Journal of Mathematics and Statistics Studies
ISSN: 2709-4200
DOI: 10.32996/jmss
Journal Homepage: www.al-kindipublisher.com/index.php/jmss

JMSS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Nonparametric Density Estimation in Survey Sampling

Dr Reuben Cheruiyot Lang'at, PhD
University of Kabianga P.O. Box 2030-20200 Kericho
Corresponding Author: Dr Reuben Cheruiyot Lang'at, PhD, **E-mail**: rlangat@kabianga.ac.ke

| ABSTRACT
Nonparametric methods for estimating probability densities are popular because they provide flexible tools for exploratory analysis, model checking, and inference when little is known about the underlying distributional form. In the context of sample surveys where data arise from complex designs involving stratification, clustering, and unequal inclusion probabilities, naive application of standard nonparametric estimators can, however, produce biased and inconsistent results. This paper reviews foundations of nonparametric density estimation and use of kernel and local polynomial methods and discusses their adaptation to design-based and model-based survey frameworks. Practical implementation issues involving bandwidth selection, boundary correction, and computational considerations are made. Throughout, emphasis is placed on methods that respect survey design information, and on trade-offs between design-based validity and model-based efficiency. The paper concludes with recommendations for practice and directions for future research.

| KEYWORDS

Nonparametric density estimation, kernel density estimator, survey sampling, design-based inference, weighted estimators, variance estimation, bandwidth selection

| ARTICLE INFORMATION

ACCEPTED: 01 February 2026 **PUBLISHED:** 16 February 2026 **DOI:** 10.32996/jmss.2026.7.2.2

1. Introduction

Density estimation is a fundamental task in statistics: summarising the distribution of a variable, identifying multimodality, tail behaviour, and informing model-building. Classical density estimators, such as the kernel density estimator (KDE) introduced in early form by Rosenblatt and Parzen, have matured into robust, well-understood tools in iid settings (Rosenblatt, 1956; Parzen, 1962; Silverman, 1986). However, survey data differ from independent, identically distributed (*iid*) samples in important ways. Complex survey designs commonly employ stratification, clustering, unequal-probability sampling, and sometimes multistage selection. Moreover, survey data are typically accompanied by sampling weights that is the reciprocals or calibrated versions of inclusion probabilities, constructed to produce unbiased design-based estimates of population totals and means (Horvitz & Thompson, 1952; Särndal, Swensson, & Wretman, 1992; Lohr, 2010).

This discrepancy raises two central questions: firstly, how to adapt nonparametric density estimators so they remain valid under complex sampling designs; and secondly how to quantify their variability accurately, accounting for the survey design. This paper synthesizes the literature and presents practical guidance on density estimation for survey data, emphasizing kernel-based methods while touching on alternative approaches.

2. Classical nonparametric density estimation (iid setting)

Density estimation aims to estimate an unknown probability density f on $\Re^d$ given a sample $X_1, X_2,..., X_n$. The kernel density estimator at a point x∈ $\Re^d$ is defined as:

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad (1)$$

where *K* is a kernel (a bounded integrable function integrating to one) and *h>0* is a bandwidth parameter controlling smoothness (Rosenblatt, 1956; Parzen, 1962; Silverman, 1986; Wand & Jones, 1995). The asymptotic properties of $\hat{f}_h$ for iid samples are well-understood: bias scales like h^2 (for kernels with zero first moment), variance scales like $\frac{1}{nh}$, and optimal mean integrated squared error (MISE) balances these terms to give $h \alpha n^{-1/4}$ in smooth one- or multi-dimensional settings (Scott, 1992; Silverman, 1986).

Bandwidth selection is the critical practical issue in that under-smoothing or using a small *h* yields noisy estimates while over-smoothing or use of a large *h* obscures features. Data-driven selectors include rule-of-thumb plug-in methods, cross-validation, and plug-in estimation of unknown smoothness (Silverman, 1986; Wand & Jones, 1995). Boundary issues require special treatment when support is bounded or restricted. This can be achieved using reflection, boundary kernels, or local polynomial methods.

Alternative nonparametric estimators include orthogonal-series estimators, which expand f in a basis and truncate the expansion, and nearest-neighbour estimators, which adapt bandwidth locally based on distances to neighbours (Devroye & Györfi, 1985; Loader, 1999). While these methods have advantages in certain settings, kernel estimators remain widely used due to simplicity and strong theoretical and practical performance.

