoing research. Fortunately, our empirical results are not sensitive to the choice of b (see Appendix C.5, online supplementary material), but this may not always be the case. Further, while the present example focussed on discrete X for comparability to other approaches, a benefit of our approach, and the Gaussian kernel, is that it applies well—and perhaps more naturally—with continuous X. Nevertheless, other choices of kernels, and a means of choosing among them, is a fruitful area of future work.

Next, in our implementation, we use only the sample observations to form the columns of $\mathbf{K}$, a decision driven by feasibility constraints. Since the number of units in the target population is typically very large, using all sample and population units to form the columns of $\mathbf{K}$ is typically infeasible. Future work could consider ways of augmenting the columns of $\mathbf{K}$ by selecting population units that are poorly represented by sample units and using these additionally in the formation of the bases for calibration. Finally, our method relies on individual-level population data, either from administrative data or a high-quality representative survey, which may not be accessible to all researchers.

In summary, *kpop* is a kernel-based approach for weighting samples to be representative of target populations, while reducing reliance on user discretion and domain expertise to determine what covariates—and functions of those covariates—should be used for calibration. It does so by estimating a flexible, nonlinear set of basis functions through a kernel transformation and achieving approximate balance on this representation of the covariates. As shown in our application to the 2016 US Presidential election, this method has great promise for reducing bias in nonrepresentative samples.

Acknowledgments

The authors thank Avi Feller, Kosuke Imai, Luke Miratrix, Santiago Olivella, Alex Tarr, Baobao Zhang, and participants in the UCLA Causal Inference Reading Group. The `kbal` package for the `R` computing language implements this method and is freely available.

Conflicts of interest: None declared.

Funding

This work is partially supported by a grant from Facebook Statistics for Improving Insights and Decisions and a University of California, Los Angeles Transdisciplinary Seed Grant.

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Data availability

Data derived from sources in the public domain. Replication materials, including links to source data, can be found at https://github.com/csterbenz1/kpop_rep.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Ansolabehere S., & Schaffner B. F. (2017). CCES Common Content, 2016. https://doi.org//10.7910/DVN/GDF6Z0

Ben-Michael E., Feller A., & Hartman E. (2021). Multilevel calibration weighting for survey data. *Political Analysis*, *32*(1), 65–83. https://doi.org/10.1017/pan.2023.9

Berinsky A. J. (2006). American public opinion in the 1930s and 1940s: The analysis of quota-controlled sample survey data. *International Journal of Public Opinion Quarterly*, *70*(4), 499–529. https://doi.org/10.1093/poq/nfl021

Caughey D., Berinsky A. J., Chatfield S., Hartman E., Schickler E., & Sekhon J. S. (2020). *Target estimation and adjustment weighting for survey nonresponse and sampling bias*. Cambridge University Press.

Chen J. K. T., Valliant R. L., & Elliott M. R. (2019). Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), 657–681. https://doi.org/10.1111/rssc.12327

Deville J.-C., & Särndal C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. https://doi.org/10.1080/01621459.1992.10475217

Dorie V., Hill J., Shalit U., Scott M., & Cervone D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Sciences*, *34*(1), 43–68. https://doi.org/10.1214/18-STS667

Fuller W. A. (1975). Regression analysis for sample survey. *Sankhya*, *37*(3), 117–132.

Hainmueller J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Hainmueller J., & Hazlett C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, *22*(2), 143–168. https://doi.org/10.1093/pan/mpt019

Hartman E., & Huang M. (2023). Sensitivity analysis for survey weights. *Political Analysis*, *32*(1), 1–16. https://doi.org/10.1017/pan.2023.12

Hartman E., & Levin I. (2019). Accounting for complex survey designs: Strategies for post-stratification and weighting of internet surveys. In E. Suhay, B. Grofman, & A. H. Trechsel (Eds.), *Oxford handbook of electoral persuasion*. Oxford University Press.

Hastie T., Tibshirani R., Friedman J. H., & Friedman J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. *2*). Springer.

Hazlett C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*, *30*, 1155–1189. https://doi.org/10.5705/ss.202017.0555

Hill J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Kallus N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, *21*(62), 1–54.

Kalton G., & Flores-Cervantes I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81–97.

Kennedy C., Blumenthal M., Clement S., Clinton J. D., Durand C., Franklin C., McGeeney K., Miringoff L., Olson K., Rivers D., Saad L., Witt G. E., & Wlezien C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, *82*(1), 1–33. https://doi.org/10.1093/poq/nfx047

Kennedy C., & Hartig H. (2019). *Response rates in telephone surveys have resumed their decline*. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.

Kern H. L., Stuart E. A., Hill J., & Green D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127. https://doi.org/10.1080/19345747.2015.1060282

Kott P. S. (2016). Calibration weighting in survey sampling. *WIREs Computational Statistics*, *8*(1), 39–53. https://doi.org/10.1002/wics.2016.8.issue-1

Kott, P. S., & Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, *105*(491), 1265–1275. https://doi.org/10.1198/jasa.2010.tm09016

Linzer D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis*, *19*(2), 173–187. https://doi.org/10.1093/pan/mpr006

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026

Little R. J., & Rubin D. B. (2019). *Statistical analysis with missing data* (Vol. *793*). John Wiley & Sons.
Long J. (1997). Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques in the Social Sciences*, *7*.
McConville K. S., Breidt F. J., Lee T. C., & Moisen G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, *5*(2), 131–158. https://doi.org/10.1093/jssam/smw041
Mercer A. W., Kreuter F., Keeter S., & Stuart E. A. (2017). Theory and practice in nonprobability surveys. *Public Opinion Quarterly*, *81*(S1), 250–271. https://doi.org/10.1093/poq/nfw060
Micchelli C. A., Xu Y., & Zhang H. (2006). Universal kernels. *Journal of Machine Learning Research*, *7*(12), 2651–2667.
Opsomer J. D., & Erciulescu A. L. (2021). Replication variance estimation after sample-based calibration. *Survey Methodology*, *47*(2), 265–278.
Pew Research Center. (2016). As election nears, voters divided over democracy and 'respect'. (Technical Report). Pew Research Center, October 27.
Särndal C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, *33*(2), 99–119.
Särndal C.-E., & Lundström S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.
Wong R. K., & Chan K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, *105*(1), 199–213. https://doi.org/10.1093/biomet/asx069
Wu C., & Lu W. W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, *84*(1), 79–98. https://doi.org/10.1111/insr.v84.1
Yeying Z., Savage J. S., & Ghosh D. (2018). A kernel-based metric for balance assessment. *Journal of Causal Inference*, *6*(2). https://doi.org/10.1515/jci-2016-0029
Zhao Q., & Percival D. (2016). Entropy balancing is doubly robust. *Journal of Causal Inference*, *5*(1). https://doi.org/10.1515/jci-2016-0010

Downloaded from https://academic.oup.com/jrsssa/article/188/3/875/7747430 by University of California, Los Angeles user on 06 January 2026