

Nonparametric and semiparametric estimation in complex surveys

F. Jay Breidt and Jean D. Opsomer

September 4, 2008

1 Introduction

1.1 Nonparametric and semiparametric methods

Nonparametric and semiparametric methods are rich classes of statistical tools that have gained acceptance in most areas of statistics. They make it possible to analyze data, estimate trends and conduct inference without having to fully specify a parametric model for the data. In the survey context, their use is much less widespread. In this article, we will focus on nonparametric and semiparametric methods in two important statistical areas: estimation of densities, and estimation of regression functions. Both of these areas have applications in survey estimation, for both descriptive and analytical uses.

We begin by a necessarily brief overview of the main nonparametric and semiparametric methods relevant to survey estimation. In this section, we will describe them for the case of independent and identically distributed (*iid*) data in order to introduce the methods. Subsequent sections will deal with the situation in which the observations are obtained from a complex survey.

We would like to note that the terms “nonparametric” and “semiparametric” have not been used consistently in the statistical literature, so there is no agreement

on which methods exactly fall into each of these two categories. Generally speaking, nonparametric methods are those that do not assume a parametric form for the main features of interest in the data (though there might be parametric assumptions on some of the “nuisance features”, e.g. the variance in the case of regression). In contrast, semiparametric methods use a combination of parametric and nonparametric specification for the main features of interest. Clearly, these descriptions are somewhat subjective and open to interpretation, so one person’s nonparametric method is another person’s semiparametric approach.

1.2 Kernel methods

Kernel methods are used for both density estimation and regression. We begin by describing the kernel density estimator, and restrict ourselves to the univariate case. Suppose we observe X_1, \dots, X_n , and we assume these x_i are *iid* from an unknown density $f_x(\cdot)$. The density is assumed to be a smooth function of x but otherwise unspecified. *Kernel density estimation* methods aim to estimate the density $f_x(\cdot)$ nonparametrically. Wand and Jones (1995) give a good introduction to these methods, and we only describe the main idea here. For a given value x , a simple kernel density estimator $\hat{f}_x(x; h)$ is defined as

$$\hat{f}_x(x; h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right), \quad (1)$$

where $K(\cdot)$ denotes the *kernel function* and the constant h is referred to as the *bandwidth*. In order to estimate the density function $f_x(\cdot)$ using (1), $\hat{f}_x(\cdot; h)$ is computed at each value x for which an estimator is needed, for instance on a dense grid of x -values, which can then be plotted or interpolated.

The kernel $K(\cdot)$ is usually a symmetric probability density, with the standard normal density being a common choice, but other functions can be used as well. The crucial feature of the kernel function is that it determines distance-based weights for the sample observations, to be used in the construction of (1). The bandwidth h determines the smoothness of the estimator $\hat{f}_x(x; h)$, with small values of h leading

to more “wiggly” estimates and large values resulting in smoother estimates. More precisely, the bandwidth determines the bias-variance trade-off for the kernel density estimator $\hat{f}_x(x; h)$, with large h having potentially larger bias but smaller variance than small h . A large literature is devoted to the determination of the best value for the bandwidth, and we will briefly return to this issue in later sections.

We now turn to the kernel-based regression estimation problem. Suppose that we have a dataset with observations $(X_1, Y_1), \dots, (X_n, Y_n)$, and we are interested in estimating the function $m(\cdot)$ in the model

$$Y_i = m(x_i) + \varepsilon_i, \quad (2)$$

where $m(\cdot)$ is smooth but not further specified, and for simplicity we assume that the ε_i are *iid* with mean 0 and variance σ^2 . The most commonly used kernel method of nonparametric estimation of $m(\cdot)$ is *local polynomial regression*, with local linear regression a popular choice.

Let q represent the degree of the local polynomial regression. For a given value x , the estimator $\widehat{m}(x)$ is defined as $\widehat{\beta}_0$, where $\widehat{\beta}_0, \dots, \widehat{\beta}_q$ are found by solving the following weighted least squares problem:

$$\min \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) (Y_i - \beta_0 - \beta_1(x_i - x) - \dots - \beta_q(x_i - x)^q)^2.$$

This estimator can be written in matrix notation as

$$\widehat{m}(x) = \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}_x(h) \mathbf{X}_x \right)^{-1} \mathbf{X}_x^T \mathbf{W}_x(h) \mathbf{Y}, \quad (3)$$

with $\mathbf{e}_1 = (1, 0, \dots, 0)^T$, $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{W}_x = \text{diag}\{K((x_1 - x)/h), \dots, K((x_n - x)/h)\}$ and

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \dots & (x_1 - x)^q \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \dots & (x_n - x)^q \end{bmatrix}. \quad (4)$$

As for kernel density estimation, the function $m(\cdot)$ is estimated by computing $\widehat{m}(x; h)$ for any value x where an estimator of the function is needed. The bias-variance

tradeoff for $\widehat{m}(x)$ also again depends on the bandwidth h . It is clear from (3) that the local polynomial estimator can be written as a linear combination of the Y_i , $\widehat{m}(x) = \sum w_i(x)Y_i$, which will be useful when applying this nonparametric method in the survey context. We refer to Wand and Jones (1995) for further information on local polynomial regression, including its theoretical properties.

1.3 Spline methods and other methods

While kernel methods span both density and regression function estimation, spline methods are typically only used for the latter problem. We therefore again consider a dataset with observations $(X_1, Y_1), \dots, (X_n, Y_n)$ which are assumed to follow the model (2) with *iid* errors. While the function $m(\cdot)$ in (2) is still assumed to be smooth but otherwise unspecified, we now make the additional assumption that it is well approximated by a *spline function*. Polynomial spline functions are defined as

$$m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{j=1}^J \beta_{p+j} (x - \kappa_j)_+^p, \quad (5)$$

where $p \geq 1$ is the order of the spline, $\kappa_1, \dots, \kappa_J$ are a set of pre-specified breakpoints called *knots* and the function $(\cdot)_+^p$ denotes

$$(x - \kappa)_+^p = \begin{cases} (x - \kappa)^p & \text{if } x > \kappa \\ 0 & \text{otherwise.} \end{cases}$$

The linear ($p = 1$) and cubic ($p = 3$) spline models are common choices in practice. The linear splines are simple and continuous, and the cubic splines match up with a common type of smoothing splines, the natural cubic splines, which arise from a penalized optimization with penalty on the squared second derivative of the function. Other formulations of the spline function $m(x; \boldsymbol{\beta})$ are possible, in which the set of *basis functions* $\{1, x, \dots, x^p, (x - \kappa_1)_+^p, \dots, (x - \kappa_J)_+^p\}$ are replaced by a different set. For instance, *B-splines* (de Boor, 2001) are a widely used set of basis functions with better numerical properties than polynomial splines. Most of these formulations,

including B -splines, can be equivalently rewritten into the above polynomial spline, so that we will restrict our attention to (5).

