REWEIGHTING SURVEY OBSERVATIONS
BY MONTE CARLO INTEGRATION ON A CENSUS

by

M.A. Keyzer
Content

Abstract iii

1. Introduction 1
2. Mean estimation by integration over a regression curve 2
3. Determining weights by kernel density regression 4
4. Effectiveness of the procedure: a simulation exercise 10
5. Conclusion 11

Appendix: Computing the weights 13

References 15
Abstract

The possible non-representativity of household surveys can be addressed by reweighting the survey observations through a correction factor. This factor is usually computed on the basis of the frequency of a combination of household characteristics, say, the age and sex of the respondent, in the survey relative to its value in a population census. In this paper, a generalization of this technique is proposed that makes it possible to account for several household characteristics including real-valued ones. It applies kernel density regression to estimate the joint density over the survey of selected characteristics that are also recorded in the census. The weights are estimated by Monte Carlo integration of the estimated density over the census distribution. The efficacy of the procedure is tested in a simulation experiment that creates a large number of survey data sets as biased samples from the Ghana Living Standards Survey, applies the reweighting procedure to every sample, and confronts the resulting estimated mean with the known true mean of the census.

Keywords: Monte Carlo integration, kernel density regression, weighting scheme
1. **Introduction**

In many countries national development policies increasingly rely on household survey information (e.g. Grosh and Munoz, 1996; Deaton, 1997), for several reasons. First, policy analysts often need timely information on, say, nutritional status, which can not be collected within a short period for a large number of households. Second, to trace social change, panel surveys that follow individuals over time play an important role. Third, under rapid economic change, even the construction of macro-indicators such as GDP requires more frequent collection of underlying distributions than before. Finally, rather than inferring perceptions and expectations from observed behavior econometric analysis increasingly relies on interviews to measure these variables directly.

As full enumeration is generally too costly, especially in poor developing countries, it has become common practice to rely on household surveys of limited coverage and specialized scope. This has led to large number of surveys being conducted in parallel, with reduced attention for sample design and pilot testing due to scarcity of funds, lack of information and time pressure and consequently representativity suffers. Moreover, since representativity is not independent of the questions asked, and surveys seldom repeat earlier exercises in full, it can rarely be established accurately beforehand. Therefore, unless the survey project can proceed in stages, gradually expanding and redefining its sample, the problem of representativity cannot be dealt with satisfactorily beforehand, and corrections have to be made after the (current round of) data collection has been completed, by confronting these data to other information. Whereas a vast and advanced literature has been devoted to the design of procedures to address the selectivity bias at the stage of model estimation (Heckman et al., 1998; Manski, 1995), far less attention was paid to correction of the selectivity bias on the data themselves. And yet the issue is important, as a good correction method would allow to keep samples smaller and more focused, and to deepen the scope of the questionnaires. It would also facilitate the integration of findings from different surveys into a coherent policy framework and thus contribute to improved policy co-ordination.

The common approach to perform a representativity correction on the survey is to relate it to a large-scale survey, that we will refer to as a census even though it will seldom be exhaustive. Class-specific inflation factors as *weights* are determined on the basis of the occurrence in the census of various combination of characteristics, say, of female-headed rural households relative to the survey (Deaton, 1997, p. 44). A higher weight will be attributed to respondents whose combination of characteristics has a high relative rate of incidence in the census. The use of class-specific inflation factors has the obvious limitation that some of the variables shared among the census and the survey might not be qualitative. The age of the respondent, the surface of land owned, or the housing space occupied have cardinal measurement, and for such variables any aggregation into discrete classes is arbitrary and necessarily implies a loss of information. In addition, as the number of shared properties grows larger, say, more than three, the evaluation and storage of the discrete frequency distribution of the census becomes very cumbersome. And when the number of, possibly empty, classes grows very large it becomes more difficult to assign a class to given survey observations.

In this paper, we propose a general technique for calculating weights on the basis of shared variables of arbitrary dimension, which may be qualitative as well as quantitative (i.e. discrete or real-valued). We show that kernel density regression leads to weights that are independent of the unshared variables, i.e. of the answers specific to the survey and can be computed for every survey.

