

## DENSITY ESTIMATION WITH COMPLEX SURVEY DATA

Sharon Lohr and Trent Buskirk<sup>1</sup>

### ABSTRACT

Kernel density estimation has been used with great success with data that fit the usual iid framework. The methods for iid data, however, do not directly apply to data from stratified multistage samples. We develop and present finite-sample and asymptotic properties of a modified density estimator introduced in Buskirk (1998) and Bellhouse and Stafford (1999); this estimator incorporates both the sampling weights and the kernel weights. The estimator is illustrated using data from the U.S. National Crime Victimization Survey and the U.S. National Health and Nutrition Examination Survey.

KEY WORDS: Smoothing; Sampling weights; Kernel estimation; Quantile estimation.

### RÉSUMÉ

L'estimation d'une fonction de densité du noyau a été utilisée avec succès pour des données obtenues à partir d'échantillons aléatoires simples. Par contre les techniques pour données iid ne s'appliquent pas directement à des données d'échantillons stratifiés à degrés multiples. On présente ici des propriétés asymptotiques ainsi que des propriétés d'échantillonnage finies pour un estimateur présenté par Buskirk (1998) et Bellhouse et Stafford (1999); cet estimateur utilise les poids d'échantillonnage et les poids du noyau. On applique cet estimateur à des données provenant de la U.S. National Crime Victimization Survey et de la U.S. National Health and Nutrition Examination Survey.

MOTS CLÉS : Lissage; poids d'échantillonnage; estimation du noyau; estimation des quantiles.

### 1. INTRODUCTION

Nonparametric kernel density estimation is commonly used to display the shape of a data set without relying on a parametric model. Rosenblatt (1956) and Parzen (1962) provided early results on kernel density estimation; since then, much research has been done in the area. Wand and Jones (1995) summarized some of the work done through the mid-1990's.

In most previous work, it is assumed that  $Y_1, \dots, Y_n$  are independent and identically distributed (iid) continuous random variables with common density  $f$ . Using a kernel function  $K$  and a positive bandwidth  $h$ ,  $f(x)$  is estimated by

$$\hat{f}_{iid}(x; h) = (hn)^{-1} \sum_{i=1}^n K[(x - Y_i)/h].$$

Density estimates are of interest in survey samples for many of the same reasons they are of interest in the iid setting: they provide a snapshot of the shape of the data and a means for comparing different populations. The iid assumptions, however, are often not valid for data from a complex survey. Stratification may reflect a violation of the identically distributed assumption, and clustering may violate the independence assumption.

To adapt density estimation to the survey setting, Buskirk (1998, 1999) and Bellhouse and Stafford (1999) independently proposed incorporating the survey weights into a density estimator of the form of  $\hat{f}_{iid}$ . Buskirk (1998, 1999) concentrated on a direct analogue; Bellhouse and Stafford (1999) also

<sup>1</sup> Sharon Lohr, Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 USA, sharon.lohr@asu.edu.  
Trent Buskirk, Department of Mathematics and Statistics, University of Nebraska, Lincoln, NE 68588-0323 USA, tbuskirk@math.unl.edu

considered the use of binned survey data in density estimation.

The research on density estimation in survey data has roots in previous work on estimating quantiles using an empirical cumulative distribution function (ECDF) for the finite population; this approach was pioneered by Woodruff (1952). Chambers and Dunstan (1986) introduced a model-based approach with auxiliary information, and Rao et al. (1990) proposed an estimator that is robust to model misspecification because it incorporates the sampling weights as well as auxiliary variables. Francisco and Fuller (1991) established design-based asymptotic properties within the context of an assumed superpopulation model. With this work as foundation, Korn et al. (1997) introduced a method for smoothing the ECDF. Korn and Graubard (1998a) suggested nonparametric smoothing as a way to display bivariate relations from survey data, and Crowley et al. (1996) proposed smoothing methods for spatial survey data.

In this paper, we present properties of a density estimator for complex survey data. The estimator and some finite population properties are given in Section 2. Consistency and asymptotic normality under design- and model-based frameworks are addressed in Section 3. Section 4 presents applications to the U.S. National Crime Victimization Survey and to the National Health and Nutrition Examination Survey.