A number of different spline regression methods exist, but we will focus here on *penalized spline regression*, because of its ease of use and relevance to the applications in survey estimation. An excellent overview of this method and its applications in a wide range of regression contexts is provided in Ruppert et al. (2003). It is clear from (5) that $m(x; \boldsymbol{\beta})$ is essentially a parametric function (albeit a complicated one), and deviations from a global p -th order polynomial can only occur at the knots, so that the flexibility of the spline as a representation of an unknown function is determined by the number and location of the knots. In order to ensure that $m(x; \boldsymbol{\beta})$ is sufficiently flexible, the penalized spline approach sets the number of knots J to be large, say as high as $J = n/4$, and places them at the appropriate quantiles of the x_i .

Fitting of the spline model to the observations is done by least squares minimization, but with a penalty added to ensure the existence of a solution and to reduce the potential increase in variance due to the large number of parameters needing estimation. Specifically, the estimator of $m(\cdot)$ is $m(\cdot; \widehat{\boldsymbol{\beta}})$ using expression (5), where $\widehat{\boldsymbol{\beta}}$ is the minimizer of

$$\sum_{i=1}^n \left(Y_i - \beta_0 - \beta_1 x_i - \dots - \beta_p x_i^p - \sum_{j=1}^J \beta_{p+j} (x_i - \kappa_j)_+^p \right)^2 + \lambda \sum_{j=1}^J \beta_{p+j}^2 \quad (6)$$

and λ is a fixed penalty. The penalty λ plays an analogous role to the bandwidth h in the kernel regression methods, in that it determines the bias-variance trade-off for the estimator $m(\cdot; \widehat{\boldsymbol{\beta}})$, with large values of λ resulting in potentially larger bias and smaller variance than small values. Since only the non-polynomial part of the spline coefficients is penalized in (6), λ determines the amount of deviation from a p -th degree polynomial function.

Because of the very flexible nature of the spline function (5) and the presence of the penalty λ that serves as a tuning constant, penalized spline regression is typically considered a nonparametric method. Nevertheless, it also shares many

characteristics of parametric regression, because the number of parameters is fixed (at $J + p + 1$) and the estimator is found as the solution to a global least squares problem.

Other spline regression methods are (unpenalized) spline regression and smoothing spline regression. In the former, a spline function with a small number of knots is specified and the function is fitted without penalization, so that careful attention needs to be paid to knot placement to avoid bias. In smoothing spline regression, the formulation of the approach is different from the above, but the estimator is essentially equivalent to a polynomial spline as in (5) but with a knot at every observation point x_i , and a penalty term on the derivative of the function. We do not pursue these methods further here, and instead refer to Ruppert et al. (2003, ch. 3).

Other important classes of nonparametric methods are available, many of which could be adapted for use in survey estimation. Orthogonal decompositions, in particular *wavelet* decomposition (Vidaković, 1999), is a nonparametric regression method with good statistical properties that is applicable in situations where the mean function is not necessarily smooth. Neural networks (Ripley, 1996) are a class of methods conceptually related to penalized spline regression, in which the parameters are found by nonlinear regression. Finally, methods based on classification such as classification and regression trees (CART; Breiman et al. 1984) and multivariate adaptive regression splines (MARS; Friedman, 1991) can be used as nonparametric regression methods.

1.4 Fitting more complex models

So far, we have discussed the situation in which the x_i are univariate observations. In surveys, the number of variables is typically large, so we would like to be able to apply nonparametric and semiparametric methods for multivariate data. In principle, it is indeed possible to directly extend all of the methods from the previous

sections to the multivariate case, but a number of constraints make this impractical for more than two or three dimensions. One issue is the so-called “curse of dimensionality,” which implies that model flexibility has to decrease as the dimension of the covariate space increases in order to obtain satisfactory fits. This could be done by increasing the amount of smoothing (by using a larger bandwidth or penalty) or using a reduced number of knots (in the case of splines), but more useful approaches are to replace the fully nonparametric model itself by more restricted model specifications. We discuss two important special cases of such models here: additive models and semiparametric models.

Let $\mathbf{X}_i = (X_{1i}, \dots, X_{Di})$ represent a vector of D covariates. In additive models, model (2) is replaced by

$$Y_i = m_1(X_{1i}) + \dots + m_D(X_{Di}) + \varepsilon_i, \quad (7)$$

where the functions $m_d(\cdot)$ are (typically) univariate and smooth but otherwise not restricted to belong to a specific parametric family. This model was made popular by Hastie and Tibshirani (1990), who proposed estimation methods based on *backfitting*. This approach, which is implemented in S-Plus and R, relies on iteratively applying one-dimensional methods such as local polynomial regression or spline regression to the residuals from the fits with respect to the other covariates. While other methods for fitting model (7) have since been proposed, backfitting remains popular today. When penalized spline regression is used as the fitting method, it is possible to fit model (7) without iterating by writing it as a penalized multiple regression problem, from which the spline parameters can be estimated directly (see Ruppert et al. (2003) for details). The package SemiPar (Wand et al. 2005) implements this approach in R.

In a semiparametric model, a nonparametric term is combined with parametrically specified components. Let $\mathbf{Z}_i = (Z_{1i}, \dots, Z_{Pi})$ represent the additional covariates to be modeled parametrically. A typical example of a semiparametric model

is

$$Y_i = m(X_i) + \mathbf{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad (8)$$

where the nonparametric term $m(X_i)$ could itself be multivariate and modeled as an additive model. Backfitting can be applied to fit model (8) as well, but other methods specially designed for semiparametric models are available. The semiparametric model is particularly useful when some of the covariates in a dataset are categorical, which by definition cannot be smoothed.

In addition to nonparametric regression for multivariate data, another important extension is for models with more complex mean structures, including nonparametric equivalents of generalized linear models. The generalized additive model (GAM) described in Hastie and Tibshirani (1990) has mean structure

$$E(Y_i | \mathbf{X}_i, \mathbf{Z}_i) = g(m_1(X_{1i}) + \cdots + m_D(X_{Di}) + \mathbf{Z}_i^T \boldsymbol{\beta}), \quad (9)$$

which combines a known link function $g(\cdot)$ with a mean additive model or a semiparametric model. This model makes it possible to perform common types of regression such as Poisson or logistic regression nonparametrically. The most common fitting method for this type of models is an iterative algorithm called *local scoring*, a combination of Fisher scoring and backfitting. Just like for additive models, this method uses univariate regression methods such as local polynomial and spline regression for the component functions.

2 Nonparametric Methods in Descriptive Inference from Surveys

We now consider the use of nonparametric methods in making inference about a finite, labelled population $U = \{1, \dots, i, \dots, N\}$. Associated with each label i are study variables y_i, z_i , etc (possibly vector-valued), which can in principle be observed without error if label i were sampled. Assume that for each $i \in U$, an auxiliary vector

\mathbf{x}_i is observed. Let $t_x = \sum_{i \in U} \mathbf{x}_i$. A probability sample $s \subset U$ is drawn according to a fixed-size sampling design $p(\cdot)$, where $p(s) = \Pr[\text{sample } s \text{ is selected}]$. Let $\pi_i = \Pr[i \in s] = \sum_{s:i \in s} p(s) > 0$ and $\pi_{ij} = \Pr[i, j \in s]$ for all $i, j \in U$.