---

1 The author thanks Peter Albersen, Bart van den Boom, Chris Elbers, Martin Ravallion, Tugrul Temel and Wim van Veen for comments on an earlier draft.
respondent in the survey by (Monte Carlo) integration of a kernel estimate of the survey density over the census. The estimated weight is statistically consistent and converges to unity. It can also be used for the design of the sample in a next round of the survey. Generally, the approach would suggest collecting additional data at points where the unexplained variation (the error in regression) is large, rather than at points that deviate strongly from the mean but lie close to the regression line.

The paper proceeds as follows. The procedures for computing means and variances are described sections 2 and 3, respectively. In section 4, we study their efficacy by means of numerical experiments in which survey data sets are created as biased samples from a given census. We evaluate how well the weighted mean can redress the sample bias. Section 5 concludes, and the appendix gives guidelines for using the associated computer program.

2. Mean estimation by integration over a regression curve

Survey exercises usually follow a fixed sequence of tasks starting from questionnaire design, followed by sample design, tests, and revision of the design, data processing, tabulations, possibly modeling exercises, and finally reporting. Most rely on external statistics for the sample design, but thereafter proceed in isolation. The common, often implicit, justification is that once a survey can be considered to be representative, its tabulation only needs to print arithmetic averages over households, and its statistical analysis can treat the survey answers as a self-contained data set. As an introduction we briefly review the basic steps of sample design. The survey exercise starts with the formulation of a questionnaire and the demarcation of the population of intended respondents (owners of a car, farmers, voters, etc. in a particular geographical area). Next, a sampling frame has to be defined, consisting of a preferably complete list of potential respondents in that area. Given this frame, one determines the size of the survey and draws a random sample of this size from this list by a lottery. In the simplest case the first sample drawn defines the sample. Alternatively, a stratified procedure is applied. Respondents are characterized by several properties, available on the list, and a sample of given size is collected for every characteristic. First, of all the car owners a sample is taken. Next in the sample one chooses a subset with, say, cars of a particular age, then of a particular color, etc. Statistical theory only enters when it comes to the determination of the sample size in the various strata.

The aim of a survey is generally to obtain a ‘reliable’ estimate of the population mean by taking the arithmetic average from the random sample. Formally, the model of the mean is:

\[ \mu = \int y g(y) dy, \]  

(2.1)

where \( g(y) \) is the unknown density of \( y \). If the sample is of size \( S \), the population mean can be estimated as the sample average and this estimate is unbiased since the sample is random, while consistency is ensured by the Law of Large Numbers. Moreover, the Central Limit Theorem implies that if \( g(y) \) has variance \( \sigma^2 \), the sample average will be have an asymptotically normal distribution with mean \( \mu \) and variance \( \sigma^2/S \). Hence, the mean variance falls with sample size, and if the sampling cost is low and the loss from wrong estimation of the mean is high the sample size will be chosen higher. The common procedure is to conduct a preparatory survey of size \( S_0 \) that yields an estimate \( \mu_0 \) of the mean and \( \sigma_0 \) of the standard deviation, while the ratio \( \mu_0/\sigma_0 \) is asymptotically t-distributed. This allows to compute the sample size \( S \) which guarantees that the sample mean has a given probability \( \alpha \) of lying within given a percentage range around the true mean. Generally, the size of the sample can be chosen so as to balance the reduction in risk
(expected loss from variance of the mean) with the cost of additional observations. However, besides the fact that the distribution of the ratio is actually unknown for small samples, the approach also neglects all a priori information on the density \( g(y) \).

To include such information, sample design often proceeds by subdividing the population into separate strata, say, age cohorts, each with their own variance. For given sample size \( S \), the variance of the mean is then minimized (approximately) if the sample is subdivided in relation to both size and heterogeneity (Fisz, 1970, p. 602):

\[
S_i = S \frac{N_i \sigma_i}{\sum_i N_i \sigma_i}.
\]  

Yet, this refinement also neglects much of the a priori information. In fact, the variance in observations of individual variables is hardly indicative of the risk. A survey contains generally many questions, the answers to which are not independent. Therefore, \( y \) will usually be a vector and the covariance among the elements of this vector has to be accounted for when the optimal sample size is determined, which is generally unknown. Furthermore, it seems unrealistic to assume that the data generating process that produces the answers is an unconditioned random sample. It is presumably better described as a conditional process, i.e. as a deterministic model with an error term, that associates household characteristics (the conditioning variables \( x \)) with answers to specific questions \( y \).