## 2. THE SAMPLE WEIGHTED KERNEL DENSITY ESTIMATOR

Let  $U = \{1, \dots, N\}$  denote the index set of the finite population of  $N$  units. A probability sample  $S$  of size  $n$  is taken from  $U$  with  $P_D \{i \in S\} = \pi_i$  for  $i = 1, \dots, N$  (the subscript  $D$  indicates the probability distribution induced by the design). The probability that units  $i$  and  $j$  are both in the sample is denoted by  $\pi_{ij} = P_D \{i \in S, j \in S\}$ . The sampling weight for observation  $i$  is  $w_i = 1/\pi_i$ . The quantity  $y_i$  is observed on unit  $i$ .

If every population unit were observed, then a density estimator corresponding to the iid estimator would be

$$\hat{f}_U(x; h) = (hN)^{-1} \sum_{i=1}^N K[(x - y_i)/h].$$

Buskirk (1998, 1999) and Bellhouse and Stafford (1999) proposed estimating  $\hat{f}_U(x; h)$  by using the sampling weights. Define the sample weighted kernel density estimator (SWKDE) by

$$\hat{f}_S(x; h) = (hw)^{-1} \sum_{i \in S} w_i K[(x - y_i)/h]$$

where  $w = \sum_{i \in S} w_i$ . We assume throughout that the kernel function  $K$  satisfies the following conditions:

$$(K1) \quad K(x) > 0 \text{ and } K(x) = K(-x) \text{ for all } x.$$

$$(K2) \quad \int_{-\infty}^{\infty} K(x) dx = 1.$$

$$(K3) \quad \int_{-\infty}^{\infty} x^4 K(x) dx < \infty.$$

$$(K4) \quad K(x) \text{ is unimodal with mode at } 0.$$

Two properties follow immediately from assumptions (K1)-(K4). If a constant bandwidth  $h$  is used for all  $x$ , then  $\hat{f}_S$  is itself a probability density function. Moreover, if the support of  $K$  is contiguous, then

$$\int_{-\infty}^{\infty} x \hat{f}_S(x; h) dx = w^{-1} \sum_{i \in S} w_i y_i.$$

Thus the expected value of  $X$ , if  $X$  has density  $\hat{f}_S$  and the bandwidth is constant, is the usual ratio estimate of the finite population mean.

The SWKDE also provides a method for estimating quantiles. Let  $\theta$  represent 100<sup>th</sup> percentile. Denote the ordered sample values by  $y_{1:n}, \dots, y_{n:n}$  and let  $w_{1:n}, \dots, w_{n:n}$  be the weights corresponding to  $y_{1:n}, \dots, y_{n:n}$ . Then  $\theta$  may be estimated by  $\hat{\theta}$  satisfying  $\int_{-\infty}^{\hat{\theta}} \hat{f}_S(x) dx = q$ . As a consequence,  $\hat{\theta}$  is between  $y_{t:n}$  and  $y_{t+1:n}$ , where  $\sum_{i \leq t} w_{i:n} / w \leq q$  and  $\sum_{i > t} w_{i:n} / w \leq 1 - q$ . This differs from many of the other methods proposed for estimating quantiles in that we use the smoothed density to interpolate between  $y_{t:n}$  and  $y_{t+1:n}$ . Francisco and Fuller (1991) use a step function to estimate the cumulative distribution function, and estimate  $\theta$  by  $y_{t+1:n}$ . Under their regularity conditions,  $\hat{\theta}$  has the same asymptotic properties as the estimators in Francisco and Fuller (1991). However, we expect that our smoothed estimate will behave better in small samples when the underlying superpopulation density is continuous.

Chambers and Dunstan (1986), Rao et al. (1990), Chambers et al. (1993), and Rao (1994) all use auxiliary information in estimating the cumulative

distribution function, so their estimates of quantiles can be expected to be somewhat more efficient if the auxiliary variables are correlated with  $y$ . We note that auxiliary information can be incorporated into the smoothed estimate of the quantile by calibrating the sampling weights. Likewise, the weights in the SWKDE may also be modified to incorporate nonresponse adjustments.

### 3. ASYMPTOTIC PROPERTIES AND INFERENCE

We now examine consistency of  $\hat{f}_S(x; h)$  under different sampling designs and superpopulation models. Because of space limitations, the reader is referred to Buskirk (1999) for proofs. For design-based inference, we use the set-up of Isaki and Fuller (1982), with a sequence of nested finite populations  $U(t)$  and corresponding samples  $S(t)$  as  $t \rightarrow \infty$ . The corresponding population and sample sizes are  $N(t)$  and  $n(t)$ . Note that the samples from successive superpopulations need not be nested.