We first consider *descriptive inferences* for this finite population, often done in terms of a point estimate and an associated confidence interval for a finite population parameter such as a total $t_y = \sum_{i \in U} y_i$ or mean $\bar{y}_U = N^{-1}t_y$. A proportion is a special case of the mean, with y_i equal to an indicator on some event. In particular, the finite population distribution function, denoted $F_y(z) = N^{-1} \sum_{i \in U} I_{\{y_i \leq z\}}$ with $I_{\{A\}} = 1$ if the event A is true, and 0 otherwise, is a proportion for each fixed z . Other interesting finite population parameters include ratios $\sum_{i \in U} y_i / \sum_{i \in U} z_i$ and vectors of regression coefficients,

$$\mathbf{B} = \left(\sum_{i \in U} \mathbf{x}_i^T \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i y_i.$$

Each of these examples is built up from finite population totals, and so a canonical problem of interest is estimation of the population total for a generic study variable y .

The Horvitz-Thompson estimator of t_y ,

$$\hat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{10}$$

(Horvitz and Thompson, 1952) provides an unbiased estimator for the population total t_y , with variance under the sampling design

$$\text{Var}_p(\hat{t}_y) = \sum_{i, j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j}. \tag{11}$$

If auxiliary variables are available for a survey, it might be possible to obtain estimators that are more efficient than \hat{t}_y .

It is of interest to improve upon the efficiency of the Horvitz-Thompson estimator by using the auxiliary information x_i . Motivation for such estimators is often provided by modeling the finite population of y_i 's as a realization from an infinite superpopulation, ξ , relating \mathbf{x}_i to y_i via

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i, \tag{12}$$

where ε_i is an independent sequence of random variables with mean zero and variance $\nu(\mathbf{x}_i)$. Standard superpopulation models are parametric, and typically linear, that is $\mu(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. The potential disadvantage of estimators motivated by a superpopulation model is inefficiency under model misspecification. If the regression model does not fit the data well, there is no improvement over a simple Horvitz-Thompson estimator and potentially even a loss of efficiency. To avoid the consequences of model misspecification, it is natural to replace the parametric specification by a nonparametric specification, in which $\mu(\cdot)$ is a smooth function of \mathbf{x} , and $\nu(\cdot)$ is smooth and strictly positive.

Once the model (whether parametrically or nonparametrically specified) is fitted to the sample data, there are at least two ways to incorporate its predictions into estimation of the finite population total. The first is a *model-based* approach, in which model-fitted values $\tilde{\mu}(\mathbf{x}_i)$ are used to predict only the non-sampled values of y :

$$\hat{t}_{MB} = \sum_{i \in U \setminus s} \tilde{\mu}(\mathbf{x}_i) + \sum_{i \in s} y_i. \quad (13)$$

Typically, model-based estimators of this type are asymptotically model unbiased and highly efficient when $\mu(\mathbf{x}_i)$ and $\nu(\mathbf{x}_i)$ are correctly specified, but biased and even inconsistent if the model is wrong. Inspired by the general applicability of nonparametric models, Kuo (1988), Dorfman (1992), and Chambers et al. (1993) have developed model-based estimators using nonparametric regression.

The second way to incorporate model predictions is *model-assisted*, and avoids the potential problems of model misspecification through a design bias adjustment. Model-assisted estimation relies on a model-fitted prediction $\hat{\mu}_i$ for all of the population elements but then corrects the possible design bias in that prediction. The resulting model-assisted regression estimator is of the form

$$\hat{t}_{MA} = \sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{y_i - \hat{\mu}_i}{\pi_i}. \quad (14)$$

An intuitive explanation of the design properties of the model-assisted regression estimator proceeds as follows. Let μ_i represent the regression fit for $\mu(\mathbf{x}_i)$ if the entire

population were observed. If these μ_i 's were known, then an exactly design-unbiased estimator of t_y would be the generalized difference estimator

$$t_y^* = \sum_{i \in U} \mu_i + \sum_{i \in s} \frac{y_i - \mu_i}{\pi_i} \quad (15)$$

(see Särndal *et al.* (1992), p. 221, for the parametric case). The design variance of the estimator would be

$$\text{Var}_p(t_y^*) = \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i - \mu_i}{\pi_i} \frac{y_j - \mu_j}{\pi_j}, \quad (16)$$

which we would expect to be smaller than (11), because the y_i 's should be “close to” the μ_i 's for any reasonable smoothing procedure under the superpopulation model.

In practice the μ_i 's are not known, but they are well-defined “parameters” of the finite population that can be estimated by the $\hat{\mu}_i$'s even if the superpopulation model (12) does not hold. As will be discussed further below, the resulting nonparametric model-assisted estimator can share many of the properties of linear model-assisted estimators familiar to survey statisticians, including design consistency.

As noted at the beginning of this section, the finite population distribution function $F_N(z) = N^{-1} \sum_{i \in U} I_{\{y_i \leq z\}}$ for each z is a special case of a population mean. The nonparametric model-based and model-assisted methods discussed below can thus be used without further modification to improve the precision of estimators of the finite population distribution function. The advantage of doing so is that the same survey weights can be used for estimating $F_N(z)$ for any z as well as the population mean \bar{y}_U , for all the survey variables. However, a number of special estimation methods have also been developed that take advantage of the special structure of $F_N(z)$. We refer to Chapter 36 for further information on this topic.

2.1 Nonparametric survey regression estimation using kernels

We now describe a number of ways in which nonparametric estimation can be implemented for descriptive inference. We begin by considering local polynomial re-

gression (LPR) for scalar x_i , as in Section 1.2. Let $\mathbf{y}_s = [y_i]_{i \in s}$ be the vector of y_i 's in the sample and define the local design matrix

$$\mathbf{X}_{si} = \left[\begin{array}{cccc} 1 & x_j - x_i & \cdots & (x_j - x_i)^q \end{array} \right]_{j \in s}, \quad (17)$$

corresponding to the design matrix in (4) evaluated at $x = x_i$, and the diagonal weighting matrix

$$\mathbf{W}_{si} = \text{diag} \left\{ \frac{1}{h} K \left(\frac{x_j - x_i}{h} \right) \right\}_{j \in s}.$$

The unweighted *local polynomial regression* estimator of $\mu(x_i)$ is then given by the intercept in the local, weighted least squares fit of the polynomial:

$$\tilde{\mu}(x_i) = (1, 0, \dots, 0) \left(\mathbf{X}_{si}^T \mathbf{W}_{si} \mathbf{X}_{si} \right)^{-1} \mathbf{X}_{si}^T \mathbf{W}_{si} \mathbf{y}_s. \quad (18)$$

Plugging these model fits into (13) then yields the model-based kernel regression estimator of Dorfman (1992).