In this case one has to account explicitly for these conditioning variables and their distribution, and this suggests to work with a regression model rather than with sample means. For example, suppose that we have conducted a survey among the farm households in a particular country and that we collected observations \( \{y^s, x^s\} \), for a sample \( \{1, ..., S\} \) of households, where the scalar \( y \) is the main dependent variable (say, income) and the vector \( x \) refers to household characteristics (age, schooling etc.). Assume also that we know from a population census that the explanatory variables have a known continuous joint density \( f(x) \) with compact support \( X \), and we want to compute the average income at census level.

Computing the arithmetic mean over \( y^s \) amounts to accepting the representativity of the empirical distribution of the sample and neglects the a priori information that \( y \) is a function of \( x \) and that the density of \( x \) is known. If an exact function \( y(x) \) was available, the mean could be calculated by averaging the function over its domain.\(^2\)

\[
\mu = \int y(x) f(x) \, dx.
\]  

This would make it possible to redress imbalances in sample design afterwards, at the stage of survey tabulation. The problem would purely be one of computation. However, there are two problems. One is that in general no such function is available, and the other that the integral in (2.3) has to be coped with numerically.

\(^2\) We limit attention to calculation of the population mean, i.e. to integration over all elements of \( x \). In geographical applications \( x \) could be partitioned as \( x = (x_1, x_2, x_3) \) where the first to elements denote latitude and longitude. Integration over \( x_3 \) only will generate a map of means \( \mu(x_1, x_2) \). For applications and visualisation of kernel density regressions in a geographical context see Sonneveld and Keyzer, 1997.
3. Determining weights by kernel density regression

Consider the deterministic function \( y_\theta(x, \varepsilon) \), where \( x \) is a given vector of explanatory variables and \( \varepsilon \) a random variable and \( \theta \) a vector of parameters. A regression function is the conditional expectation of this function.

\[
y_\theta(x) = E[y_\theta(x, \varepsilon) \mid x].
\] (3.1)

Parametric regression techniques usually postulate an additive error with zero conditional expectation \( E[\varepsilon \mid x] = 0 \), hence:

\[
y_\theta(x, \varepsilon) = y_\theta(x) + \varepsilon.
\] (3.2)

In contrast, non-parametric kernel density regression (e.g. Haerdle, 1991) does not postulate any given functional form and parameters for the function and but it makes further assumptions on the density of \( \varepsilon \), and assumes an error in independent variables rather than an error in \( y \). Hence the expectation (3.1) becomes:

\[
y_\theta(x) = \int \psi(\varepsilon) d\varepsilon.
\] (3.3)

Here we develop the non-parametric formulation. For the kernel density we choose

\[
\psi_\theta(\varepsilon) = \psi(\varepsilon / \theta),
\] (3.4)

to be a multivariate normal density with window size \( \theta \) (the elements \( \varepsilon_i \) are independently distributed with standard deviation \( \sigma_i / \theta \)). Given the finite sample of the survey, the function \( y_\theta(x) \) is approximated as the probability weighted mean:

\[
\hat{y}_\theta(x) = \sum P_\theta'(x) y',
\] (3.5)

for given probability weights

\[
P_\theta'(x) = \frac{\psi_\theta(x') - x}{\sum \psi_\theta(x' - x)},
\] (3.6)

and window size \( \theta \) depending on the number of observations and the number of variables only. This yields a consistent, though possibly biased estimate, whose variance vanishes as the number of observations goes to infinity and the window size goes to zero.\(^3\) Approximation (3.5) can be used in (2.3) and leads to:

\[
\hat{\mu} = \int \hat{y}_\theta(x) f(x) dx.
\] (3.7)

\(^3\) To check the consistency, divide numerator and denominator of (3.6) by \( S \). The denominator goes to unity for \( S \) going to infinity, and for the remaining terms the limit property follows via the law of large numbers. The bias is essentially due to the fact that the probability is actually the ratio of two estimates and that the window size falls as \( S \) goes to infinity.
Clearly, this form would also apply to parametric regression but the kernel density specification yields a further simplification:

\[ \hat{\mu} = \frac{1}{S} \sum_{i=1}^{S} y^i w^i, \]  

for

\[ w^i = \int P_{\theta}^i(x) f(x) \, dx. \]  

(3.8)  

(3.9)

Hence, the weights \( w^i \) of (3.9) enable us to recalculate the means and possibly higher moments for every variable \( y \) of the survey, as in (3.8). It is important to note that in (3.9) the weights are independent of \( y^i \). This circumvents the difficulty that weights might differ across regression functions, and that a poor fit for some of the regressions would affect them. However, we also notice that the kernel density method implies that the estimate of the mean will always be a convex combination of the observations \( y^i \) and is therefore unable to compute extrapolated \( y \)-values, unlike parametric regression.