For model-based inference in the sample survey setting, we assume  $Y_1, \dots, Y_{N(t)}$  are distributed according to some joint probability distribution  $g$ , and that  $y_1, \dots, y_{N(t)}$  is a realization of  $Y_1, \dots, Y_{N(t)}$  that gives the measurements in the  $t^{\text{th}}$  finite population. Probabilities in the model-based setting are denoted by the subscript  $M$ . Since interest is often in a theoretical underlying density  $f(x)$  rather than in the finite population quantity  $\hat{f}_U(x; h)$ , we also examine asymptotic properties under the combined framework discussed in Pfeffermann (1993). The main interest in this framework is in estimating the superpopulation density  $f$ , but an estimator that is design-based consistent for the corresponding finite population quantity will be consistent under the combined distribution if the model holds for the finite population; this approach can provide protection against model misspecification.

In the design-based setting, we note that  $\hat{f}_S(x; h) = w^{-1} \sum_{i \in S} w_i u_i$ , where  $u_i = h^{-1} K[(x - y_i)/h]$ . Thus, as pointed out in Bellhouse and Stafford (1999),  $\hat{f}_S(x; h)$  is approximately design-unbiased for  $\hat{f}_U(x; h)$  for standard designs, and the design-based variance of  $\hat{f}_S(x; h)$ ,  $V_D[\hat{f}_S(x; h)]$ , may be calculated by

standard survey methods described in Lohr (1999). Asymptotic results, however, depend on the sample size, the sampling design, and the bandwidth  $h$ , which is assumed to converge to 0 as  $n(t) \rightarrow \infty$ . In a simple random sample, for example, the SWKDE is the sample average of the  $u_i$  for units in the sample. The observations  $u_i$ , though, are a function of the bandwidth  $h$  which is converging to 0; consequently, standard results on consistency and central limit theorems for finite population sampling which treat the  $u_i$  as fixed quantities (see for example Krewski and Rao, 1981) do not directly apply. In the following, we write  $h = h(t)$  to treat the convergence of  $h$  to 0. We then specify conditions for the rate of convergence of  $h$  to zero in order to obtain consistency in the different frameworks for inference.

#### 3.1 Stratified Sampling

First, consider stratified random sampling with  $L(t)$  strata in population  $U(t)$ . Stratum  $k$  has  $N_k(t)$  population and  $n_k(t)$  sample units, with  $N(t) = \sum_{k=1}^{L(t)} N_k(t)$  and  $n(t) = \sum_{k=1}^{L(t)} n_k(t)$ . Let  $W_k(t) = N_k(t)/N(t)$ . We allow for asymptotic inference in stratified random sampling when either the sample sizes within each stratum increase with  $t$ , or when the number of strata  $L(t)$  increases with  $t$ . We need the following assumptions for design-based consistency:

- (S1)  $n_k(t) \geq 2$  for all  $k$
- (S2)  $\max_{1 \leq k \leq L(t)} N_k(t)/n_k(t) = O(1)$
- (S3)  $N(t)h^2(t) \rightarrow \infty$  and  $h(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

Theorem 1. Suppose that conditions (S1)-(S3) hold in stratified random sampling. Then  $V_D[\hat{f}_S(x; h)] \rightarrow 0$  as  $t \rightarrow \infty$ , uniformly in  $x$ .

Condition (S2) of the theorem ensures that all of the strata are represented in the sample as  $t \rightarrow \infty$ . Condition (S3) is stronger than the usual condition for pointwise consistency in the iid model; namely, that  $nh \rightarrow \infty$ . The condition  $nh \rightarrow \infty$  is also the one used in Bellhouse and Stafford (1999). The stronger condition (S3) is needed here because pointwise consistency is required for every possible finite population. Buskirk (1999) gives a counterexample that shows  $nh \rightarrow \infty$  is not sufficient for design-based pointwise consistency.