3. Survey sampling basics and inferential frameworks

3.1 Complex survey designs and sampling weights

It may be common to have simple random sampling for simple cases, however, survey data might on many occasions require other more complex techniques such as stratification which involves dividing the population into homogeneous strata and applying simple random sampling within each stratum, cluster sampling where primary sampling is done on unit clusters, then subunits within, and unequal-probability sampling where units have different inclusion probabilities, often to oversample rare subpopulations. Associated with each sampled unit *i* is a sampling weight w_i which is often the inverse of the inclusion probability π_i, and sometimes calibrated weights adjusted to known population totals (Särndal et al., 1992; Lohr, 2010).

3.2 Design-based versus model-based inference

Two dominant paradigms exist for survey inference. The design-based perspective treats the finite population as fixed and randomness as arising solely from the sampling design. Estimators are constructed to be unbiased or approximately unbiased under the sampling design; variance estimators must reflect the actual design. The Horvitz–Thompson (HT) estimator for population totals is the conventional example (Horvitz & Thompson, 1952).

The model-based or super-population perspective posits that the finite population arises from a stochastic model; inference conditions on the observed sample but uses the model for estimation and prediction. Model-based approaches can provide efficiency gains but require correct model specification or robust model-assisted techniques (Särndal et al., 1992; Rao, 2000).

For nonparametric density estimation in surveys, both perspectives are relevant. Weighted estimators that use design weights are natural from the design-based standpoint; model-assisted smoothing such as smoothing of residuals under a working model can leverage auxiliary information and improve efficiency, while still allowing design-based variance estimation in some cases.

4. Density estimation for survey data

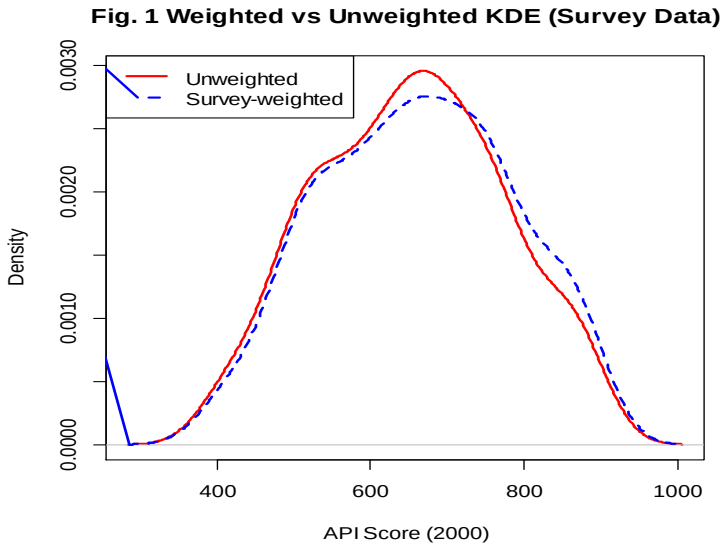
4.1 Weighted kernel density estimator

A natural adaptation of the KDE to survey data is the weighted kernel density estimator:

$$\hat{f}_w(x) = \frac{1}{Nh} \sum_{i=1}^{n} w_i K\left(\frac{x - X_i}{h} \right) \tag{2}$$

where the sum is over sampled units *s*, w_i are sampling weights (often $\frac{1}{\pi_i}$), and $\hat{N} = \sum_{i \in s} w_i$ estimates the population size. Equivalently one can use weights normalized to sum to one. This estimator generalizes the unweighted KDE and reduces to it when weights equal unity. The weighted KDE has been proposed and studied in the literature as a design consistent estimator of the population density under suitable regularity available in literature (Horvitz & Thompson, 1952; Särndal et al., 1992).

Under a design-based framework, $\hat{f}_w(x)$ can be shown to estimate the finite-population density (viewing the finite population as empirical measure), and its bias will depend on the kernel smoothing as in the *iid* case, while the variance must account for the sampling correlations induced by design such as clustering and unequal weights. If weights are informative (i.e., inclusion probabilities correlated with the variable of interest beyond what auxiliary information accounts for), failure to use weights can result in biased estimates (Lohr, 2010). In Fig. 1, this phenomenon of weighting and its effect is shown. Survey-weighted kernel density estimation illustrated in Fig. 1 was implemented using probability-weighted kernel methods rather than the svysmooth() function, which performs nonparametric regression rather than density estimation. Probability weights were normalized to ensure proper density scaling.