**Interpreting the weights**

Monte Carlo simulation is a natural technique for computing, or actually estimating the integral in (3.9), and will be applied in section 4 below. Here we seek to interpret expression (3.9). We compare the densities in the survey and the sample. For this, we define the estimated density function of the sample:

\[ \varphi(x) = \frac{1}{S} \sum_{s=1}^{S} \psi_{\theta} (x^s - x), \]  

(3.10)

and rewrite (3.8) as

\[ w^i = \int \psi_{\theta} (x^i - x) \frac{f(x)}{\varphi(x)} \, dx. \]  

(3.11)

If the ratio of “actual census” to “estimated sample survey” densities is equal to unity for all \( x \), i.e. if the survey is fully representative of the census, then the weight will be equal to unity. The density ratio has the natural representation of a correction factor that increases the emphasis on \( s \) if \( x \)-values close to \( x^i \) are being under-represented in the survey. Conversely, \( x \)-values from the census that lie far away from the range of the survey will hardly affect the weights.

It may also be useful to derive a measure of the fit between the census density and the survey density, say, the standard deviation:

\[ \sigma^2 = \frac{1}{S} \sum_{i} (w^i - 1)^2. \]  

(3.12)

Using this measure, it is possible to test whether on average the weights differ significantly from unity and by it whether the reweighting could be dispensed of. We notice, however, that the measure actually checks whether the density in the numerator of (3.6) is more or less uniform across respondents, which is insufficient to ensure a close correspondence between survey and census. For example, if all \( x^i \) coincide, the low variance only signifies that there is no information
available to effectuate the correction. At the same time, whenever the $x^s$ show variation, a low variance will usually indicate close correspondence.

**The case of discrete variables**

We need to consider the case where $x$ partitions into quantitative (real-valued) and qualitative (discrete) elements: $x = (x_1, x_2)$. So far, we have assumed continuity of the census density. In case of discrete values, the true density is a spike i.e. has infinite value and zero width. Yet to deal with this case it is practical to assume a finite value and positive width. This is actually what the kernel density does since it treats real and integer values alike and attributes a bounded density to any observation, as long as the window size is positive. Indeed, if in calculations an explicit form of the density is used it will be practical to choose it with a positive window size and a positive density (e.g. the normal density) so as to avoid infinite ratios and divisions by zero in (3.6).

Clearly, in case all variables are discrete it is in principle possible to avoid regression altogether and return to the usual weights mentioned in the introduction. Suppose that the variable $x$ is a scalar, classified in discrete categories $j = 1, \ldots, J$. Then, $F_j$ is the given incidence of $j$ in the census and the weight can now be computed as:

$$w^s = \sum_j \psi_j \frac{F_j}{\varphi_j},$$

with

$$\psi_j = \delta^s_j, \quad \varphi_j = \frac{S_s}{S}, \quad S_j = \sum \delta^s_j, \quad S = \sum_j S_j,$$

for the relative incidence in the survey, and the fraction in the survey, respectively, where $\delta$ is equal to unity if respondent $s$ has $x$-value falling in class $j$, and zero otherwise. This illustrates how (3.11) reduces to (3.13) in case of discrete variables. At the same time it also shows that we can always approximate (3.13) by (3.11), replacing a spike by a kernel density, and that for a small enough window size $\theta$, any desired precision can be achieved. Nonetheless, the continuous representation offers the additional advantage that all ranges of $x$ in the census will affect the estimate, whereas the discrete assignment will only account for classes $j$ that have some $s$ in the survey associated to them.

**Computation by Monte Carlo integration**

It would be possible to calculate the weights in (3.9) by numerical integration but this is cumbersome. The next step is to find a numerically efficient approximation and we propose to apply Monte Carlo integration. Various alternatives present themselves, the practicality of which depends on how the census data are made available, and on the size of the census. We specify procedures that correspond to three typical cases: (i) the raw census data are available and the census is very large (say, over 1 million households); (ii) the raw data are available and the census is not very large (say, less than 50000 households); (iii) census information is available as a joint density only. In the third case the procedure corresponding to (i) does not apply but since it is always possible to estimate some density on the basis of raw data, all three procedures could be used if such data are available.
(i) **Sampling from the raw census data**

Let \( x^c \) denote an observation vector from the census, \( c = 1, 2, \ldots, C \). This vector may contain real as well as discrete elements. We choose a fixed number of iterations \( T \), and circumvent the need for an explicit density function by performing random sampling over the empirical distribution, i.e. over the raw census data themselves.

1. Start from iteration \( t=1 \) and \( w_t' = 1 \).
2. Calculate the step size \( \rho_t = 1/t \).
3. Draw a random number \( c \) from the list \( \{1,C\} \).
4. Evaluate for all \( s = 1, \ldots, S \) the kernel densities \( \xi_s^t = \psi_{\rho_t}(x^s - x^c) \).
5. Evaluate the mean sample density (jointly with (i.4)) \( \phi_t = 1/S \sum_{s=1}^{S} \xi_s^t \).
6. Update the estimate of the mean \( w_{t+1}' = w_t' + \rho_t \frac{\xi_t^t}{\phi_t} \).
7. If \( t \leq T \) increment \( t \) by 1 and go to (i.2).

For \( T \) sufficiently large this procedure converges to (3.11). We note that step (2) can be modified. Convergence is maintained as long as for \( t \to \infty \), the step size \( \rho_t \) satisfies (see Ermoliev and Wets, 1988):

\[
0 < \sum_{t=1}^\infty \rho_t = \infty, \sum_{t=1}^\infty \rho_t^2 < \infty.
\]

(ii) **Single pass over all raw census data**

The procedure is as (i.1-7) above, except that \( T = C \) while step (i.3) is replaced by

3. Assign \( c = t \).

(iii). **Census density \( f(x) \) is given**

In case the joint density \( f(x) \) is given, it is always possible to generate an artificial data set and apply procedure (i) or (ii) to it. However, when the elements of \( x \) are correlated this may be a difficult task. Therefore, we also present the algorithm that will approximate the mean on the basis of the census density rather than the data. We take the census density to be either a continuous density or of the form (3.13) in which case the sampling takes place in two steps: first a drawing from the discrete probability and then one from the continuous density. For drawing from the continuous density, two cases should be distinguished. In the first case there is software available to draw a random sample from the density \( f(x) \) directly (see for example Marsaglia and Zaman, 1991; Chandler and Northrop, 1999). Then, procedure (i) applies with the modification that (i.3) is replaced by:
(i.3') Draw a random value $x^c$ from the density.

In the second case we draw from the uniform distribution, assuming that the hyperrectangle $X$ is the compact support of $f(x)$.

(iii.1) Start from iteration $t=1$ and
$$w^t_1 = 1.$$

(iii.2) Calculate the step size $\rho_t = 1/t$.

(iii.3) Draw $x^c$ uniformly from the rectangle $X$.

(iii.4) Evaluate for all $s = 1, \ldots, S$ the kernel densities
$$\zeta^t_s = \psi_\theta(x^t - x^c).$$

(iii.5) Evaluate the mean sample density (jointly with (iii.4))
$$\phi_t = \frac{1}{S} \sum_{s=1}^S \zeta^t_s.$$

(iii.6) Evaluate the population density
$$f_t = f(x^c).$$

(iii.7) Update the estimate of the mean
$$w^t_{r+1} = w^t_r + \rho_t \left( \zeta^t_r \frac{f_t}{\phi_t} - w^t_r \right).$$

(iii.8) If $t \leq T$ increment $t$ by 1 and go to (iii.2).

After running one of these algorithms, we can conduct the survey tabulation. The variability of these weights within the survey can be described through a cumulative distribution whose variance, skewness, and other moments can be computed on the basis of the estimated weights. Alternatively, it is possible to measure higher moments of each weight over the sample, by including an additional equation in step (i.6) or (iii.7) that keeps a record of the mean square value, as opposed to the mean value itself. Use of an explicit density function is to be considered when the census data are somewhat outdated, while some variables $z$ are available that can serve to update the census density.