If we use model-based inference, or inference under the combined distribution induced by the model and design, however, the condition  $nh \rightarrow \infty$  is sufficient for consistency. For a superpopulation model consistent with the motivation for a stratified design, assume that  $Y_{11}, \dots, Y_{1, N_1}, \dots, Y_{L, N_L}$  are independent and that  $Y_{kj}$  has density  $f_k$ . The overall density is a mixture,  $f_t(x; h) = \sum_{k=1}^L W_k(t) f_{kt}(x; h)$ . Then if the stratum densities and their absolute second derivatives are uniformly bounded, the SWKDE is consistent under the model and under the combined distribution when  $nh \rightarrow \infty$ . The exact conditions needed are given in Theorem 2.

Theorem 2. In the context of stratified random sampling, assume conditions (S1) and (S2) hold. Also assume that  $\sup_x \max_{1 \leq k \leq L(t)} f_{kt}[x; h(t)] = O(1)$ ,

$$\sup_x \max_{1 \leq k \leq L(t)} \left| f_{kt}''(x; h(t)) \right| = O(h(t)^{-1}),$$

$$N(t)h(t) \rightarrow \infty, \quad \text{and} \quad h(t) \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty.$$

Then  $V_M[\hat{f}_S(x; h(t))] \rightarrow 0$  and

$$E_M V_D[\hat{f}_S(x; h(t))] \rightarrow 0 \text{ as } t \rightarrow \infty.$$

We show design-based asymptotic normality under two frameworks. The first assumes that the number of strata are bounded but sample sizes within the strata increase; the second assumes that the individual stratum sizes are bounded but the number of strata increase. Assumptions (S4)-(S5) below give conditions for the former case, and assumptions (S5)-(S8) give conditions for the latter case. Let  $v[\hat{f}_S(x; h(t))]$  be the sample variance of the estimator in stratified sampling, and let  $V_k$  be the population variance for stratum  $k$ .

(S4) For each  $k$ ,  $n_k(t) \rightarrow \infty$  and  $N_k(t) - n_k(t) \rightarrow \infty$  as  $t \rightarrow \infty$

(S5)  $0 < \lim_{t \rightarrow \infty} \hat{f}_{U(t)}(x) \leq B < \infty$  for all  $t$

(S6)  $\max_{1 \leq k \leq L(t)} W_k(t) = O(L(t)^{-1})$

(S7)  $\max_{1 \leq k \leq L(t)} n_k(t) = O(1)$

(S8)  $n(t) \sum_{k=1}^{L(t)} W_k^2(t) n_k^{-1}(t) V_k \rightarrow \gamma > 0$

Theorem 3. Assume conditions (S1)-(S3) hold. Also assume that either (S4)-(S5) or (S5)-(S8) hold. Then

$[\hat{f}_{S(t)}(x) - \hat{f}_{U(t)}(x)] / \sqrt{V_D[\hat{f}_{S(t)}(x)]}$  converges in distribution to a standard normal. Furthermore,  $v[\hat{f}_{S(t)}(x)] - V_D[\hat{f}_{S(t)}(x)]$  converges to zero in probability.

Theorem 3 thus provides a basis for design-based inference for density estimates. Because the SWKDE is essentially an estimator of a sample mean, the jackknife estimate of the variance will coincide with  $v[\hat{f}_S(x; h(t))]$ . In light of Theorem 3 it follows that the jackknife provides a consistent estimate of the design-based variance.

### 3.2 Cluster Sampling

Cluster sampling presents different challenges for density estimation than does stratification. In stratification, each stratum is guaranteed to be represented in the sample; if the underlying densities  $f_k$  differ among the strata, each density will at least be represented by two data points. The danger in cluster sampling, if clusters have different underlying densities, is that a cluster making an important contribution to the overall density  $f$  may not be sampled; this situation may lead to poor estimation of  $f$ . We thus need regularity conditions to ensure that no one cluster will dominate the density estimation. We assume that in the population and sample indexed by  $t$ , that there are  $Q(t)$  primary sampling units and that psu  $j$  has size  $H_j(t)$ ;  $\pi_j$  represents the inclusion probability of the psu  $j$ . The following conditions, adapted from Isaki and Fuller (1982) and Korn and Graubard (1998b), then give design-based consistency in the cluster sampling setting:

(C1)  $2 \leq H_j(t) < B < \infty$

(C2) For some  $\delta_1, \delta_2$ ,  $0 < \delta_1 < \pi_j(t) < \delta_2 < 1$  for all  $j$

(C3)  $\pi_i(t)\pi_j(t) - \pi_{ij}(t) \leq \alpha(t)\pi_i(t)\pi_j(t)$ , where  $\alpha(t) = O(1/Q(t))$

(C4)  $Q(t) \rightarrow \infty$  and  $Q(t)h^2(t) \rightarrow \infty$  as  $t \rightarrow \infty$ .