One approach to producing a model-assisted estimator begins instead with the finite population local polynomial fit. Let $\mathbf{y}_U = [y_i]_{i \in U}$ be the vector of y_i 's for the entire finite population. Define the $N \times (q + 1)$ matrix

$$\mathbf{X}_{Ui} = \left[\begin{array}{cccc} 1 & x_j - x_i & \cdots & (x_j - x_i)^q \end{array} \right]_{j \in U},$$

and define the $N \times N$ matrix

$$\mathbf{W}_{Ui} = \text{diag} \left\{ \frac{1}{h} K \left(\frac{x_j - x_i}{h} \right) \right\}_{j \in U}.$$

The finite population local polynomial fit is then given by

$$\mu_i = \mathbf{e}_1^T \left(\mathbf{X}_{Ui}^T \mathbf{W}_{Ui} \mathbf{X}_{Ui} \right)^{-1} \mathbf{X}_{Ui}^T \mathbf{W}_{Ui} \mathbf{y}_U, \quad (19)$$

as long as $\mathbf{X}_{Ui}^T \mathbf{W}_{Ui} \mathbf{X}_{Ui}$ is invertible. The μ_i are the quantities that would be used in the difference estimation (15), if they were available. Since they are generally not available, they are estimated by design-weighted estimators $\hat{\mu}_i$, constructed by letting

$$\mathbf{W}_{si\pi} = \text{diag} \left\{ \frac{1}{\pi_j h} K \left(\frac{x_j - x_i}{h} \right) \right\}_{j \in s}.$$

and

$$\hat{\mu}_i = \mathbf{e}_1^T \left(\mathbf{X}_{si}^T \mathbf{W}_{si\pi} \mathbf{X}_{si} \right)^{-1} \mathbf{X}_{si}^T \mathbf{W}_{si\pi} \mathbf{y}_s, \quad (20)$$

provided $\mathbf{X}_{si}^T \mathbf{W}_{si\pi} \mathbf{X}_{si}$ is invertible. Plugging these fits into (14) then yields the model-assisted LPR estimator of Breidt and Opsomer (2000).

Breidt and Opsomer (2000) discuss the theoretical design and model properties of the local polynomial estimator, showing that the LPR estimator is design consistent and asymptotically design unbiased under a mild set of regularity conditions that we hereafter assume to hold. Asymptotically, the design mean squared error of \hat{t}_{MA} under LPR is equivalent to the variance of the generalized difference estimator,

$$\text{MSE}_p(\hat{t}_{MA}) = \text{E}_p \left(\hat{t}_{MA} - t_y \right)^2 \approx \sum_{i,j \in U} (y_i - \mu_i)(y_j - \mu_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}, \quad (21)$$

recalling that the μ_i have been defined as (unknown) finite population parameters. A design consistent and asymptotically design unbiased estimator of $\text{MSE}_p(\hat{t}_{MA})$ is

$$\hat{V}(\hat{t}_{MA}) = \sum_{i,j \in s} (y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}}. \quad (22)$$

Because each of the smoothed values $\hat{\mu}_i$ is a linear combination of the y_i in the sample, the LPR model-assisted estimator (14) can also be written in the same form, i.e. $\hat{t}_{MA} = \sum_s \omega_{is} y_i$ with ω_{is} not involving the y_i . It is readily checked that the weights ω_{is} are calibrated for the population size as well as for the totals of powers of the x_i up to degree q : $\sum_s \omega_{is} x_i^p = \sum_U x_i^p$ for $0 \leq p \leq q$. The LPR model-assisted estimator shares this property with the generalized regression estimators.

In simulation experiments reported in Breidt and Opsomer (2000), the local polynomial regression estimator was competitive with the classical survey regression estimator when the population regression function was linear, but dominated the regression estimator when the regression function was not linear. The estimator also performed well relative to other parametric and nonparametric estimators, both model-assisted and model-based. It generally dominated the Horvitz-Thompson estimator, and it dominated cubic regression and post-stratification estimators provided it was not oversmoothed. It was sometimes much better and never much worse

than two competing model-based nonparametric estimators. Though the efficiency of the nonparametric estimator depended on the choice of bandwidth parameter, the results were fairly insensitive to this choice, suggesting that large gains in efficiency can be attained for a variety of bandwidths.

The local polynomial method can be applied in virtually all situations where generalized regression estimation is used, as long as the value of auxiliary variable x_i is available for every element in the population and it is a continuous (not categorical) variable. Examples of the type of generalizations that are possible are provided by Deville and Goga (2004), who applied local polynomial regression to improve the efficiency of survey estimators when samples are taken on two occasions, and by Aragon et al. (2006), who considered quantile estimation.

2.2 Nonparametric survey regression estimation using splines

We next consider nonparametric survey regression estimation using splines, focusing on the special case of penalized regression splines with scalar x_i . For the superpopulation model (12), we assume now that the mean function $\mu(\cdot)$ can be written as in (5). We define $\mathbf{x}_i^T = (1, x_i, \dots, x_i^q, (x_i - \kappa_1)_+^q, \dots, (x_i - \kappa_J)_+^q)$, $\mathbf{X}_s = [\mathbf{x}_i^T]_{i \in s}$, $\mathbf{X}_U = [\mathbf{x}_i^T]_{i \in U}$ and $\mathbf{\Pi}_s = \text{diag}\{\pi_i\}_{i \in s}$. Further, define the diagonal matrix $\mathbf{A}_\alpha = \text{diag}\{0, \dots, 0, \alpha, \dots, \alpha\}$, with $q + 1$ zeros on the diagonal followed by J penalty constants α , corresponding to the J truncated polynomial terms in (5).

The unweighted sample spline estimator of $\mu(x_i)$, corresponding to the solution to the penalized least squares minimization (6), is then

$$\tilde{\mu}(x_i) = \mathbf{x}_i^T (\mathbf{X}_s^T \mathbf{X}_s + \mathbf{A}_\alpha)^{-1} \mathbf{X}_s^T \mathbf{y}_s. \quad (23)$$

Using $x_i = \pi_i$ in (23) and plugging in to (13), Zheng and Little (2003) have proposed a model-based survey regression estimator that uses penalized splines to account for the effect of non-ignorable design weights. They have further extended the penalized spline model-based survey regression estimator to the case of two-stage sampling in Zheng and Little (2004).

A model-assisted survey regression estimator based on penalized splines begins by first defining the population fit

$$\mu_i = \mathbf{x}_i^T (\mathbf{X}_U^T \mathbf{X}_U + \mathbf{A}_\alpha)^{-1} \mathbf{X}_U^T \mathbf{y}_U$$

and then estimates this finite population parameter with a sample-weighted version,

$$\hat{\mu}_i = \mathbf{x}_i^T (\mathbf{X}_s^T \mathbf{\Pi}_s^{-1} \mathbf{X}_s + \mathbf{A}_\alpha)^{-1} \mathbf{X}_s^T \mathbf{\Pi}_s^{-1} \mathbf{y}_s.$$

Plugging these design-weighted fits into (14) yields the penalized spline model-assisted survey regression (PSP) estimator proposed by Breidt et al. (2005).

This estimator has theoretical properties similar to those of the LPR estimator discussed above, including calibration for population totals of powers of x_i up to degree q , design consistency and asymptotic design unbiasedness (under mild conditions), and its design mean squared error can also be written as in (21). In simulation experiments reported in Breidt et al. (2005), it is shown that the PSP estimator is most often very similar to the LPR estimator. However, penalized spline regression offers a number of advantages over kernel-based methods that make it an attractive smoothing method in the model-assisted context. Incorporating multiple covariates as well as combinations of categorical variables, parametric and non-parametric terms, is straightforward, as shown in Aerts et al. (2002). Another important advantage is the relative ease with which PSP estimators can be computed, even for large datasets or datasets with regions of sparse data. Finally, an important practical consideration is that, since they are more closely related to parametric models, estimators based on spline models are easier to implement in existing survey estimation procedures.