4.2 Normalization and population size

Normalization ensures that the total area under the estimated curve integrates to one. In Kernel Density Estimation (KDE), this will help to create a valid probability density function (PDF), while population size (sample size) influences the smoothness and reliability of the estimate. KDE works by placing a kernel - a small, pre-defined density function, such as Gaussian, over each data point. Each individual kernel also normalized so as to have an area of one. These kernels are summed up and the total sum is divided by *n* - the number of data points so that the resulting, composite curve also has a total area of one. Normalization allows the height of the curve to represent the *density*, not the raw count, making it possible to compare distributions across different datasets, even if they have different total counts or scales.

The size of sample data or population is a critical factor that affects the quality of the KDE depending on whether it is large or small. For instance, with limited data, the KDE can be "noisy" or "wiggly". The estimation is more sensitive to the specific locations of the few data points collected, and the chosen bandwidth parameter has a more pronounced effect. On the other hand when number of data points increases, the KDE becomes a more stable and smooth approximation of the true underlying population distribution. Features in the distribution become clearer, and the estimate is made more robust.

It should be noted that calculating KDE can be computationally intensive for extremely large datasets because the value at any single point requires processing all data points. This may therefore require use of alternative approaches like data binning to reduce computational demands.

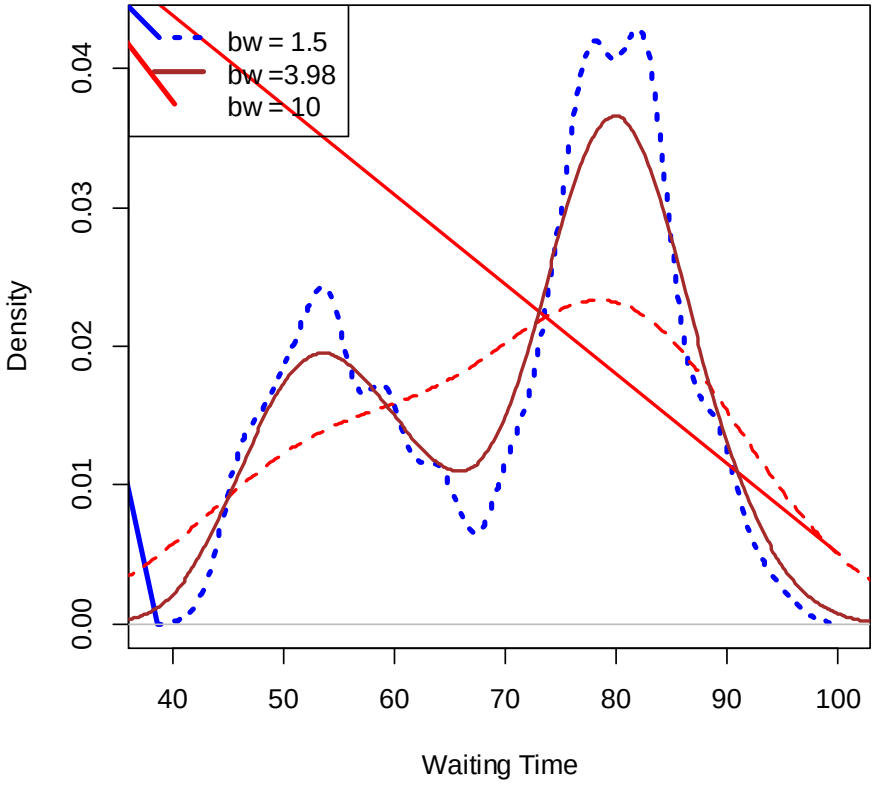
4.3 Bandwidth selection under survey designs

In nonparametric estimation, the most important parameter is bandwidth (*h*) also called the smoothing parameter. Its selection is more delicate with weighted data. Standard plug-in or cross-validation selectors assume iid structure and equal weights; direct application can under-smooth or over-smooth in the presence of unequal weights and cluster dependence. Practical strategies include using the weighted sample size. The Kish effective sample size can be used to adjust bandwidth selectors (Kish, 1965), or employing design-based cross-validation that computes a leave-one-out criterion weighted by survey weights. Model-assisted approaches can estimate optimal bandwidths by estimating integrated squared bias and variance terms under a superpopulation model and plugging in estimates that account for design effects.