Often, in inference under the combined distribution, the random variables generating the finite population are assumed to be iid. This assumption causes difficulty for density estimation, however, because it implies that there are no clustering effects in the superpopulation. It seems reasonable that units in the same cluster would be considered dependent in the superpopulation. We introduce here a model for the

densities that supports the concept of cluster sampling. Let  $Y_{jk}$  represent the  $k^{\text{th}}$  unit in cluster  $j$ . Let  $(Y_{jk}, Y_{jl})$  have joint density  $g$ , with the property that  $\int g(x, u)du = \int g(u, x)du = f(x)$

and assume that  $Y_{ik}$  and  $Y_{jl}$  are independent if  $i \neq j$ . This property is satisfied, for example, if the random variables generating the finite population satisfy the conditions for the one-way random effects model with normal errors. Buskirk (1999) derives the mean squared error of the SWKDE under this general model for the superpopulation structure and under the combined distribution. If conditions (C1)-(C3) are met, and if  $Q(t)h(t) \rightarrow \infty$  as  $t \rightarrow \infty$ , then the SWKDE is consistent for  $f$  under the presumed model and under the combined distribution as well.

#### 4. APPLICATIONS

Violence against women is of worldwide concern; one issue of interest is the age distribution of victims of sexual assault. Do sexual assault victims have the same age profile as victims of other types of assault? To investigate this question, we estimated the density for ages of sexual assault victims, and for ages of female victims of non-sexual assault, using data from the U.S. National Crime Victimization Survey (NCVS). The NCVS is an ongoing stratified multistage sample with rotating panel design; although it is designed to be approximately self-weighting, nonresponse and ratio adjustments vary the final weights. Figures 1 and 2 show the estimated density function for ages of female victims of sexual assault and for victims of other assaults, using the 1994 NCVS.

Figure 1. Density estimate for the ages of female victims of non-sexual assault crimes, using a triweight kernel and a four-year bandwidth.

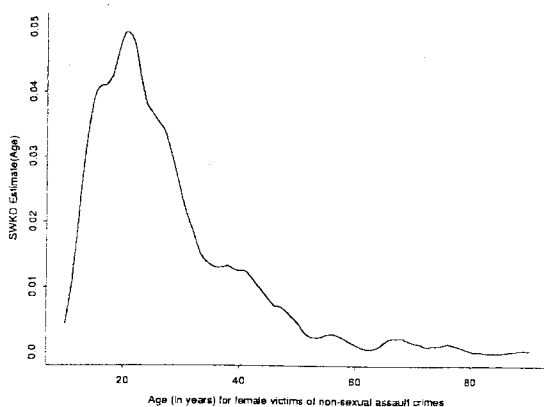


Figure 2. Density estimate for the ages of female victims of sexual assault, using the Epanechnikov kernel function and a five-year bandwidth.

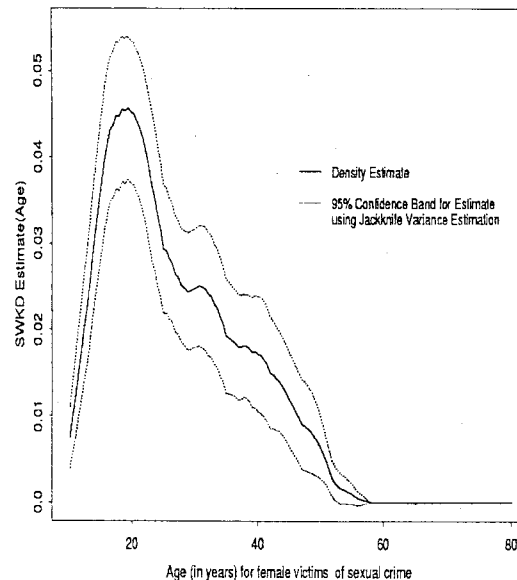
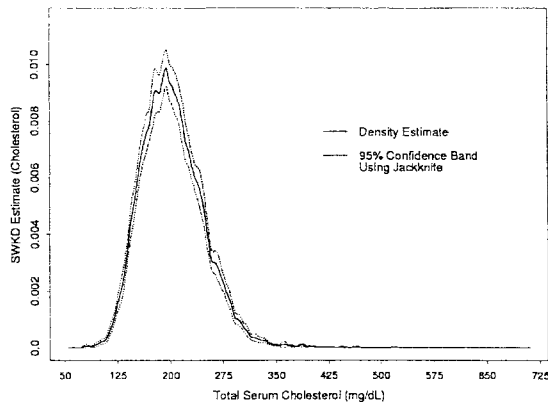