**Prediction error**

The definition of weights (3.9) is based on kernel density prediction (3.5). The estimated error in regression is at observation point $s$:
$$\eta^s = y^s - \hat{y}_\theta(x^s), \quad (3.14)$$

Using the formulation as (3.5) it is possible to estimate the error at intermediate points:
$$\hat{f}_\theta(x) = \sum_r P^s_\theta(x) \eta^s, \quad (3.15)$$

Thus, the mean square error can be calculated as:
$$\hat{e}^2 = \frac{1}{S} \sum_s (\eta^s)^2 w^s, \quad (3.16)$$
while the variance would be

\[ s^2 = \frac{1}{S} \sum_s (\eta_s - \hat{\epsilon})^2 w_s, \tag{3.17} \]

where \( \hat{\epsilon} \) is the mean error. If one is willing to assume asymptotic normality, this variance can be used to characterize the reliability of this mean. We note that if we apply kernel density regression this variance can be approximate by Monte Carlo integration, but if we use in (3.14) a parametric form as in (3.2), it becomes natural to look for parameter values that minimize this variance or some other integral measure, rather than the sum of square deviations at observation points. This could be done by weighted least squares or any other parametric regression technique. However, under more general specifications the weight would also become parameter dependent. In this case, it becomes necessary to allow for parameter adjustment during the Monte Carlo iterations. This is precisely what is done in stochastic quasi-gradient algorithms.4

Two alternatives suggest themselves. The first applies the interpretation of \( w/S \) as probability of a given deviation. This makes it possible to calculate the probability of a deviation from the mean exceeding, say, 10 per cent. The second alternative relies on Monte Carlo simulation to generate a sequence of samples (surveys). This generates a series \( \mu^s \) of means, whose average and variance and higher moments can be computed, say, for varying sample sizes. Suppose that the model was estimated parametrically. Then, the left-hand term of (3.14) would only measure the idiosyncratic error and neglect the part due to errors in parameters. If the model is linear in parameters, it is possible to use an analytical expression for this error but in the nonlinear case it will often be necessary to rely on simulation in order to map out how the prediction varies with the data set sampled (see e.g. Yatchew, 1998).

**Implications for sample design**

It would be possible to treat the weight (or the correction factor) as dependent variable of a regression over the sample, so as to identify the causes of under- and overrepresentation. However, this presupposes that it would be desirable to have a sample with equal weights for all observations, whereas the discussion so far suggests that reliability of the regression function would be more important. Even though there is no “hard” number measuring the contribution of an additional observation, and therefore indicating the ideal sample size, the uncertainty of the estimate is nonetheless determined by the variance of the error in \( y \) rather than by the variance of \( y \) itself. Now suppose that the vector \( x \) can be partitioned into \( x_1, x_2 \) where \( x_2 \) refers to the elements of the sample frame, for which the sample can be controlled: say, geographical area, age or sex. These elements only have discrete values and thus define the strata for sampling. Then, we can for error terms \( \eta^c \) from (3.14) to calculate the mean square error for every stratum \( f \) in the census as:

\[ \sigma^2_f = \frac{1}{N_f} \sum_{c|x_2^c=f} (\eta^c)^2, \tag{3.19a} \]

\[ N_f = \sum_{c|x_2^c=f} 1. \tag{3.19b} \]

4 See Ermoliev and Wets, 1988, for a general introduction to the case where the minimization problem is convex and Ermoliev et al., forthcoming, for a treatment of non-convexities.
Assuming that the errors are not correlated across strata, these calculations could serve to
determine the sample size distribution (2.2).

4. Effectiveness of the procedure: a simulation exercise

So far, we only described the procedure for calculating the weights. To assess its effectiveness,
we conduct experiments for which the true census value of the mean is known. One approach to
effectuate this would be apply bootstrapping to create a series of samples of a fixed size by
sampling with replacement from the survey data and to calculate the mean, variance of the
prediction error. However, this prediction would be subject to three sources of error: (1) an error
caued by the lack of compatibility between the survey and the census data; (2) an error due to
the poor fit of the function within the survey; (3) an error in extrapolation from the survey to the
census. Since we are interested in assessing the procedure rather than the correspondence
between the survey and the census, we choose to circumvent the first type of error by sampling
the survey data from the census itself.

Specifically, a given experiment creates a series of (survey) data sets that are created as
biased samples from a given “census” (here the GLSS, Ghana Living Standard Survey, Ghana
Statistical Service, 1989, 1993, 1995). For each data set, a measure is evaluated of the extent to
which the weighted mean can redress the sample bias, for selected variables not used in the
weight calculation. We define this measure as

\[
\gamma_k^s = 1 - \frac{|\hat{\mu}_k^s - \mu^c|}{|\hat{\gamma}_k^s - \gamma^c|},
\]  

(4.1)

where the superscripts S, C refer to the unweighted survey mean, the weighted survey mean and
the mean in the census, respectively. This measure is negative when the estimation amounts to
deterioration relative to the survey mean, and positive if it yields an improvement. It has value
unity in the ideal case. The survey data are constructed as follows. A data set of size S is created
by sampling uniformly from the population, while dropping all observations that lie outside a
specific range for selected variables. For example, we can drop all households whose head is
younger than a specified age, or has less than a given number of years of education. The resulting
sample will be biased. The exercise was conducted using the Ghana Living Standard Survey, with
100 biased samples of 100 observations from a census of 6600 observations. Overlapping x-
variables are the size of the household, the age and the number of years of education of the head
of household. The y-variables for which the mean is to be estimated are per capita expenditure,
per capita food bought, per capita food that was home produced, the height, bodymass and the
years of education.

The results of the GLSS exercise are shown in Figure 1, as the frequency distribution of \( \gamma^s \).
It appears that the kernel density regression achieves remarkable improvement.
5. Conclusion

The possible non-representativeness of household surveys can be addressed by reweighting the survey observations through a correction factor. This factor is usually computed on the basis of the frequency of a combination of discrete household characteristics, such as the age and sex of the respondent, in the survey relative to its value in a population census. This procedure has the disadvantage of requiring arbitrary grouping into discrete classes for common variables with cardinal measurement. Also, as the number of shared properties grows larger, the evaluation and storage of the discrete frequency distribution of the census becomes cumbersome, the assignment of census classes to survey observations more difficult and many classes will remain without any assignment.

In this paper, we have specified a generalization that enables us to account for several household characteristics including real-valued ones, avoids complicated assignments and errors due to the arbitrariness of classification of real-valued variables. Kernel density regression was applied to estimate the joint density over the survey and how the weights could be estimated by Monte Carlo integration of the estimated density over the census distribution. A Monte Carlo experiment that uses the Ghana Living Standard Survey as census data shows a good performance under repeated biased sampling from the same data set.

A comparison of this kernel density approach with any alternative parametric regression technique for obtaining the regression model \( y(x) \), say, maximum likelihood estimation, suggests the following advantages. First, as mentioned earlier, unlike other methods, kernel density estimation produces weights that apply to all elements of the vector \( y \) of survey variables, and can be computed independently of the observations \( y_s \). This makes it possible to compute the weights prior to tabulation. Second, kernel density estimation is very flexible and easily produces a good fit to an unknown sample density. Third, the method of calculation does not involve any iterative optimization procedure and is therefore more easily applicable when a flexible form has to be adopted and also eases the design of reliable software. Fourth, because the weight does not depend on \( y \), a possibly poor fit between \( y \) and \( x \) does not impair on the tabulation. Fifth, kernel density accounts directly for errors in the common variable \( x \), whereas a maximum likelihood estimation treats \( x \) as given and requires some prior instrumentalization on \( x \) to correct for such errors. Finally, in case the vector \( x \) only consists of discrete variables, then the formulation
reduces to the standard weighting procedure on the basis of relative frequencies mentioned in the introduction, whereas parametric estimations would yield different results.

The kernel density approach has the limitation that its prediction might have a bias which persists even after integration over \( x \) and that since the kernel estimate is a convex combination of the observations \( y^i \), it cannot adopt a value outside this range. It is therefore less appropriate in cases where a significant fraction of the census values fall outside the range of the survey data. Clearly, improvement by reweighting presupposes the availability of a census data set that is sufficiently up to date. It can, for example, be used to study the implications of a newly compiled census for the outcomes from previous surveys. However, in practice the converse situation will often apply, where one would like to infer the implications at census level of a new, specialized survey, given an older census. Then, it will often be necessary to perform some adjustments that can account for, say, demographic change and migration in the intermediate period.

Follow up activities of the present research are scheduled to include parallel exercises for other sets of census data as well as an application within a geographical context where extrapolation of kernel density regressions to census level will be used to estimate geographical distributions rather than only a mean.
Appendix: Computing the weights

Calculation of the weights relies on two data sources, the survey and the census. To keep the computing time and the overhead to a minimum, a Fortran program was written that performs the computational tasks. A normal density is chosen as kernel because it has the advantage of keeping the value of the density positive at all x-values. It uses two data files as input, one with the overall control parameters, and one with the census data. The Fortran program can also be called within a SAS-interface program,\(^5\) that can be used to prepare the data and to retrieve the resulting weights for further processing. Here we only describe the files of the Fortran program, to be called as

```
WEIGHT.EXE <Census><Survey><Results><Other_output>,
```

where <..> denotes user-specified file names.

The census file consists of consecutive records, one for every respondent, with data in ASCII form, separated by blanks or by commas. The survey file is also in ASCII form and starts as follows:

```
* Survey data for reweighting program
* Number of variables in survey
  15
* Number of observations in survey
  703
* Theta
  1.000
* Number of variables in census
  16
* Approximate number of observations in census
  6500
* Fraction of census to be used for Monte Carlo
  .5
* Numerical code for missing data in census
  -9.100
* The following table is in fixed format: A9, A25, I3, A1, 2F8.2
*234567890123456789012345678901234567890123456
* Labels     Order    Lower Upper
V_42  Sex                                   2     0.00 1.00
V_52 Age                               13       0.00       1.00
V_54 Herdsize                               11c     0.00       1.00
V_56 Household size                              5  0.00 1.00
* Data
1  1 1 56 3 2 2 0
1 2 1 7 ... 
```

Table A1: Header of the Survey file

---

\(^5\) The SAS-interface was written by Peter Albersen.
We notice the entry code “11c” in the third column. Here the c-flag inactivates the selection of this element (and so does any non-blank character). The fourth and fifth column permit to restrict the observations to values within the percentage range indicated. The header of the Result file is shown in Table A.2. The Min., Aver., and Max. columns show the minimal, average and maximum value sampled from the census, and the % missing column gives the percentage of missing observations found for the variable.

<table>
<thead>
<tr>
<th><em>Results</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Number of variables used</em></td>
<td>3</td>
</tr>
<tr>
<td><em>Number of observations in survey</em></td>
<td>699</td>
</tr>
<tr>
<td><em>Window size</em></td>
<td>.336</td>
</tr>
<tr>
<td><em>Number of drawings from census</em></td>
<td>6000</td>
</tr>
<tr>
<td><em>Standard deviation of weights</em></td>
<td>.795</td>
</tr>
</tbody>
</table>

*Variable| Label| Census:| Nr| Min.| Aver.| Max.| %Missing|
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WG</td>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>V_42</td>
<td>Sex</td>
<td>2</td>
<td>1.00</td>
<td>1.20</td>
<td>2.00</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>V_52</td>
<td>Age</td>
<td>13</td>
<td>15.00</td>
<td>43.56</td>
<td>99.00</td>
<td>.00</td>
</tr>
<tr>
<td>4</td>
<td>V_56</td>
<td>Household size</td>
<td>5</td>
<td>1.00</td>
<td>4.72</td>
<td>38.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

*Results for survey variables 1-4*

| 1| 1.4557| 1.0000| 56.000| 1.0000 |
| 2| .98320| 1.0000| 74.000| 2.0000 |
| 3| .76742| 1.0000| 35.000| 4.0000 |
| 4| .33203| ...| |

Table A2: Header of the Result file
References


The Centre for World Food Studies (Dutch acronym SOW-VU) is a research institute related to the Department of Economics and Econometrics of the Vrije Universiteit Amsterdam. It was established in 1977 and engages in quantitative analyses to support national and international policy formulation in the areas of food, agriculture and development cooperation.

SOW-VU's research is directed towards the theoretical and empirical assessment of the mechanisms which determine food production, food consumption and nutritional status. Its main activities concern the design and application of regional and national models which put special emphasis on the food and agricultural sector. An analysis of the behaviour and options of socio-economic groups, including their response to price and investment policies and to externally induced changes, can contribute to the evaluation of alternative development strategies.

SOW-VU emphasizes the need to collaborate with local researchers and policy makers and to increase their planning capacity.

SOW-VU’s research record consists of a series of staff working papers (for mainly internal use), research memoranda (refereed) and research reports (refereed, prepared through team work).

Centre for World Food Studies
SOW-VU
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

Telephone (31) 20 - 44 49321
Telefax (31) 20 - 44 49325
Email pm@sow.econ.vu.nl
www http://www.sow.econ.vu.nl/