Another class of nonparametric model-assisted estimators based on splines has been studied by Goga (2004, 2005). In both of these papers, Goga uses unpenalized regression splines for which the domain of the auxiliary variable is divided by a number of knots, a B -spline basis function is associated with each knot, and the number of knots goes to infinity so that the B -splines become dense on the domain. Goga

(2005) shows that the regression spline estimator is asymptotically design-unbiased and consistent, proposes a design-based variance approximation and shows that the anticipated variance is asymptotically equivalent to the Godambe-Joshi lower bound. Simulations show that the regression spline estimator has good properties. Goga (2004) applies this methodology to construct model-assisted estimators in the case of sampling on two occasions, with complete auxiliary information available on each occasion.

2.3 Other smoothing methods for survey regression estimation

While the estimators of Sections 2.1 and 2.2 can in principle be generalized directly to handle multivariate \boldsymbol{x}_i , the modeling approaches described in Section 1.4 are likely to be more useful in practice. Breidt et al. (2007) extend the LPR estimator to the semiparametric model (8), and show that the semiparametric model-assisted estimator is design consistent and asymptotically normal. They also show that it is calibrated for the population totals of the auxiliary variables in both the parametric and nonparametric portions of the model.

Montanari and Ranalli (2005) proposed neural networks as a multivariate smoothing technique for model-assisted estimation. Opsomer et al. (2007) considered the generalized additive model (9) as a multivariate superpopulation model specification, and fitted it by local scoring. One issue with both methods is that they do not lead to estimators calibrated to population totals of the auxiliary variables. In addition, the local scoring estimator for GAM is not a linear function of the y_i , so that the resulting model-assisted estimator cannot be written as a weighted sum, making it difficult to integrate GAMs into the usual survey estimation context. Both Montanari and Ranalli (2005) and Opsomer et al. (2007) applied *model calibration*, originally proposed by Wu and Sitter (2001) as a way to obtain calibrated weighted forms for their estimators. Letting $\hat{\mu}_i$ denote the fits obtained by either

neural network fitting or local scoring, model calibration uses the same expression as the model-assisted estimator \hat{t}_{MA} in (14) but replaces the $\hat{\mu}_i$ by $\hat{\mu}_i^* = \hat{\mu}_i \hat{\beta}$, with $\hat{\beta}$ the estimated coefficient from regressing the y_i on the $\hat{\mu}_i$ using design-weighted least squares regression. The resulting estimator is then calibrated for $\sum_U \hat{\mu}_i$. In Opsomer et al. (2007), the idea of model calibration is further extended by combining the $\hat{\mu}_i$ from the GAM with additional covariates into a multivariate linear model, with the resulting estimator calibrated for all the variables included in that linear model. This estimator can again be written as a weighted sum of the observations (ignoring the fact that the $\hat{\mu}_i$ themselves depend on the y_i).

2.4 Smoothing parameter selection

Nonparametric regression applications require the specification of one or several smoothing parameters, such as the bandwidth in kernel regression or the penalty in spline regression. Selecting the “right” amount of smoothing is a challenging topic in the model-assisted context, further complicated by the fact that in a typical survey application, a single set of survey regression weights is applied to all the survey variables. Because the best smoothing parameter choice depends on the variable being smoothed, no single parameter value (and hence single set of survey weights) will be optimal for all variables in the survey. Nevertheless, it is of interest to have a method to select the amount of smoothing for those cases when precision for a single variable or a small set of them can justify the development of a specifically targeted estimation procedure.

Opsomer and Miller (2005) proposed an automated bandwidth selection method for the LPR estimator that estimates the bandwidth h minimizing the design mean squared error (21). They note that minimizing the traditional estimator of the design mean squared error, $\hat{V}(\hat{t}_{MA})$ in (22), tends to pick bandwidths that are much too small, and instead propose a cross-validation based estimator. The estimator is

the minimizer of

$$\text{CV}(h) = \sum_{i,j \in s} (y_i - \hat{\mu}_i^{(-)})(y_j - \hat{\mu}_j^{(-)}) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}},$$

where $\hat{\mu}_i^{(-)}$ is the local polynomial regression fit computed as in (20), but with observation i removed from the sample. The criterion $\text{CV}(h)$ is a complicated function of h that needs to be evaluated numerically in order to find its minimum. Computations can be greatly simplified because $\hat{\mu}_i^{(-)}$ is easily written as a function of $\hat{\mu}_i$, so that it is only necessary to fit the local polynomial regression once for each value of h . Simulations in Opsomer and Miller (2005) show that the minimizer of $\text{CV}(h)$ is able to successfully adjust the amount of smoothing to the characteristics of the underlying function $\mu(\cdot)$, even for moderate sample sizes. Smoothing parameter selection for the other nonparametric model-assisted estimators described in this section has not been formally studied, but the principle of using a cross-validation based criterion based on the design mean squared error should apply for the PSP estimator as well as the more complicated GAM and neural network cases as well.

3 Nonparametric Methods in Analytic Inference from Surveys

In contrast to descriptive inferences about the current state of a real, finite population, analytic inferences are about model parameters for a hypothetical, infinite population considered to be a generating mechanism for the current state of the finite population. Analytic inference from surveys using nonparametric methods is relatively rare in the literature, and not always cleanly divided from descriptive inference. Both nonparametric density estimation and regression estimation have been used in analytic inferences.

3.1 Nonparametric density estimation

Because no probability density function exists for a finite population, density estimation must be regarded as asymptotic descriptive inference about a limiting sequence of finite populations (that is, the finite population distribution function $F_N(z)$ defined above is assumed to converge to a differentiable function $F(z)$ as $N \rightarrow \infty$), or as analytic inference about an infinite superpopulation. We focus here on the case of analytic inference, in which the goal is to estimate the hypothetical probability density function generating the realized y -values in the finite population.

Bellhouse and Stafford (1999) consider both asymptotic descriptive inference and analytic inference. Using a design-based approach, they compute design-weighted kernel smooths from sample data, as estimates of the corresponding finite population smooths. They also consider a binned version of the problem, with the range of y divided into equally-sized bins, leading to a histogram estimate. Finally, they consider a smoothed version of the histogram. They informally develop asymptotic integrated mean square errors of the various estimators under model and design, and illustrate with data from the Ontario Health Survey.

Breunig (2001) takes a purely model-based approach to density estimation in the context of survey data from a clustered design. This work accounts for the correlation structure induced by the clustering, but ignores all other design features.

Buskirk and Lohr (2005) extend earlier work of Buskirk (1998, 1999) to a thorough exploration of design-weighted kernel density estimation in design-based, model-based, and combined settings. They develop asymptotic theory in these various settings, discuss bandwidth selection and density estimation near boundaries, and apply the methods to data from the US National Crime Victimization Survey and from the US National Health and Nutrition Examination Survey III.