Figure 1 presents a density estimate without confidence bands; in Figure 2, approximate design-based confidence bands are computed using the jackknife estimate of the variance. The density tapers off more quickly with age for non-sexual assault victims than for the sexual assault victims; further investigation shows that this is because of sexual assaults of divorced women in their mid-thirties. The bandwidth chosen for the sexual assault density in Figure 2 is larger than that for the victims of non-sexual assault because the effective sample size is larger in the domain of non-sexual assault victims. See Buskirk (1999) for more discussion of optimal bandwidth choice under the combined distributions specified for stratified and cluster sampling.

Figure 3 estimates the density of total serum cholesterol using data from the U.S. National Health and Nutrition Examination Survey III. It is readily seen that the distribution is slightly skewed to the right. Because of the large sample size, the jackknife confidence bands are tight around the density estimate. These do not incorporate the model-based bias, however. Future research includes methods for incorporating an estimate of the squared bias into confidence bands using the combined distribution.

Figure 3. Density estimate for the total serum cholesterol levels of U.S. adults aged 17 or older, using Epanechnikov kernel function and a 6.5-unit bandwidth.



### ACKNOWLEDGEMENTS

This research was partially supported by a grant from the U.S. National Institute of Justice. The authors thank David Bellhouse for making his paper available to them prior to publication, and for helpful discussions.

### REFERENCES

Bellhouse, D. and Stafford, J. (1999). « Density estimation from complex surveys ». *Statistica Sinica*, 9, 407-424.

Buskirk, T. (1998). « Nonparametric density estimation using complex survey data ». *Proceedings of the Survey Research Methods Section, American Statistical Association*, 799-801.

Buskirk, T. (1999). *Using Nonparametric Methods for Density Estimation with Complex Survey Data*. Ph.D. dissertation, Arizona State University.

Chambers, R. and Dunstan, R. (1986). « Estimating distribution functions from survey data ». *Biometrika*, 73, 597-604.

Cowling, A., Chambers, R.L., and Parameswaran, B. (1996). « Applications of spatial smoothing to survey data ». *Survey Methodology*, 22, 175-183.

Francisco, C. and Fuller, W. (1991). « Quantile

estimation with a complex survey design ». *Annals of Statistics*, 19, 454-469.

Isaki, C.T. and Fuller, W.A. (1982). « Survey design under the regression superpopulation model ». *Journal of the American Statistical Association*, 77, 89-96.

Korn, E. and Graubard, B. (1998a). « Scatterplots with survey data ». *The American Statistician*, 52, 58-69.

Korn, E. and Graubard, B. (1998b). « Variance estimation for superpopulation parameters ». *Statistica Sinica*, 8, 1131-1151.

Korn, E., Midthune, D., and Graubard, B. (1997). « Estimating interpolated percentiles from grouped data with large samples ». *Journal of Official Statistics*, 13, 385-399.

Krewski, D. and Rao, J.N.K. (1981). « Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods ». *Annals of Statistics*, 9, 1010-1019.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Parzen, E. (1962). « On estimation of a probability function and mode ». *Annals of Mathematical Statistics*, 33, 1065-1076.

Pfeffermann, D. (1993). « The role of sampling weights when modeling survey data ». *International Statistical Review*, 61, 317-337.

Rao, J.N.K., Kovar, J., and Mantel, H. (1990). « On estimating distribution functions and quantiles from survey data using auxiliary information ». *Biometrika*, 77, 365-375.

Rosenblatt, M. (1956). « Remarks on some nonparametric estimates of a density function ». *Annals of Mathematical Statistics*, 27, 186-190.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. New York: Chapman & Hall.

Woodruff, R. (1952). « Confidence intervals for medians and other position measures ». *Journal of the American Statistical Association*, 47, 635-646.