Besides the Kish's effective sample size that provides a rough correction by replacing sample size *n* by the effective sample size n_{eff} resulting in bandwidth rules of thumb, more principled methods derive asymptotic *MISE* under the design and choose *h* to minimize estimated *MISE*; these approaches require estimating design-adjusted variance components and smoothness of the function. Generally, smaller bandwidths make the graph wiggly while a bigger one over-smoothes the graph. It therefore requires one to obtain an optimal smoothing parameter as earlier discussed. Fig. 2 illustrates the phenomenon. The graph also indicates

that bias increases with the bandwidth contrary to the variance which increases for smaller band caused by overfitting while reducing as it increases as a result of under fitting. The solution is therefore to pick the bandwidth at the equilibrium point.

Fig. 2 Effect of Bandwidth on KDE



4.4 Boundary correction and support constraints

Survey data often include variables with bounded support. For example income, expenditure, nad ages among others are truncated at zero. As in the iid case, boundary bias can distort density estimates near edges. Among the methods for bias correction developed for iid data include reflection, boundary kernels, and local polynomial methods that carry over to the weighted setting, but weights must be incorporated consistently in the local polynomials or reflected contributions.

Reflection technique is a widely used and simple method that augments the original dataset by mirroring data points across the boundaries. From this expanded dataset, the standard KDE is applied to it and the resulting density estimates within the original boundaries are used. This method does well, especially if the true density's derivative is near zero at the boundary. Another method is transformation. In this approach the bounded data is mapped onto an unbounded space by using a log or logit function for data on a [0,1] interval. It then applies the standard KDE in the transformed space, before transforming the density estimate back to the original scale. As in the other case this approach can handle boundary problem effectively, particularly for specific density shapes like those with singularities near a boundary. Additionally specialized, asymmetric kernel functions such as Gamma or Beta kernels can be designed for use near the boundaries. These kernels are inherently non-negative and adapt their shape based on the location relative to the boundary to minimize bias.

Local polynomial density estimators estimate the density (and derivatives) by locally fitting polynomials to the empirical distribution; they have superior boundary properties and can be adapted to weighted samples by performing weighted local fits where sample weights combine survey weights and kernel weights (Loader, 1999; Fan & Gijbels, 1996).

Fig. 3 Boundary Correction- Technique Comparison

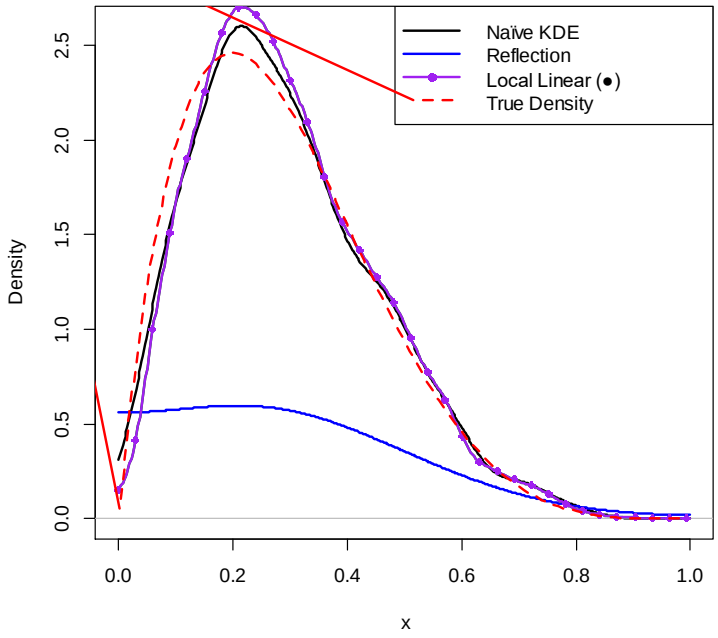


Fig. 3 illustrates boundary bias in kernel density estimation using the Old Faithful waiting-time data. The naive estimator exhibits downward bias near the lower support limit. Reflection and transformation methods mitigate this bias, while the local polynomial estimator provides adaptive boundary correction without artificial data manipulation.