3.2 Nonparametric regression estimation

One particular type of analytic inference for which nonparametric methods are well-suited is exploratory data analysis, in which nonparametric scatterplot smoothers are used to suggest the functional form of the regression relationship between y and \mathbf{x} . Korn and Graubard (1998) use local polynomial regression to smooth survey micro data, accounting for complex design. They do not describe the theoretical properties of this methodology. Bellhouse and Stafford (2001) use local polynomial regression in the same context of exploratory studies. Their goal is to make inference about the infinite population regression function $m(\cdot)$ from (2). They take a design-based approach to this inferential problem, by constructing bins on the x -variable and using the design weights to estimate bin proportions in the finite population and the y -means within bins. If x_i denotes the x -value characterizing the i th bin, with estimated bin proportion \hat{p}_i and bin mean \hat{y}_i , then the function $m(\cdot)$ is estimated using weighted local polynomial regression of \hat{y}_i on x_i , with the usual kernel weights modified by multiplying by \hat{p}_i . The authors approximate the design expectation and variance of the resulting estimator and illustrate with data from the Ontario Health Survey.

Smith and Njenga (1992) take a different approach to incorporating nonparametric regression into analytic inference. They begin by discussing robustness under both design and model-based inference and propose new methods for robust model-based analytic inference by using smoothing techniques. Specifically, they suggest kernel regression of multivariate \mathbf{y} study vectors on covariates \mathbf{x} to estimate conditional mean vectors, followed by kernel regression of the resulting multivariate residuals on \mathbf{x} to estimate conditional covariance matrices. These nonparametric estimates are robust to model misspecification and can be used for analytic inferences about regression coefficients or for multivariate analyses about the relationships among components of the \mathbf{y} vector.

Yet another approach to employing nonparametric regression in analytic infer-

ence is proposed by Chambers and Skinner (2003). This approach builds on the corresponding parametric approach developed in Pfeffermann and Sverchkov (1999). The idea of that paper is to avoid the potential bias caused by nonignorable sampling designs by using the sample distribution of the sample measurements in maximum likelihood estimation. This sample distribution is related to the conditional distribution of the study variable and the conditional distribution and conditional mean of the sample selection probabilities. These quantities are estimated parametrically in Pfeffermann and Sverchkov (1999) and nonparametrically in Chambers and Skinner (2003) under various data scenarios. The simplest of these scenarios, for example, leads to an estimator that uses a design-weighted version of the Nadaraya-Watson estimator.

The demand for flexible, robust procedures in all aspects of inference from complex surveys suggests that nonparametric methods have great potential in this area. The three approaches described in this section have tapped some of that potential, but it is clear that much further work remains to be done, and this should be a fruitful area of future research.

4 Nonparametric Methods in Nonresponse Adjustment

In Section 2, nonparametric methods were used to improve the efficiency of survey estimators, by taking advantage of the relationship between auxiliary variables available for the population and the survey variables. In this section, we describe how nonparametric methods can also be used to adjust survey estimators for the presence of nonresponse, when the response mechanism is related to an auxiliary variable available for the original sample. We focus here on the case of *unit nonresponse*.

Nonresponse is pervasive in surveys and can induce bias if it is not properly accounted for. In the context of unit nonresponse, the most commonly used approach

is to adjust the weights of the responding observations by incorporating estimates of the probabilities that the units are respondents. This can be done implicitly, as in the *weighting cell* estimator, or explicitly, by specifying and fitting a response probability function and obtaining new weights. In both cases, the response process can be viewed here as a second *phase* of sampling, with unknown probability mechanism. This nonresponse phase follows the first phase of sampling, which is determined by the original sampling design. Särndal and Swensson (1987) formally describe the two-phase framework for nonresponse and the types of approaches that can be used to adjust survey estimators for nonresponse.

Suppose that we have a sampling design $p(\cdot)$ with corresponding inclusion probabilities $\pi_i, i \in U$, and that, in the absence of nonresponse, we were planning to estimate the population total t_y by the Horvitz-Thompson estimator in (10) based on the sample s . Because of nonresponse, we only observe $r \subseteq s$. Since the random process generating r is typically unknown, we need to assume a model for the response mechanism. Let $R_i = 1$ if $i \in r$ and 0 otherwise. As is often done in the nonresponse modeling context, we will assume that the R_i are independent Bernoulli random variables with

$$\Pr\{R_i = 1\} = \phi_i, \quad 0 < \phi_i \leq 1, \quad \forall i \in U. \quad (24)$$

See also Chapter 8.

For the weighting cell estimator, the population is divided into G cells, $U = \cup_{g=1}^G U_g$, and in each cell the (average) response probability is estimated by the fraction of the sampled respondents in cell g . This fraction is usually computed as $\sum_{r_g} w_i / \sum_{s_g} w_i$, where $s_g = s \cap U_g, r_g = r \cap U_g$, with either $w_i = 1/\pi_i$ or $w_i = 1$. In what follows, we will only consider the former case. The weighting cell estimator for t_y is defined as

$$\hat{t}_{WC} = \sum_{g=1}^G \left(\frac{\sum_{s_g} w_i}{\sum_{r_g} w_i} \right) \sum_{r_g} w_i y_i. \quad (25)$$

From this expression, it is easy to see that in each cell, the estimator of the cell total is ratio-adjusted by the inverse of the weighted proportion of respondents in

the cell.

A number of authors have studied the properties of the weighting cell estimator, including Oh and Scheuren (1983), Särndal et al. (1992, p. 578) (using the term “response homogeneity group” for the cells) and Kim and Fuller (1999). A common assumption in the study of the design-based properties of \hat{t}_{WC} is that the cells are correctly specified, in the sense that they correspond to well-defined and known population groups in which the response indicators R_i are independent and identically distributed with $\Pr\{R_i = 1\} = \phi_g, i \in U_g$. While these authors showed that \hat{t}_{WC} is consistent under this assumption, it was not clear what happens when the cells are not correctly specified.

Da Silva and Opsomer (2004) investigate the theoretical behavior of \hat{t}_{WC} using nonparametric methodology. Instead of assuming that the cells correspond to known response categories in the population, they consider the situation in which the response probability $\phi_i = \phi(x_i)$, with x_i an auxiliary variable observed for all elements in the sample and $\phi(\cdot)$ an unknown smooth function. The weighting cells are formed by sorting the sample on the x_i and dividing the range of x_i into G groups. Under this scenario, the grouping can be thought of as a very simple form of smoothing, with the unknown function $\phi(\cdot)$ approximated by a piecewise constant fit, and G a “smoothing parameter” similar to the bandwidth h or the penalty λ in the previous sections. Da Silva and Opsomer (2004) prove that \hat{t}_{WC} is a consistent estimator of t_y under quasi-randomization (i.e. the combination of the sampling design and the response mechanism) under mild conditions, provided that G is allowed to increase as the sample size increases. Unlike the previous authors studying the weighting cell estimator, they do not require the cells to be correctly specified.

While the weighting cell estimator can be considered a simple nonparametric estimator, it is possible to construct nonresponse adjusted estimators that incorporate nonparametric regression methods more fully. This type of estimator will be based on explicit estimation of the unknown response probability function $\phi(\cdot)$, and the ideas of two-phase estimation. Starting again from the Horvitz-Thompson estimator

in (10), suppose the response probability function $\phi(\cdot)$ were known. The two-phase estimator

$$\hat{t}_\phi = \sum_r \frac{y_i}{\pi_i \phi(x_i)} \quad (26)$$

is unbiased and consistent under quasi-randomization. This estimator is unfeasible, so that it is replaced by

$$\hat{t}_{\hat{\phi}} = \sum_r \frac{y_i}{\pi_i \hat{\phi}_i}, \quad (27)$$

with $\hat{\phi}_i$ an estimate of $\phi(x_i)$.

A number of authors have considered parametric specifications for $\phi(\cdot)$, including most recently Kim and Kim (2007). We discuss the nonparametric case here. The use of kernel-type smoothing methods in the nonresponse context was first proposed by Giommi (1984, 1987), and further discussed by Niyonsenga (1994, 1997). Neither of these authors provided formal theoretical results on their nonparametric estimators. Recently, Da Silva and Opsomer (2006) studied the properties of the estimator in (27) with the response probability function $\phi(\cdot)$ estimated by a sample-weighted kernel regression estimator of the response indicators. The estimator is a special case of the estimator in (20), with \mathbf{y}_s replaced by the vector of response indicators R_i in the sample and the degree of the local polynomial $q = 0$, i.e. the local design matrix in (17) replaced by a vector of ones. The resulting estimator can be written as

$$\hat{\phi}_i = \left(\sum_{j \in s} K \left(\frac{x_j - x_i}{h} \right) \frac{1}{\pi_j} \right)^{-1} \sum_{j \in s} K \left(\frac{x_j - x_i}{h} \right) \frac{R_j}{\pi_j}. \quad (28)$$

The results of Da Silva and Opsomer (2006) for the estimator (27) with $\phi(x_i)$ estimated by (28) show that the nonparametric nonresponse adjusted estimator is quasi-randomization consistent for t_y under mild conditions. They also found that $\hat{t}_{\hat{\phi}}$ does not have the same asymptotic distribution as \hat{t}_ϕ , but that the estimation of the response probability function $\phi(\cdot)$ contributes additional terms in the asymptotic approximation. This implies that if the estimated response function is treated as if it were known for the purpose of inference, it is likely that the variance will be

incorrectly estimated. Kim and Kim (2007) found a similar result in the parametric case.

5 Nonparametric Methods in Small Area Estimation

As a final application of nonparametric methods in the survey context, we discuss applications in small area estimation. Cowling et al. (1996) present two applications of spatial smoothing in a small area estimation context. The first use of smoothing is in making small area estimates less variable. The procedure rids the original weights to allow for deviations from benchmark totals, then the modified weights are spatially smoothed via a kernel over geographic neighborhoods to get less spatial variability in the weights. The result is more stable small area estimates. The second application uses design-weighted kernel smooths to get maps of estimates of the characteristic over a spatial domain. No properties are derived for either of these methodologies.

In two recent developments, nonparametric methods are brought directly into classical methods for small area estimation. Mukhopadhyay and Maiti (2004) propose an extension of the area-level model in which the linear mean function is replaced by a nonparametric function to be estimated by kernel regression, while Opsomer et al. (2007) consider an element-level model and use penalized spline regression.

Suppose the population contains T small areas of interest, indexed by t . The nonparametric area-level model studied by Mukhopadhyay and Maiti (2004) is

$$y_t = m(x_t) + u_t + \varepsilon_t, \quad (29)$$

where u_t and ε_t are distributed independently as $\mathcal{N}(0, \sigma_u^2)$ and $\mathcal{N}(0, D_t)$ with D_t known. If $m(\cdot)$ is a linear function, this model is usually called the Fay-Herriot model (Fay and Herriot, 1979). The purpose of small area estimation methods for model (29) is to predict $\tilde{y}_t = m(x_t) + u_t$, and in the linear model case, empirical best

linear prediction (EBLUP) methods or hierarchical Bayesian methods are typically used. The prediction procedure starts by estimating $m(\cdot)$ by the local polynomial regression estimator (3) with $q = 0$, so that the matrix \mathbf{X}_x in (4) is replaced by a vector of ones (as was done in Section 4). The small area variance σ_u^2 is estimated by $\hat{\sigma}_u^2 = \sum_{t=1}^T \{(y_t - \widehat{m}(x_t))^2 - D_t\}/T$, possibly adjusted to ensure non-negativity, and the predictor for \tilde{y}_t is defined in analogy to the EBLUP as

$$\hat{y}_t = \hat{\gamma}_t y_t + (1 - \hat{\gamma}_t) \widehat{m}(x_t) \quad (30)$$

with $\hat{\gamma}_t = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + D_t)$. Mukhopadhyay and Maiti (2004) derive an asymptotic approximation to the prediction mean squared error of \hat{y}_t , $E(\hat{y}_t - \tilde{y}_t)^2$, and a plug-in estimator for that quantity.

Because of close connections between EBLUP and penalized spline regression (see Wand, 2003), penalized splines provide a convenient approach for integrating nonparametric models into small area estimation. Opsomer et al. (2007) extend the linear element-level mixed model small area estimation approach described in Battese et al. (1988) to the setting in which the mean function can be nonparametrically (or semiparametrically) specified. The model is

$$y_i = m(x_i) + \mathbf{d}'_i \mathbf{u} + \varepsilon_i, \quad (31)$$

where $\mathbf{d}_i = (d_{1i}, \dots, d_{Ti})'$ is a vector of indicators with $d_{ti} = 1$ if element i is in the small area t and zero otherwise, $\mathbf{u} = (u_1, \dots, u_T)'$ is a vector of mutually independent small area effects with mean 0 and variance σ_u^2 , and ε_i is the random error with mean 0 and variance σ_ε^2 , independent of \mathbf{u} . The nonparametric function $m(\cdot)$ is expressed as a spline function as in (5). Following Wand (2003), we rewrite this as $m(x_i) = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\gamma}$, with $\mathbf{x}_i = (1, x_i, \dots, x_i^p)'$, $\mathbf{z}_i = ((x_i - \kappa_1)_+^p, \dots, (x_i - \kappa_J)_+^p)'$, $\boldsymbol{\beta}$ a vector of unknown parameters, and $\boldsymbol{\gamma}$ a vector of independent random variables with mean 0 and variance σ_γ^2 . The full model is therefore

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \boldsymbol{\gamma} + \mathbf{d}'_i \mathbf{u} + \varepsilon_i. \quad (32)$$

The term $\mathbf{z}'_i\boldsymbol{\gamma}$ is a random deviation from the fixed linear trend in the population, and $\mathbf{d}'_i\mathbf{u}$ is the random effect for small area i . The goal of the small area estimation is now the prediction of $\tilde{y}_t = \bar{\mathbf{x}}'_t\boldsymbol{\beta} + \bar{\mathbf{z}}'_t\boldsymbol{\gamma} + u_t$, where we assume that $\bar{\mathbf{x}}_t, \bar{\mathbf{z}}_t$ are known.

The critical point of the formulation (32) is that we have once again expressed the model as a linear element-level mixed effect model, so that the full range of EBLUP methods can be applied. Opsomer et al. (2007) propose restricted maximum likelihood estimation (REML) to estimate the parameters $\boldsymbol{\beta}, \sigma_\gamma^2, \sigma_u^2, \sigma_\varepsilon^2$, and predict \tilde{y}_t by

$$\hat{y}_t = \bar{\mathbf{x}}'_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{z}}'_t\hat{\boldsymbol{\gamma}} + \hat{u}_t, \quad (33)$$

with

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{Y} \\ \hat{\boldsymbol{\gamma}} &= \hat{\sigma}_\gamma^2\mathbf{Z}'\widehat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \hat{\mathbf{u}} &= \hat{\sigma}_u^2\mathbf{D}'\widehat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{aligned}$$

where $\mathbf{X} = (\mathbf{x}_1 \dots, \mathbf{x}_n)'$ and \mathbf{Y}, \mathbf{Z} and \mathbf{D} are defined analogously. Further, $\widehat{\mathbf{V}}$ is the estimated variance-covariance matrix of \mathbf{Y} obtained by plugging the REML estimates of the variance parameters in the variance-covariance matrix of \mathbf{Y} . The asymptotic approximation to the prediction mean squared error of \hat{y}_t is shown to directly generalize that obtained in the absence of the spline random effect, and a bias-corrected estimator for the prediction mean squared error is provided. Opsomer et al. (2007) also discuss likelihood ratio testing for the variances of the random effects and propose a simple nonparametric bootstrap for inference.

References

- Aerts, M., G. Claeskens, and M. Wand (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference* 103, 455–470.

- Aragon, Y., C. Goga, and A. Ruiz-Gazen (2006). Estimation non-paramétrique de quantiles en présence d'information auxiliaire. In P. Lavellée and L.-P. Rivest (Eds.), *Méthodes d'Enquêtes et Sondages. Pratiques Européenne et Nord-américaine*, pp. 377–382. Dunod, Paris-Sciences Sup.
- Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83, 28–36.
- Bellhouse, D. R. and J. E. Stafford (1999). Density estimation from complex surveys. *Statistica Sinica* 9, 407–424.
- Bellhouse, D. R. and J. E. Stafford (2001). Local polynomial regression in complex surveys. *Survey Methodology* 27(2), 197–203.
- Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* 92(4), 831–846.
- Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics* 28, 1026–1053.
- Breidt, F. J., J. D. Opsomer, A. A. Johnson, and M. G. Ranalli (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology* 33, 35–44.
- Breiman, L., J. H. Friedman, R. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- Breunig, R. V. (2001). Density estimation for clustered data. *Econometric Reviews* 20(3), 353–367.
- Buskirk, T. D. (1998). Nonparametric density estimation using complex survey data. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 799–801. American Statistical Association.
- Buskirk, T. D. (1999). *Using nonparametric methods for density estimation with complex survey data*. Ph. D. thesis, Department of Mathematics, Arizona State

University.

- Buskirk, T. D. and S. L. Lohr (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference* 128, 165–190.
- Chambers, R. L., A. H. Dorfman, and T. E. Wehrly (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* 88, 268–277.
- Chambers, R. L. and C. J. Skinner (Eds.) (2003). *Analysis of Survey Data*. Chichester, U. K.: John Wiley & Sons.
- Cowling, A., R. Chambers, R. Lindsay, and B. Parameswaran (1996). Applications of spatial smoothing to survey data. *Survey Methodology* 22, 175–183.
- Da Silva, D. N. and J. D. Opsomer (2004). Properties of the weighting cell estimator under a nonparametric response mechanism. *Survey Methodology* 30, 45–55.
- Da Silva, D. N. and J. D. Opsomer (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *Canadian Journal of Statistics* 34, 563–579.
- de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). New York: Springer-Verlag.
- Deville, J.-C. and C. Goga (2004). Estimation par régression par polynômes locaux dans des enquêtes sur plusieurs échantillons. In P. Ardilly (Ed.), *Echantillonnage et Méthodes d’Enquêtes*, pp. 156–162. Dunod, Paris-Sciences Sup.
- Dorfman, A. H. (1992). Non-parametric regression for estimating totals in finite populations. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 622–625. American Statistical Association (Alexandria, VA).
- Fay, R. E. and R. A. Herriot (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American*

- Statistical Association* 74, 269–277.
- Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* 19, 1–141.
- Giommi, A. (1984). On the estimation of the probability of response in finite population sampling (Italian). *Societa Italiana di Statistica, Atti della Riunione Scientifica della Societa Italiana* 32(1), 275–284.
- Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology* 13, 127–134.
- Goga, C. (2004). Estimation de l'évolution d'un total en présence d'information auxiliaire : une approche par splines de régression. *Comptes Rendus de l'Académie des Sciences Paris Ser. I* 339, 441–444.
- Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche non paramétrique par splines de régression. *Canadian Journal of Statistics* 33, 163–180.
- Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Washington, D. C.: Chapman and Hall.
- Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663–685.
- Kim, J.-K. and W. A. Fuller (1999). Jackknife variance estimation after hot deck imputation. In *1999 Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 825–830. American Statistical Association.
- Kim, J. K. and J. J. Kim (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics* 35, To appear.
- Korn, E. L. and B. I. Graubard (1998). Scatterplots with survey data. *The American Statistician* 52, 58–69.

- Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 280–285. American Statistical Association (Alexandria, VA).
- Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* 100(472), 1429–1442.
- Mukhopadhyay, P. and T. Maiti (2004). Two-stage nonparametric approach for small area estimation. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 4058–4065. American Statistical Association.
- Niyonsenga, T. (1994). Nonparametric estimation of response probabilities in sampling theory. *Survey Methodology* 20, 177–184.
- Niyonsenga, T. (1997). Response probability estimation. *Journal of Statistical Planning and Inference* 59, 111–126.
- Oh, H. L. and F. J. Scheuren (1983). Weighting adjustments for unit non-response. In W. G. Madow, I. Olkin, and D. B. Rubin (Eds.), *Incomplete data in sample surveys (Vol. 2): Theory and bibliographies*, pp. 143–184. Academic Press (New York; London).
- Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007). Model-assisted estimation of forest resources with generalized additive models (with discussion). *Journal of the American Statistical Association* 102, 400–416.
- Opsomer, J. D. and C. P. Miller (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Journal of Nonparametric Statistics* 17, 593–611.
- Pfeffermann, D. and M. Sverchkov (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B* 61, 166–186.

- Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
- Särndal, C. E. and B. Swensson (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* 55, 279–294.
- Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Smith, T. M. F. and E. Njenga (1992). Robust model-based methods for analytic surveys. *Survey Methodology* 18, 187–208.
- Vidaković, B. (1999). *Statistical Modeling by Wavelets*. New York: John Wiley & Sons, Inc.
- Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics* 18, 223–249.
- Wand, M. P., B. A. Coull, J. L. French, B. Ganguli, E. E. Kammann, J. Staudenmayer, and A. Zanobetti (2005). *SemiPar 1.0. R package*. <http://cran.r-project.org>.
- Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall.
- Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* 96, 185–193.
- Zheng, H. and R. J. A. Little (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics* 19, 99–117.

Zheng, H. and R. J. A. Little (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology* 30, 209–218.