While reflection correction improves bias near the boundary, it may introduce mild over-smoothing away from the boundary due to artificial symmetry. In the graph of Fig. 3, this case is clear.

5. Variance estimation and inference

Accurately estimating variability is crucial for inference and comparison of densities across domains. Several approaches are available that include linearization, replication methods (Jackknife, Balance Repeated Replication-BRR and Bootstrap). Linearization (Taylor series) is a method which treats the estimator as a smooth functional of population totals and approximates its variance by linearizing around expectation. For the weighted KDE, one can express $\hat{f}_w(x)$ as a Horvitz–Thompson-type estimator of a total of kernel contributions and apply standard variance formulas for totals under complex designs (Horvitz & Thompson, 1952; Särndal et al., 1992). Linearization yields closed-form variance approximations that can be computed using inclusion probabilities π_{ij} where applicable. However, deriving and implementing linearization for bandwidth selection or functionals of densities such as modes or quantiles can be algebraically involved.

Replication methods create pseudo-samples (replicates) by deleting or reweighting parts of the sample and computing the estimator on each replicate. The variability across replicates estimates the variance. Common schemes include the delete-1 jackknife, Balanced Repeated Replication (BRR), and replicate-weight bootstrap methods. Replication is flexible and handles complex designs (stratification, clustering, unequal probabilities) provided replicate construction respects design structure (Shao & Sitter, 1996). When using replication with KDEs, one recomputes the entire smooth estimate on each replicate, then measures variability pointwise (or for derived functionals), yielding design-consistent variance estimates.

Bootstrap methods adapted to survey data have been developed, rescaling bootstrap, Rao–Wu bootstrap, and other adaptations that recreate sampling variability consistent with the design (Rao & Wu, 1988; Shao & Sitter, 1996). For kernel-based estimators these methods require recomputing the KDE for each bootstrap replicate; care is needed with bandwidth choice in samples obtained through the process of re-sampling by fixing the bandwidth from the full sample or reselecting bandwidth per replicate. The two strategies obviously have pros and cons.

When interest focuses on features like modes, peaks, or tail probabilities, variance estimation can be challenging. Delta-method approximations (linearization) can work for smooth functionals; resampling methods are often preferred for complex or non-smooth functionals. It is also important to note that replication methods are easy to implement when survey software provides replicate weights. They automatically incorporate calibration adjustments if replicate weights are generated after calibration. Linearization may be more computationally efficient for point-wise variance estimates but requires derivations that may not be straightforward.

6. Discussion and conclusions

Nonparametric density estimation in survey sampling is both conceptually straightforward and practically subtle. Weighted kernel estimators provide a simple and design-consistent extension of classical KDEs, but correct inference requires careful attention to sampling design, weight variability, and dependence structures. Bandwidth selection remains the crucial practical problem; design-adjusted selectors or sensitivity analysis are recommended. Variance estimation benefits from replication methods when possible, while linearization offers efficient approximations when tractable.

Future research directions include more systematic development of bandwidth selectors explicitly minimizing design-based MISE, improved resampling algorithms tailored to high-weight variability scenarios, and integration of modern nonparametric techniques (e.g., adaptive kernels, deep-learning-based density estimators) with survey calibration constraints. As large-scale surveys increasingly incorporate complex designs and auxiliary big-data sources, robust nonparametric tools that respect design principles will be essential for reliable population inference.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1]. Devroye, L., & Györfi, L. (1985). *Nonparametric density estimation: The L1 view*. Wiley.
[2]. Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall.
[3]. Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.
[4]. Kish, L. (1965). *Survey sampling*. Wiley.
[5]. Loader, C. (1999). *Local regression and likelihood*. Springer.
[6]. Lohr, S. (2010). *Sampling: Design and analysis* (2nd ed.). Brooks/Cole.
[7]. Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3), 1065–1076.
[8]. Rao, J. N. K. (2003). Small area estimation. *Wiley-Interscience*.
[9]. Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 231–241.
[10]. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3), 832–837.
[11]. Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer.
[12]. Shao, J., & Sitter, R. R. (1996). The bootstrap and other resampling methods in survey sampling. *Canadian Journal of Statistics*, 24(2), 315–335.
[13]. Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.
[14]. Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall.