Journal of Official Statistics, vol. 29, n. 3 (2013)

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 329–353

Unit Nonresponse and Weighting Adjustments: A Critical Review

J. Michael Brick[1]

This article reviews unit nonresponse in cross-sectional household surveys, the consequences of the nonresponse on the bias of the estimates, and methods of adjusting for it. We describe the development of models for nonresponse bias and their utility, with particular emphasis on the role of response propensity modeling and its assumptions. The article explores the close connection between data collection protocols, estimation strategies, and the resulting nonresponse bias in the estimates. We conclude with some comments on the current state of the art and the need for future developments that expand our understanding of the response phenomenon.

Key words: Response propensity; bias; data collection; calibration.

1. Introduction

This article critically reviews aspects of unit nonresponse in sample surveys, where unit nonresponse is defined as the failure to obtain a valid response from a sampled unit. We emphasize the consequences of unit nonresponse and methods of adjusting for it in circumstances that are typical of cross-sectional household surveys. Establishment surveys and attrition nonresponse in panel surveys are also subject to unit nonresponse, and issues reviewed here pertain to these surveys. However, the data collection design options, reasons for nonresponse, and auxiliary data available for adjustment differ dramatically across types of surveys. Because these features are critical to dealing with nonresponse and nonresponse bias, we have chosen to focus on situations frequently arising in cross-sectional household surveys.

Unit nonresponse is just one form of missing data in surveys. Other types of missing data include incomplete coverage of the target population, item nonresponse, and partial nonresponse such as wave nonresponse in panel surveys and failure to obtain second-phase responses in two-phase surveys. While these are all important, they are beyond the scope of this review.

Most surveys, especially government surveys, employ large sample sizes and design-based theory to make inferences from the sample to the target population. This theory assumes complete response. While surveys employ methods to minimize nonresponse and its effects on estimates, in almost every survey some sampled units do not respond.

[1] Westat, 1600 Research Blvd, Rockville, MD 20850, U.S.A. Email: MikeBrick@westat.com
Acknowledgments: We would like to thank Graham Kalton, Sharon Lohr, and Douglas Williams, who provided helpful comments on earlier drafts.

Model assumptions and adjustments are made in an attempt to compensate for missing data. Because the mechanisms that cause unit nonresponse are almost never adequately reflected in the model assumptions, survey estimates may be biased even after the model-based adjustments. Nonresponse also causes a loss in the precision of survey estimates, primarily due to reduced sample size and secondarily as the result of increased variation of the survey weights. However, bias is the dominant component of the nonresponse-related error in the estimates, and nonresponse bias generally does not decrease as the sample size increases. Thus, bias is often the largest component of mean square error of the estimates even for subdomains when the sample size is large.

The classification of nonresponse by reason is important because the effects and methods of dealing with nonresponse may be directly tied to the reason (Lin and Schaeffer 1995; Steele and Durrant 2011). Reasons for unit nonresponse are usually classified as the failure to contact the sampled unit, the inability to persuade the sampled unit to respond, and other reasons (Brick and Montaquila 2009). Noncontact or inaccessibility nonresponse may occur for a variety of reasons. For example, the sampled unit may not be at home during the times the data collector visits or calls, the survey schedule may limit the number of contact attempts, or data to locate the sampled unit may be incorrect or out of date. Refusal nonresponse may occur because the sampled person does not wish to participate in the particular survey, or because someone else such as a gatekeeper refuses to provide access to the sampled person. For example, in a telephone survey the person answering the telephone may not be willing to give the telephone to the sampled person. While noncontact was a larger component of total nonresponse in earlier times, refusals now constitute the majority of total nonresponse in most surveys (Atrostic et al. 2001; Brick and Williams 2013). The other nonresponse category includes assorted reasons such as language problems and health problems that may prevent the sampled unit from responding. These other problems are typically a small proportion of the total nonresponse in a survey, but may be important in some cases (see Feskins et al. 2011; Brick et al. 2012).

2. Background

Unit nonresponse has been recognized as a potential problem since the early days of probability sampling. Colley (1945), Hansen and Hurwitz (1946), Ferber (1949), Yates (1946), and Deming (1953) are examples of early research that examined data collection and weighting methods to deal with nonresponse[2]. As research on nonresponse and its effects accumulated and worries about increasing nonresponse rates were expressed, the Committee on National Statistics in the United States convened a Panel on Incomplete Data in 1977 to consolidate this research and develop new approaches. The Panel's work resulted in a three-volume set in 1983 (Vol. 1 edited by Madow, Nisselson, and Olkin; Vol. 2 edited by Madow, Olkin, and Rubin; Vol. 3 edited by Madow and Olkin) that was the first monograph dedicated to nonresponse in surveys. Around the time of the Panel, the way nonresponse was conceived and adjustments were motivated began to shift to treat

[2] The references in this section are examples and useful summaries of a body of work and are not intended to assign precedence for ideas.

response as a random rather than fixed outcome. In our review, several references to chapters from one of the three volumes reflect some of these changes.

In the years following the Panel's meetings, several published books were devoted largely to survey nonresponse. These include Kalton (1983), Goyder (1987), Brehm (1993), Groves and Couper (1998), Tourangeau et al. (2000), Groves et al. (2002), Särndal and Lundström (2005), Stoop (2005), Stoop et al. (2010), and Bethlehem et al. (2011). Journals have dedicated special issues to survey nonresponse, including the *Journal of Official Statistics* and *Public Opinion Quarterly*. International workshops and symposiums have been also been held; the Groves et al. (2002) monograph is a product of one of these.

To provide some context for this research, we identify three major themes in nonresponse research (although there is considerable overlap among them). One theme is the study of the response mechanism that causes nonresponse. This research seeks to understand important psychological and sociological factors that dispose some units to respond and others to fail to respond. Goyder (1987) is an example of this work that takes a sociological perspective on the causes of nonresponse; Tourangeau et al. (2000) is an example taking the psychological view. Most of the psychological and sociological research examines the willingness or amenability of the sampled unit to participate in the survey by looking at factors such as the interviewer, the survey materials, and the characteristics of the respondent that might influence response.

A second theme is data collection methods to reduce nonresponse. Dillman's (1978) tailored design method illustrates one branch within this theme. He offers general approaches to the design of data collections to increase cooperation rates and improve the chances of reaching respondents to deliver the survey request. The literature on incentives is another such example (Singer 2002). The other branch within this theme describes a set of methods for following up nonrespondents; survey methods to gain the cooperation of those who refuse the initial survey request or who are never contacted are important topics in this area. Switching modes for nonresponse follow up is an example within this area (Dillman et al. 2009).

Statistical adjustment of the survey weights to adjust for survey nonresponse is a third theme while retaining the design-based mode of inference. Särndal and Lundström (2005) is an example. They examine statistical models to adjust the estimates from the survey after the nonresponse has been realized. The aim of all of this research is to reduce the level of nonresponse and develop methods to minimize nonresponse bias in the estimates.

For many years, nonresponse bias and response rates were often treated as equivalent, or at least surveys with low response rates were thought likely to have the potential for high nonresponse bias in the estimates. Data collection efforts that increased response rates were assumed to reduce nonresponse bias. This presumed relationship is especially pronounced in the literature on incentives, where effects of incentives on response rates are carefully described and nonresponse bias is often not assessed directly (Singer and Ye 2013). The reasons for this assumption are easy to understand. Response rates are easy to compute, provide a single measure for an entire survey, and have face validity.

A spate of articles in the last decade forced researchers to reconsider this presumption. These articles show that the empirical relationship between response rates and nonresponse bias is not strong (e.g., Keeter et al. 2000; Curtin et al. 2000;

Groves 2006). Of course, even long ago we knew that a single measure like a response rate could not be used to predict nonresponse bias. Ferber (1949, p. 672) noted "The problem of response bias must be considered with specific reference to a particular question or characteristic. The presence of bias in one question does not mean *a priori* that the replies to other questions on the same questionnaire are also biased."

Falling response rates in most countries across the developed world, especially in the past few decades, are documented in various reviews (e.g., Stoop 2005; Steeh et al. 2001; Atrostic et al. 2001; de Leeuw and de Heer 2002; Smith 1995; and Synodinos and Yamada 2000). Furthermore, the trend toward lower response rates is happening despite additional procedures aimed at increasing response in many surveys. Some of these procedures are designed to increase contact rates and others are aimed at reducing refusals. However, none of these methods appear to be capable of reducing the level of nonresponse, and reliance on adjustments to the survey weights is increasing.

Although response rates may not be predictive of nonresponse bias, the declines in response rates have raised the level of concern among survey methodologists and prompted new developments. Some debate whether low response rate probability samples are qualitatively different from nonprobability samples; others have sought to find different measures that are more predictive of nonresponse bias. Schouten et al. (2009) propose R-indicators to serve as a substitute for response rates. These indicators attempt to measure how similar the respondents are relative to the full sample by estimating the variability in the estimated response propensities, where the response propensity (ϕ_i) for every sampled unit i is its probability of responding to the survey. Schouten et al. (2009) define the R-indicator as

$$R(\phi(\mathbf{x})) = 1 - 2S(\phi(\mathbf{x})), \tag{1}$$

where $S(\phi(\mathbf{x}))$ is the population standard deviation of the response propensities and $\mathbf{X}$ is a vector of auxiliary variables known for the full sample. If the R-indicator is close to unity, the respondent set is more 'representative' of the target population, at least as measured with respect to $\mathbf{X}$, and has a lower potential for nonresponse bias. Schouten et al. (2011b) extend these results.

Särndal and Lundström (2005) and Särndal (2011a) propose what they refer to as balance indicators that are intended to measure the similarity between the respondents and the sample. Some of these indicators are like the R-indicators in that they measure variation in subgroup response rates, where the subgroups are formed based on auxiliary variables. Wagner (2010) proposes using the fraction of missing information as an alternative to the response rate because this measure permits the inclusion of auxiliary variables in the determination of the influence of the missingness on the estimate.

All of the alternatives for response rates are only able to measure representativeness of the respondents in relation to $\mathbf{X}$, the auxiliary variables available. Different choices of $\mathbf{X}$ lead to different values of the indicators. Although using these data is an improvement over response rates that do not consider any auxiliary data, the measures are only useful when powerful auxiliary variables for the specific estimates are available.

Some of these measures were influenced by the desire to continually monitor the data collection process for responsive designs (Groves and Heeringa 2006). Responsive and adaptive designs are two data collection approaches that have been proposed as a way to

reduce nonresponse bias. Responsive design makes changes to data collection strategies during data collection when one recruitment protocol is no longer successful in getting responses from sampled units, especially units with differing characteristics (Groves and Heeringa 2006). For responsive designs, data for making these decisions must be collected and analyzed rapidly during the field period. Adaptive design is similar, but the analysis of response patterns may be done from previous or similar collections (Schouten et al. 2011a). Both responsive and adaptive designs contemplate data collection strategies that are tailored for specific sampled units, whereas the standard data collection procedure for many years has been to apply a single protocol to all units.

3. Bias Representations

The rationale for the design, data collection, and estimation approaches mentioned above is based on models of nonresponse bias. Two models dominate the way we think about nonresponse bias. The models are most often presented in terms of the bias of an unadjusted estimator of the mean, where unadjusted implies using the full sample estimator with just the respondent data. The unadjusted Horvitz-Thompson estimator of the total is

$$\hat{y}_{un} = \sum_{i \in s_r} d_i y_i, \tag{2}$$

where d_i is the inverse of the probability of selection of unit i and the sum is over s_r, the set of respondents. The ratio mean is $\hat{\bar{y}}_{un} = \hat{y}_{un} \Big/ \sum_{i \in s_r} d_i$.

The deterministic representation of bias partitions the population into respondent and nonrespondent strata (Cochran 1977), and nonresponse bias is then a function of the nonresponse rates and the characteristics of the units in these strata. In the deterministic approach, response is a fixed outcome of the survey (and the procedures used in data collection) and is not subject to random variation other than the variation due to sampling the response strata. The nonresponse bias of the unadjusted estimator of the mean is

$$bias\left(\hat{\bar{y}}_{un}\right) \approx (1 - P)\left(\overline{Y}_r - \overline{Y}_m\right), \tag{3}$$

where P is the proportion of units in the respondent stratum, $\overline{Y}_r$ is the mean in the respondent stratum, and $\overline{Y}_m$ is the mean in the nonrespondent stratum (Thomsen 1973). The expression shows that bias depends on the response rate and the distribution of each characteristic as discussed by Ferber (1949). However, a difficulty with Expression (3) is that the response strata definition is *post hoc* so it is difficult to use this in advance of data collection.

The alternative stochastic model has become more popular since the late 1970s, although its origins go back as early as Politz and Simmons (1949) and Hartley (1946). It assumes that response is a random variable and the probability of response is like the probability in an additional phase of sampling, but the probability of response for every unit i in this phase is unknown.

The nonresponse bias of an estimated ratio mean under the stochastic model is

$$bias\left(\hat{\bar{y}}_{un}\right) \approx \bar{\phi}^{-1} \sigma_\phi \sigma_y \rho_{\phi,y}, \tag{4}$$

where $\bar{\phi}$ is the population mean of the response propensities, σ_ϕ is the standard deviation of ϕ, σ_y is the standard deviation of y, $\rho_{\phi,y}$ is the correlation between ϕ and y, and $\phi_i > 0$ for all i (Bethlehem 1988). The estimated respondent mean is unbiased if ϕ and y are uncorrelated.

The two expressions are appropriate for the Horvitz-Thompson of the unadjusted mean, but different relationships hold for totals, correlations, and other statistics as well as for different estimators. Brick and Jones (2008) extend these results to other types of statistics and some calibrated estimators.

Both models are useful for estimating the potential bias under particular circumstances. For example, if data are available for all units in the population, then the bias can easily be computed using (3) or (4) after data collection is complete. Both bias expressions are equivalent in this case. The two models also lead to similar conclusions about how to attempt to adjust for biases due to nonresponse. We find the stochastic model to be generally more helpful when speculating about the potential magnitude of bias prior to data collection. It expresses bias in terms of a correlation so it is bounded, and correlations computed from other surveys may be useful guides for speculating about the magnitude of correlation.

Thus far, we have discussed bias in the simple situation in which no other information is known about the sampled units. In practice, we often have other data available for either the sampled units or the entire population. Thus, the expressions given above can be revised slightly to account for the auxiliary information. For example, the response propensity can be written more formally as

$$\phi_i = \phi(\mathbf{x}_i) = \Pr(R_i = 1 | \mathbf{X} = \mathbf{x}_i), \tag{5}$$

where $\mathbf{X}$ consists of the set of variables known for the full sample and $R_i = 1$ if unit i responds (Rosenbaum and Rubin 1983). The bias expressions for both the deterministic and stochastic models can also be modified to account for auxiliary data. For example, suppose auxiliary data are available and used for poststratification. The stochastic expression for the bias of the poststratified estimator of the mean is

$$bias\left(\hat{\bar{y}}_{ps}\right) \approx N^{-1} \sum_h \bar{\phi}_h{}^{-1} \sigma_{\phi_h} \sigma_{Y_h} \rho_{\phi_h, Y_h}, \tag{6}$$

where h denotes the poststratification classes. See Kalton (1983), Brick and Kalton (1996), and Bethlehem et al. (2011) for such expressions and their implications.

The auxiliary variables are very valuable for adjusting the design weights to account for nonresponse. Kalton (1983, p. 63) states: "Among the potential variables for use in forming weighting classes, the ones that are most effective in reducing nonresponse bias are those that are highly correlated both with the survey variables and the (0,1) response variable." Both (3) and (4) explicitly contain the characteristic being estimated, suggesting that adjustments could be developed by modeling the distribution of the characteristic.

Two types of auxiliary variables can be used: if the auxiliary variables are known for all sampled units, then the adjustment is called sample-based or Info-S; if they are known for the entire population, the adjustment is population-based or Info-U (Kalton and Kasprzyk 1986; Lundström and Särndal 1999). The population-based adjustment is especially useful when characteristics for the entire sample are not available but the population totals are

known, because these adjustments only require capturing the data from the respondents. Population-based adjustments may also reduce noncoverage error and sampling error. Sample-based adjustments need data for the full sample but do not require knowing control totals for the entire population. Sample-based and population-based adjustments are equally effective for dealing with nonresponse bias (Särndal and Lundström 2005; Brick and Jones 2008).

4. Modeling and Missing Data Mechanisms

As noted above, modeling either the response propensity or the outcome variable can be effective for reducing nonresponse bias. Nevertheless, this section discusses only response propensity modeling, for two reasons. First, modeling outcomes and using design-based calibration estimators like the generalized regression estimator can be extremely valuable for improving the precision of the estimates even when there is full response. Ratio and regression estimators were originally developed exactly for these reasons. These estimators are also beneficial at reducing nonresponse bias when the same variables are correlated to response (e.g., Bethlehem 1988; Fuller et al. 1994). Our perspective is that powerful auxiliaries for key outcomes should be included in the estimator when they are available, irrespective of their relationship to response.

Second, in our experience most cross-sectional household surveys produce multiple characteristics and there are few auxiliary variables that are related to any of these outcomes. In this situation, response propensity modeling may be the only remaining tool to reduce nonresponse bias. It has the potential to reduce bias for variables that cannot be modeled directly because powerful correlates of the variable are not available. Of course, this approach is not a panacea by any means. Often, bias is reduced by response propensity weight adjustments, but only partially, as shown by Micklewright et al. (2012).

We also concentrate on nonresponse where the data are missing at random (MAR). In our notation, the missing data mechanism is MAR (see Rubin1976; Little and Rubin 2002) when

$$\Pr(R_i = 1|Y_i, \mathbf{X}_i) = \Pr(R_i = 1|\mathbf{X}_i) \tag{7}$$

for all sampled units. Roughly speaking, under the MAR assumption the missing data mechanism may depend on observed data but not on unobserved data. When (7) does not hold, the missing data mechanism is called not missing at random (NMAR). Although this dichotomy is useful, in practice it is not possible to assess whether the data mechanism is MAR or NMAR without obtaining additional data for the nonrespondents.

Two approaches have been proposed for handling nonresponse when researchers assume the mechanism is NMAR. The first is called the selection model approach; it postulates a model that relates the missing data to the distribution of the outcome. Heckman (1979) is probably the best-known example of an explicit selection model. Greenlees et al. (1982) also use this approach. A second approach is the pattern mixture model (PMM), where the distribution of Y is conditioned on the missing data and mixed or averaged over different populations (Little 1993). Andridge and Little (2011) have recently expanded on the PPM approach using a proxy variable. Nearly all researchers using NMAR models strongly urge sensitivity analyses to determine whether the estimates

are robust to the modeling assumptions, since generally there is no other way to assess these assumptions.

Molenberghs et al. (2008) show that for every NMAR model there is a MAR counterpart that has an equal fit to the observed data. This means that the NMAR model cannot be distinguished from its MAR counterpart based on the observed data. Even though they have equal fits, the models do not necessarily produce the same estimates. In a similar vein, David et al. (1986) re-examine the NMAR approach of Greenlees et al. (1982) using a MAR model and find that the MAR model is adequate. Molenberghs et al. (2008) show an example where the estimates from the NMAR models and their MAR counterparts are very different. They use a series of MAR counterparts corresponding to NMAR models for sensitivity analysis. Since MAR models are usually easier to understand and describe, in the following sections we generally restrict our attention to MAR models. We will return to this concept later.

5. Response Propensity Weight Adjustment

One approach to weight adjustment is to model the response propensities for the sampled units individually, and the adjustment factor is the inverse of the estimated propensities of the respondents. The idea is to replace the unknown probability of response by an estimate. The propensity-adjusted estimator of the total is

$$\hat{y}_{rp} = \sum_{i \in s_r} d_i \hat{\phi}_i^{-1} y_i \tag{8}$$

where $\hat{\phi}_i$ is the estimated propensity for unit i where i is a respondent. The $\hat{\phi}_i$ are usually estimated by logistic regression, but probit and nonparametric methods are also used (Little 1986; Da Silva and Opsomer 2009; Phipps and Toth 2012).

As mentioned above, Politz and Simmons (1949) pioneered thinking about stochastic response models when they estimated propensities by collecting data on how often the respondent would be at home on different days. These data provide a basis for estimating contact propensities to account for noncontact nonresponse. Related methods such as those proposed by Bartholomew (1961) and Dunkelburg and Day (1973) have not generally proven to be effective, especially as contact rates have risen due to increased data collection efforts.

Rather than estimating individual response propensities, the approach most surveys use is to form groups and adjust the weights in each group by the inverse of the observed group response rate. Särndal et al. (1992) describe these as response homogeneity groups (RHGs). Weighting classes is an alternative term that has been used for decades. Important outcome statistics or domains may also be considered when forming RHGs. If all the units within an RHG have the same response propensity so that MAR holds, any nonresponse bias is eliminated (see Da Silva and Opsomer (2004) for extensions). In this case, (8) is a weighting class estimator and can be written as

$$\hat{y}_{wc} = \sum_g \sum_{i \in s_{r_g}} d_{gi} \hat{\phi}_g^{-1} y_{gi} \tag{9}$$

where $g = 1,2,\ldots,G$ are the RHGs, $i \in s_{r_g}$ is a respondent in RHG g, and $\hat{\phi}_g$ is the estimated response propensity in g. One issue that often arises with weighting class

estimators is the need to have large enough respondent counts in each cell to avoid unstable estimates. For this reason, Little (1986) proposes using cells based on the estimated propensity scores rather than individually estimated propensities.

A third general approach is to use calibration estimation (Deville and Särndal 1992) for adjustment. Lundström and Särndal (1999) extend calibration estimators to encompass estimators to include both sample-based and population-based information for nonresponse adjustment. The calibration estimator is

$$\hat{y}_{ca} = \sum_{i \in s_r} d_i^* y_i, \tag{10}$$

where the sum is over the respondents, d_i^* is the adjusted weight that satisfies the calibration equation $\sum_{i \in s_r} d_i^* \mathbf{x_i} = \mathbf{X}$, $\mathbf{x_i}$ is a vector of auxiliary variables, and $\mathbf{X}$ is a vector of totals (sample based, population based, or a combination of the two) of those auxiliary variables. Since the weights are not uniquely defined by these conditions, other constraints may be imposed, such as $d_i^* = d_i v_i$, where v_i is a linear regression estimate (Bethlehem 2002; Särndal and Lundström 2005). A wide variety of nonresponse adjustment estimators are in this class, including poststratification, raking, and generalized linear regression estimators. Lumley et al. (2011) give insight into the relationship between calibration estimators and nonresponse bias for different estimators.

Poststratification is a simple calibration estimator that has a single dimension and has been used for decades (Holt and Smith 1979). Assume that poststrata are defined by the number of persons in age categories (N_h) and that N_h is known for the entire population. In this case, (10) simplifies to $\hat{y}_{ps} = \sum_h \frac{N_h}{\hat{N}_h} \sum_{i \in s_{r_h}} d_i y_i$, where $\hat{N}_h = \sum_{i \in s_{r_h}} d_i$ and the sum is over the respondents in poststratum h. The calibration equation forces the estimator for the age groups to match the known population total for that group.

It is easy to see that the weighting class estimator given by (9) is a sample-based calibration estimator – the calibration equation in this case forces the adjusted weight to reproduce the weighted (using d_i) distribution of the weighting classes from the sample. A related estimator that only uses $\mathbf{x_i} = 1$ for all $i \in s$ is called the primitive estimator by Särndal (2011b) and is given by

$$\hat{y}_{pr} = \left(\sum_{i \in s} d_i \right) \left(\sum_{i \in s_r} d_i \right)^{-1} \left(\sum_{i \in s_r} d_i y_i \right). \tag{11}$$

Estimators of this nature have a substantial effect on the bias of the estimated total but have no effect on the ratio mean.

Details on specific nonresponse adjustment techniques are covered in several articles and texts, including Särndal and Lundström (2005), Kalton and Flores-Cervantes (2003), Chang and Kott (2008), Brick and Montaquila (2009), and Bethlehem et al. (2011). Generally, the specific form of the adjustment is not highly related to the bias reduction, except when the form limits the ability to take advantage of all the information in the auxiliary data. For example, poststratification may be less effective than linear calibration or raking when many variables are available because poststratification has one dimension.

In addition, any method that results in large variability in the nonresponse adjustments due to instability in the estimated adjustments should be avoided since that may increase the variance of the estimates without further reducing bias.

The basic theory underlying the adjustment methods described above is formalized by Cassel et al. (1983), who treat response as an additional phase of "sampling" (see also Oh and Scheuren 1983). According to this theory, the adjusted estimator should have desirable statistical properties such as unbiasedness and consistency when expectations are taken over both sampling and response mechanisms, provided that the response propensities can be adequately estimated. Suppose that RHGs are formed and the adjustment to the sampling weight is the inverse of the response rate in the RHG, $\hat{\phi}_g^{-1}$. The heuristic interpretation is that each respondent in an RHG g "represents" $\left(\hat{\phi}_g^{-1} - 1\right)$ nonresponding units in the group. Within this framework, the goal is to identify groups of units with the same probability of responding to the survey at the end of data collection, so that the MAR assumption is satisfied. The methods employed to create the RHGs and the choice of variables for creating these groups are an essential feature of nonresponse weighting.

6. Choosing Auxiliaries and Alternative Metrics

Traditionally, auxiliary variables and weighting classes were developed based on the availability of variables and the judgment of the statisticians (Madow, Nisselson, and Olkin Vol. 1, ch. 4, 1983). Predictors of response, key outcome statistics, and domains are considered in this process. Demographic variables such as age, sex, race, and geography were, and still are, frequently chosen even though they may not be effective in reducing bias (Peytcheva and Groves 2009). Many of these are population-based adjustments using data from a recent census for the controls. When the number of respondents in a cell of the cross-classification of the variables is below a threshold set for the survey, then cells are collapsed to avoid large adjustment factors.

When many variables are available, other methods of choosing which variables to include are needed. Search algorithms and regression models are sometimes used in this setting (Brick and Kalton 1996). These methods divide the sample into cells that discriminate between response and nonresponse or variables correlated with key outcome variables. The main advantage of these methods, especially the search algorithms, is the ability to identify interactions among the variables that may be important for nonresponse reduction. Regression models can also be used to examine interactions, although practitioners often rely on main effect models. Brick and Jones (2008) show the importance of interactions in some situations.

New methods for choosing auxiliary variables to reduce nonresponse bias in the estimates have been recently developed. Schouten (2007) and Särndal and his colleagues (Särndal and Lundström 2005, 2008, 2010; Särndal 2011a) suggest two approaches. These approaches do not assume that the data are missing at random, but to be effective they do require powerful predictors of the response mechanism. The methods are also described in terms of searching for main effects and including or excluding variables. Extensions are needed to deal with interactions among the variables.

Schouten et al. (Schouten 2007; Schouten et al. 2009) use indicators for choosing variables for weighting that are related to R-indicators. Schouten (2007) gives a forward-backward selection strategy for choosing variables, similar to stepwise regression. He starts with the variable that minimizes an estimate of maximal bias (which is linked to the R-indicator) and iteratively adds and removes other variables. The maximal bias is computed based on a generalized regression estimator.

Särndal and Lundström (2008) approach the choice of auxiliary variables by focusing on the estimation phase, although they are explicit about the importance of the design and data collection stages also (see Särndal and Lundström 2010; Särndal 2011a). Särndal and Lundström (2010) propose survey-specific indicators that account for the sample design, the set of observed respondents, and the specific calibration estimator. Their indicators choose auxiliaries based on the distance between the calibrated estimator and the primitive calibration estimator ($\hat{y}_{ca}$ and $\hat{y}_{pr}$) and may be outcome specific or generic. These authors describe an "all vectors procedure" that chooses the auxiliaries that are in the list of vectors that has the highest indicator. They also offer a "stepwise" procedure that builds the vector one variable at a time.

Särndal and Lundström (2010) compare the two approaches and find that they do not always include the same set of auxiliary variables in the estimator. They attribute some of the difference to the different perspectives, especially the fact that Schouten's (2007) approach uses population-level measures while theirs are sample-level. When choosing among many possible auxiliary variables to include or exclude in the estimation phase, the indicators of Särndal have the advantage of assessing improvements in estimators for the specific sample.

In some countries, especially in northern Europe, population registers may provide the types of data needed for using these methods. However, in household surveys in countries like the United States and Canada, these methods are less pertinent because there are few powerful auxiliary variables. When the information available for the sample does not predict response well, researchers have resorted to creating paradata from the survey itself (Beaumont 2005; Bates et al. 2008). The use of paradata is a rapidly developing area, but initial findings reveal that this may be a difficult task (Kreuter et al. 2010).

7. Response Propensity Models in Surveys

Because response propensity scores play such a large role in nonresponse adjustment methods, we describe the underlying theory and assumptions here. We begin with a few observations. First, response propensities are unknown, unlike probabilities associated with an additional phase of sampling. In fact, they are latent variables and cannot be observed directly – we observe only the binary outcome of response or nonresponse. Second, we assume that $\phi_i > 0$ for all i. Deming (1953) explicitly considers units with zero response propensities. He calls those that never participate "permanent refusers." Third, as Brick and Montaquila (2009) note, response propensities are specific to both the units sampled and the survey conditions. The same units may have different response propensities depending on key survey conditions. The survey conditions may be manipulated to increase response rates during data collection.

Rosenbaum and Rubin (1983) provide the framework for the application of propensity scores in observational studies for estimating causal effects. In observational studies, propensity scores are used to approximate unbiased estimates of the average effect of a treatment (the difference in outcomes between those subject to a treatment and those not treated) when the treatment assignment is not randomized. Rosenbaum and Rubin show that the propensity score is the coarsest balancing score and that, at any value of a balancing score, the average treatment effect can be estimated without bias when certain assumptions hold.

Response propensity theory has been used in a wide variety of applications, including survey nonresponse adjustment. Little (1986) applied propensity score theory to surveys, primarily utilizing the property that propensity score is the coarsest balancing score. In surveys, all sampled units are subject to a data collection protocol – as opposed to the observational setting where units are subject to more than one treatment (one of which may be the null treatment). In surveys, the response propensities are primarily used to form groups to satisfy the MAR. In terms of propensity scores, MAR implies that

$$\Pr(R_i = 1|Y_i, \mathbf{X}_i) = \Pr(R_i = 1|\phi(\mathbf{X}_i)). \tag{12}$$

Thus, by conditioning on the groups based on estimated response propensities, we hope to be able to justify the assumption that missing data are independent of the outcome characteristic. The response propensity score is just the dimension-reducing function that facilitates using multiple auxiliary variables in forming groups.

David et al. (1983) outline a structure using the framework of Rosenbaum and Rubin (1983) and define the treatment as the survey response and the outcome as the characteristic being estimated. In observational studies, we are interested in differences in outcomes when subjects self-select into different treatments and outcomes are observed for those with different treatments. In surveys, we do not observe outcomes for those who do not respond. Despite this difference, David et al. (1983) use this structure only to take advantage of theorems of Rosenbaum and Rubin (1983) showing that the propensity score has the dimension-reducing property.

Two assumptions in Rosenbaum and Rubin's (1983) development are the strongly ignorable treatment assignment assumption and the stable unit treatment value assumption (SUTVA). The strongly ignorable assumption roughly translates into the MAR assumption in the survey context, and it is considered in most applications of propensity scores in nonresponse adjustment. In many cross-sectional household surveys, the lack of powerful predictors means that the strongly ignorable or MAR assumption is tenuous. Of course, the effectiveness of propensity scores to satisfy the MAR assumption is bounded by the power of the auxiliary data used to create the score. Researchers appreciate this limitation and have sought to find better variables or to collect them using paradata.

The second key assumption in propensity score theory, SUTVA, is rarely discussed in the nonresponse adjustment literature. In observational studies, SUTVA is sometimes summarized as a lack of interference between units. One way to translate this into the survey situation is to state that the response propensities of the sampled units are not affected by those of other units, at least within the subsets or groups of units used to estimate the propensities. The typical approach to estimate propensities is to assume that

the response for a sampled unit is independent of responses for other units. For the multistage, clustered samples used in many household surveys, this practice seems problematic. For example, interviewers in face-to-face surveys are typically clustered in areas to reduce travel costs. There is ample evidence showing that interviewers and supervisors may influence response (Groves and Couper 1998). Skinner and D'Arrigo (2011) use multilevel models and find some bias in estimates of response propensities that ignore clustering. They suggest using conditional maximum likelihood for estimating propensities rather than the standard logistic modeling. They see the problem as a failure to satisfy the strong ignorability assumption rather than the SUTVA. Other examples are clearer failures of SUTVA, such as when sampling more than one adult per household or multiple teachers from a school. In this case, the sampled units may influence other sampled units directly.

Finally, an issue we think is likely to have even greater importance is related to the definition of the propensity in the nonresponse setting. The propensity is often treated as a fixed attribute of a sampled unit. This conceptualization of response propensities prompted Dalenius (1983, p. 412) to take a "dim view" on estimating response propensities because "it appears utterly unrealistic to postulate fixed response probabilities which are independent of the varying circumstances under which an effort is made to elicit a response." In large measure, we agree and believe a more refined definition of response propensities is needed.

We prefer to express the propensity so that the survey conditions are explicit, such as

$$\phi_i = \phi(a_i, \mathbf{X}_i) = \phi(a_{i1}, a_{i2}, \ldots, \mathbf{X}_i) = \Pr\left(R_i = 1 | \mathbf{a}_i, \mathbf{X}_i\right), \tag{13}$$

where the effort or activity vector (**a**) indicates the relevant data collection activities. The components of the activity vector encompass all forms of data collection, such as the number of call attempts, the use of incentives, the modes of data collection, and refusal conversion attempts. Schouten et al. (2011b) and Olsen and Groves (2012) are also explicit about including fieldwork as well as other variables known for all sampled units when defining the propensity. The quantity that should be estimated to create a nonresponse adjustment factor is $\phi'(\mathbf{a}_i, \mathbf{X}_i)$, where the prime denotes the actual activities at the end of data collection. Defining the propensities as in (13) does not simplify the task, but at least it better defines the quantity being estimated.

Olsen and Groves (2012) and Schouten et al. (2011b) both postulate that response propensities are dynamic, with the response propensity of a sampled unit varying as the recruitment protocol changes. They show that response propensities are influenced by the data collection protocol. In our terminology, they demonstrate that the response propensities are not constant when at least some components of **a** are altered.

Olsen and Groves (2012) also plot conditional response propensities and show that these decline over the field period during which a stable data collection protocol is in place. They argue that this decline implies that the individual's response propensity decreases over repeated applications of the same recruitment protocol. While their explanation is consistent with our perception and with the discussion in Schouten et al. (2011b), there is an alternative explanation that highlights our concern about the unobservable nature of response propensities. Assume that the persons in the sample are

members of two different RHGs, with 70 percent of the sample having fixed response propensities of 0.4 and 30 percent having propensities of 0.2. The dotted lines in Figure 1 show the constant propensities over the data collection (effort) for each of the RHGs, and the solid line shows the decreasing propensity of the entire sample. The solid line approximates the shape observed by Olsen and Groves (2012), suggesting that combining RHGs with different propensities could produce the effect they observed even though the conditional response propensities for individuals are constant. Because the response propensities are unknown even after data collection, it is impossible to assess whether the propensities are changing or whether we are mixing groups with different, but constant, response propensities.

8. Response Propensities and Data Collection

The importance of the connection between data collection and nonresponse adjustments can be illustrated by simple examples. We begin with an example inspired by Olsen and Groves (2012). A sample is selected and a standard data collection protocol is applied to all sampled units; some units respond at the end of the first phase of data collection. For the second phase, a subsample of nonrespondents is selected and given a new protocol (e.g., a large incentive, more highly trained interviewers, a different mode), which increases response. We assume that all the units in the sample have identical response propensities, ϕ (**a**, **X**), but that only those in the subsample are given the additional effort.

One approach to estimation (Approach A) is to exclude those units not in the second-phase subsample; weight the first-phase respondents by the inverse of their selection

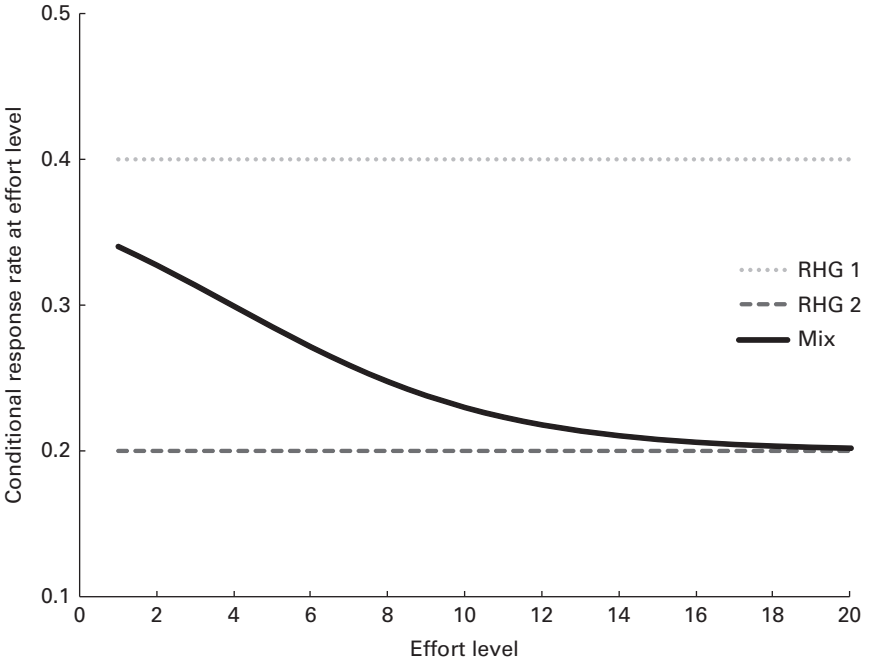


Fig. 1. Observed response propensities for a sample composed of two RHGs

probabilities, d_i; and weight the second-phase respondents by d_i times the inverse of the product of the subsampling rate and the response rate within the subsample. For example, if half of the nonrespondents are subsampled and 40 percent of these respond, the weight for the second-phase respondents would be $5d_i$ ($5 = .5^{-1} \cdot .4^{-1}$ is the adjustment factor). Under the sample-response mechanism, this estimator is unbiased.

In practice, it may be tempting to use an alternative Approach B that computes the nonresponse adjustment across respondents to both the first phase and the second phase to reduce the nonresponse adjustment factor and its impact on the variance of the estimates. In this case, all respondents get the same final weight – d_i times the inverse of the response rate, where the response rate is computed over the entire sample rather than the subsample. Essentially, this estimator ignores the subsampling. The Approach B estimator is biased if the characteristics of the second-phase respondents differ from those of the first-phase respondents. The problem is that the Approach B estimator combines two groups that have different response propensities at the end of data collection. In other words, while all the sampled units have the same $\phi(\mathbf{a}, \mathbf{X})$, they have different values of $\phi'(\mathbf{a}, \mathbf{X})$ because the activity vectors are not identical for the first- and second-phase units. The MAR assumption holds only when the groups are defined by the data collection activity.

Now consider a slightly revised example with the same structure. Suppose we want to estimate the proportion with a characteristic ($y_i = 1$), and assume that the units with $y_i = 1$ have a response rate of 60 percent at the end of the first phase while units with $y_i = 0$ have a first-phase response rate of 40 percent. This is a classic example of topic salience bias. We assume that no auxiliary data are available to identify those with and without the characteristic. A second-phase protocol is implemented by giving *all* nonrespondents an incentive, and the conditional response rate for the second phase is 60 percent for those with $y_i = 1$ and 50 percent for those with $y_i = 0$. The two adjustment methods used above are applied; Approach A computes the nonresponse adjustment factor over just the second-phase respondents (there is no subsampling here); Approach B computes it over all respondents. Figure 2 shows the bias associated with two adjustment approaches. Neither method eliminates the bias completely because the additional phase does not eliminate the difference in the response rates between units with $y_i = 1$ and $y_i = 0$. Thus, this is an example of NMAR. However, Approach A produces estimates that are less biased in this situation because the difference in rates or response propensities is reduced by the second phase of data collection. This result is not always obtained, as discussed below.

In both examples, the data collection activities applied to the units affect the response propensities at the end of data collection. In the first example, the response propensities for all the units are identical but the adjustment groups must be defined by phase for MAR to hold. In the second example, the response propensities differ for those with and without the characteristic, and we must "know" that the incentive applied at the second phase reduces the differences in response rates for these groups to justify using Approach A. We can observe that the percentage with $y_i = 1$ is greater in the second phase than the first phase, but there is no test to show that this reduces bias (an example below has the opposite effect). Of course, the rationale for the second-phase incentive "should" have been that it would reduce bias, otherwise it is hard to justify its application to the second-phase data collection protocol. Unfortunately, in many surveys these factors are not fully considered in data collection, and the main concern is increasing the overall response rate.

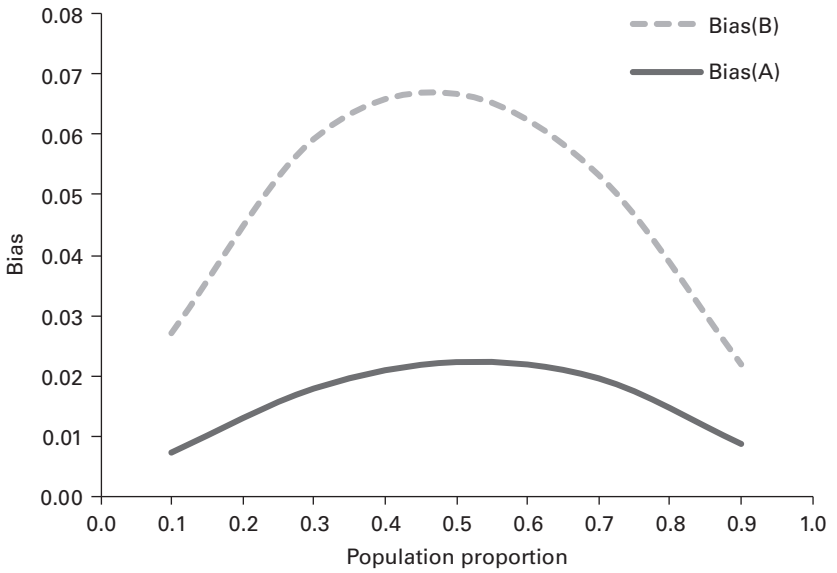


Fig. 2. Bias in estimated proportion using two adjustment methods

These hypothetical examples may seem overly simple or unrealistic, so some examples from real surveys are presented. We begin with two examples with desirable outcomes. The first is taken from Mohadjer et al. (1997). They show that providing incentives in an adult literacy survey improved responses more from low education and minority adults, resulting in reduced bias in key outcomes such as literacy scores by race and education level. A second example is provided by Groves and Heeringa (2006), who offered incentives in the second phase of a responsive design. They too show differential improvements in response rates and reduced nonresponse bias for some statistics.

Examples of changes in data collection protocols that have little or no effect on the estimates appear to be more numerous. This suggests that these results are surprising because publishing null results is generally difficult. One of the first of the recent examples of this genre is Keeter et al. (2000), who substantially increased response rates in a telephone survey by increasing the level of effort (number of call attempts, length of data collection period, etc.). Despite the higher response rates, however, almost none of the estimates from the survey changed significantly. Similar outcomes have been observed numerous times, for example by Curtin et al. (2005), Haring et al. (2009), and Ingen et al. (2009). While there are several possible explanations for the lack of an effect on the estimates, these examples point out gaps in our understanding of the effects of data collection efforts on biases.

There are also examples of data collection efforts that increase both response rates and nonresponse bias. Wetzels et al. (2008) document a survey where incentives increased response rates and had little effect on most estimates. They also report that response rates of non-Western foreigners did not increase with the use of incentives, possibly increasing the biases of estimates related to this subgroup. Merkle et al. (1998) describe an experiment with incentives in an exit poll survey where increased response rates were

accompanied by increased nonresponse bias. They suggest that the incentives appealed differentially to voters by party. Schmeets (2010) examines changes in data collection procedures to increase response rates for the Dutch Parliamentary Election Study. He concludes that the changes increased the survey response rates but also might have increased the bias for some of the estimates.

These examples lead us to consider how we can use effort data from the survey to form RHGs for adjustment purposes. Clearly, the data collection activities do not have to be the same for all units; rather, the objective is to classify units with the same final response rate into a RHG irrespective of how they get to this final state.

The traditional approach to forming groups is to use the auxiliary data to identify groups with different response propensities by logistic regression models. Olsen and Groves (2012) suggest using discrete hazard models because response propensities vary over the data collection period. Because the goal is to identify groups of sampled units that have the same value of $\phi'(\mathbf{a}_i, \mathbf{X}_i)$ at the end of data collection, we believe hazard models might be valuable only when the sequencing of data collection activities is important to the response process. The Skinner and D'Arrigo (2011) findings indicate that conditional maximum likelihood estimation might account for clustering.

A perhaps more important realization is that, for most surveys, regression models may not be useful in assigning sampled units to RHGs based on data collection effort. For example, suppose all the units in the sample have the same response propensity for a three-contact data collection protocol. Some units respond at each contact level and some do not respond after all three contacts. If we model response based on the number of calls it took to get a response, we would form RHGs giving different adjustment factors to the respondents by the number of calls it took to respond. These RHGs would only increase the variation in the weights and could, in some situations, introduce bias. Contrast this with the first hypothetical example given above, where bias is reduced by weighting only those units given additional effort. Why should we not adjust the weights only for the cases that responded on the third call? The difference is that we assume in the three call example that the response propensities at the end of the protocol are the same regardless of when the unit responded. The data themselves do not inform us which assumption is correct. Modeling of effort will not reveal this. We would argue that if we subsampled nonrespondents at the end of the first contact, then forming RHGs based on effort would be appropriate in most surveys. The most troubling fact is that the real examples cited above show that we do not always know which assumptions are most reasonable. Although forming RHGs with logistic regression models based on $\mathbf{X}$ is valuable, modeling based on data collection activity may not be as effective without a more complete theory of response.

9. Discussion

As we have mentioned several times, there is a substantial literature that shows the effectiveness of data collection strategies for enhancing response rates. Such strategies include changing modes of data collection, providing incentives, and converting reluctant respondents. When these strategies reduce nonresponse bias, however, is less clear. Without a better understanding of these effects, it is difficult to design effective data collection and estimation strategies to combat nonresponse bias for surveys.

Responsive and adaptive designs have been proposed as a way to reduce nonresponse bias, but these are predicated on making changes to data collection strategies either during data collection (Groves and Heeringa 2006) or from analyses of response patterns in previous collections (Schouten et al. 2011a). Because these designs implement data collection protocols that may vary at the sample case level, they require a refined understanding of the effects these data collection protocols have on nonresponse bias. These types of designs have the potential to increase nonresponse bias if the design, data collection, and estimation stages are not fully integrated.

For example, suppose increased effort is given to some sampled units identified during data collection based on paradata collected in the initial contacts. How should this be handled in forming RHGs? Should units getting extra effort be identified as separate RHGs, or should we assume that the extra effort for those units equalizes response rates so that separate groups are not needed? The answer depends on the assumptions made about the effect of the efforts on nonresponse bias. Surveys that use responsive or adaptive designs need to explain the rationale for their nonresponse adjustment procedures sufficiently so that others can assess the assumptions underlying their estimation methods.

The central problem, in our opinion, is that even after decades of research on nonresponse we remain woefully ignorant of the causes of nonresponse at a profound level. This may be a harsh critique given all the progress we have made in many areas. We better understand methods to reduce nonresponse due to noncontact in surveys and have made substantial strides in this area. We also have a much better understanding of correlates of nonresponse. Over time, studies have replicated the correlations between demographic and geographic variables and nonresponse rates (e.g., Groves and Couper 1998; Stoop et al. 2010). These are important developments but have not led to a profound understanding of the causes of nonresponse.

Stoop (2005) reviews some of the areas of research on noncooperation in surveys, but her review shows few lessons that can be generalized and used to reduce nonresponse bias. For example, some research has shown that certain types of people – outgoing and altruistic people – seem to cooperate in surveys more than others. However, utilizing these findings to mitigate nonresponse bias remains a challenge. Another example is the practice of asking people why they refuse to participate in surveys. These requests produce uninformative responses such as being "too busy," and the distribution of these responses has been constant for years (Brick and Williams 2013). Even though we know that sampled units will never be able to answer our analytic questions about the response process directly, we continue to ask these questions. To better understand the response process we need to reformulate our approach, use less direct questions, and ask both respondents and nonrespondents similar items to support comparative analysis (Singer and Ye 2013).

Some research approaches do appear to have promise and could lead to improvements in our practices and our understanding. For example, if we can increase the perceived value of the survey to the respondent and make the response process simple and enjoyable, then we could potentially lower nonresponse bias (e.g., Dillman et al. 2009). Additional research into ways of increasing the value and making the process more enjoyable is needed. Another promising development is by Groves et al. (2006), who report on an innovative approach to try to generate nonresponse bias in surveys by manipulating factors

thought to be related to nonresponse bias. Many practitioners were surprised that their results showed less bias than might have been expected. The idea of prospectively manipulating factors in a controlled manner could increase our understanding of the response mechanism.

One of the difficulties preventing a deeper understanding of nonresponse in surveys is the complexity of the survey process. Many factors in a survey contribute to complexity and may affect nonresponse. These factors include the target population, sponsorship, survey content, interviewer training and experience, mode of data collection, incentives, length of interview, the available field period, and regulatory limitations. Complex systems are inherently more difficult to analyze than simple ones.

One of the ways that other sciences have made progress in studying complex systems is to conduct basic research, often in a laboratory setting, to isolate important main effects. Survey research seems to lack that type of basic research. The exception is statistical design and estimation work that is not as constrained as data collection. Nearly all survey research is empirical, and most of our knowledge comes from experiences in specific surveys. This makes it harder to generalize the findings.

Cognitive research methods were originally introduced into surveys with some of these issues in mind. Over time, this movement has largely devolved into a set of tools to improve questionnaires. Tanur (1999) reviews the origins and evolution of cognitive research in surveys. Today, there are few, if any, settings or laboratories where survey methodologists and psychologists can postulate and explore response theories without being tethered to the needs of a particular survey. The reasons that the cognitive movement has gone in this direction seem clear in hindsight: The research is situated in survey organizations, and those organizations need to justify the allocation of scarce resources. As a result, the application to specific surveys is a higher priority than basic research.

Perhaps the time is ripe for new approaches to the vexing and important question of why people do and do not respond to surveys. Interdisciplinary and basic research may prove profitable if the structural issues can be addressed. But substantive progress cannot be guaranteed by any single approach. Research on making the process more respondent friendly, experiments to induce nonresponse bias, and comparative analysis of respondents and nonrespondents using indirect assessments of attributes of response may have merit. Until we have methods to better understand the relationships between survey requests and response, we are unlikely to be able to structure survey designs, data collection protocols, and estimation schemes to minimize nonresponse bias.

10. References

Andridge, R.H. and Little, R.J. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. Journal of Official Statistics, 27, 153–180.

Atrostic, B.K., Bates, N., Burt, G., and Silberstein, A. (2001). Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights. Journal of Official Statistics, 17, 209–226.

Bartholomew, D.J. (1961). A Method of Allowing for 'Not-at-Home' Bias in Sample Surveys. Applied Statistics, 10, 52–59.

Bates, N., Dahlhamer, J., and Singer, E. (2008). Privacy Concerns, too Busy, or Just not Interested: Using Doorstep Concerns to Predict Survey Nonresponse. Journal of Official Statistics, 24, 591–612.

Beaumont, J.F. (2005). On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment. Survey Methodology, 31, 227–231.

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias Through Regression Estimation. Journal of Official Statistics, 4, 251–260.

Bethlehem, J.G. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). Handbook in Nonresponse in Household Surveys. New York: Wiley.

Brehm, J. (1993). The Phantom Respondents: Opinion Surveys and Political Representation. Ann Arbor: University of Michigan Press.

Brick, J.M. and Jones, M.E. (2008). Propensity to Respond and Nonresponse Bias. Metron-International Journal of Statistics, LXVI, 51–73.

Brick, J.M. and Kalton, G. (1996). Handling Missing Data in Survey Research. Statistical Methods in Medical Research, 5, 215–238.

Brick, J.M. and Montaquila, J.M. (2009). Nonresponse and Weighting. Handbook of Statistics. Sample Surveys: Design, Methods, and Applications, D. Pfeffermann and C.R. Rao (eds). Vol. 29A. Amsterdam: Elsevier-North Holland, 163–186.

Brick, J.M., Montaquila, J., Han, D., and Williams, D. (2012). Improving Response Rates for Spanish-Speakers in Two-Phase Mail Surveys. Public Opinion Quarterly, 76, 721–732.

Brick, J.M. and Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. The ANNALS of the American Academy of Political and Social Science, 645, 36–59.

Cassel, C., Särndal, C.-E., and Wretman, J. (1983). Some Uses of Statistical Models in Connection With the Nonresponse Problem. Incomplete Data in Sample Surveys, W.G. Madow and I. Olkin (eds). Vol. 3. New York: Academic Press.

Chang, T. and Kott, P.S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika, 95, 557–571.

Cochran, W. (1977). Sampling Techniques, (3rd edition). New York: Wiley.

Colley, R.H. (1945). Don't Look Down Your Nose at Mail Questionnaires. Printers' Ink, March, 16, 21–108.

Curtin, R., Presser, S., and Singer, E. (2000). The Effects of Response Rate Changes on the Index of Consumer Sentiment. Public Opinion Quarterly, 64, 413–428.

Curtin, R., Presser, S., and Singer, E. (2005). Changes in Telephone Survey Nonresponse Error Over the Past Quarter Century. Public Opinion Quarterly, 69, 87–98.

Da Silva, D.N. and Opsomer, J.D. (2004). Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism. Survey Methodology, 30, 45–55.

Da Silva, D.N. and Opsomer, J.D. (2009). Nonparametric Propensity Weighting for Survey Nonresponse Through Local Polynomial Regression. Survey Methodology, 35, 165–176.

Dalenius, T. (1983). Some Reflections on the Problem of Missing Data. Incomplete Data in Sample Surveys, W.G. Madow and I. Olkin (eds). Vol. 3. New York: Academic Press, 411–413.

David, M., Little, R., Samuhel, M., and Triest, R. (1983). Nonrandom Nonresponse Models Based on the Propensity to Respond. Proceedings of the Business and Economic Statistics Section of the American Statistical Association, 168–173.

David, M., Little, R.J.A., Samuhel, M., and Triest, R. (1986). Alternative Methods for CPS Income Imputation. Journal of the American Statistical Association, 81, 29–41.

De Leeuw, E. and De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 41–54.

Deming, W. (1953). On a Probability Mechanism to Attain an Economic Balance Between Resultant Error of Response and the Bias of Nonresponse. Journal of the American Statistical Association, 48, 743–772.

Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Dillman, D. (1978). Mail and Telephone Surveys: The Total Design Method. New York: Wiley.

Dillman, D., Smyth, J., and Christian, L. (2009). Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, (3rd edition). New York: Wiley.

Dunkelburg, W. and Day, G. (1973). Nonresponse Bias and Callbacks in Sample Surveys. Journal of Marketing Research, 10, 160–168.

Ferber, R. (1949). The Problem of Bias in Mail Returns: A Solution. Public Opinion Quarterly, 12, 669–676.

Feskins, R., Hoop, J., Lensvelt-Mulders, G., and Schmeets, H. (2011). Collecting Data Among Ethnic Minorities in an International Perspective. Field Methods, 18, 284–304.

Fuller, W.A., Loughin, M.M., and Baker, H.D. (1994). Regression Weighting for the 1987-88 National Food Consumption Survey. Survey Methodology, 20, 75–85.

Goyder, J. (1987). The Silent Minority: Nonrespondents on Sample Surveys. Boulder, CO: Westview Press.

Greenlees, J., Reece, W., and Zieschang, K. (1982). Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed. Journal of the American Statistical Association, 77, 251–261.

Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. Public Opinion Quarterly, 70, 646–675.

Groves, R.M. and Couper, M.P. (1998). Nonresponse in Household Interview Surveys. New York: Wiley.

Groves, R.M., Couper, M., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P., and Nelson, L. (2006). Experiments in Producing Nonresponse Bias. Public Opinion Quarterly, 70, 720–736.

Groves, R., Dillman, D., Eltinge, J., and Little, R. (2002). Survey Nonresponse. New York: Wiley, 41–54.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. Journal of the Royal Statistical Society, Series A, 169, 439–457.

Hansen, M.H. and Hurwitz, W.N. (1946). The Problem of Non-Response in Sample Surveys. Journal of the American Statistical Association, 41, 517–529.

Haring, R., Alte, D., Völzkea, H., Sauer, S., Wallaschofski, H., John, U., and Schmidt, C. (2009). Extended Recruitment Efforts Minimize Attrition but not Necessarily Bias. Journal of Clinical Epidemiology, 62, 252–260.

Hartley, H.O. (1946). Discussion of "A Review of Recent Statistical Developments in Sampling and Sample surveys.". Journal of the Royal Statistical Society, 109, 37–38.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47, 153–162.

Holt, D. and Smith, T.M.F. (1979). Post-Stratification. Journal of the Royal Statistical Society, Series A, 142, 33–46.

Ingen, E., Stoop, I., and Breedveld, K. (2009). Nonresponse in the Dutch Time Use Survey: Strategies for Response Enhancement and Bias Reduction. Field Methods, 21, 69–90.

Kalton, G. (1983). Compensating for Missing Survey Data. Ann Arbor: University of Michigan Press.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. Journal of Official Statistics, 18, 81–97.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a Large National Telephone Survey. Public Opinion Quarterly, 64, 125–148.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. Journal of the Royal Statistical Society, Series A, 173, 389–407.

Lin, I.-F. and Schaeffer, N.C. (1995). Using Survey Participants to Estimate the Impact of Nonparticipation. Public Opinion Quarterly, 59, 236–258.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Little, R.J.A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association, 88, 125–134.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis With Missing data, (2nd edition). New York: Wiley.

Lumley, T., Shaw, P., and Dai, J. (2011). Connections Between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. International Statistical Review, 79, 200–220.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. Journal of Official Statistics, 15, 305–327.

Madow, W.G., Nisselson, H., and Olkin, I. (1983). Incomplete Data in Sample Surveys, Vol. 1. New York: Academic Press.

Madow, W.G. and Olkin, I. (1983). Incomplete Data in Sample Surveys, Vol. 3. New York: Academic Press.

Madow, W.G., Olkin, I., and Rubin, D.B. (1983). Incomplete Data in Sample Surveys, Vol. 2. New York: Academic Press.

Merkle, D., Edelman, M., Dykeman, K., and Brogan, C. (1998). An Experimental Study of Ways to Increase Exit Poll Response Rates and Reduce Survey Error. Paper presented at the Annual Conference of the American Association for Public Opinion Research, St. Louis, MO.

Micklewright, J., Schnepf, S., and Skinner, C. (2012). Non-Response Biases in Surveys of Schoolchildren: The Case of the English Programme for International Student Assessment (PISA) samples. Journal of the Royal Statistical Society, Series A, 175, 915–938.

Mohadjer, L., Berlin, M., Rieger, S., Waksberg, J., Rock, D., Yamamoto, K., Kirsch, I., and Kolstad, A. (1997). The Role of Incentives in Literacy Survey Research. Adult Basic Skills: Innovations in Measurement and Policy Analysis, A. Tuijnman, I. Kirsch, and D. Wagner (eds). Creskill, NJ: Hampton Press.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008). Every Missingness not at Random Model has a Missingness at Random Counterpart With Equal Fit. Journal of the Royal Statistical Society: Series B, 70, 371–388.

Oh, H.L. and Scheuren, F.J. (1983). Weighting Adjustments for Unit Nonresponse. Incomplete Data in Sample Surveys, W.G. Madow, I. Olkin, and D.B. Rubin (eds). Vol. 2. New York: Academic Press, 143–184.

Olsen, K. and Groves, R.M. (2012). An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period. Journal of Official Statistics, 28, 29–51.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

Phipps, P. and Toth, D. (2012). Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data. Annals of Applied Statistics, 6, 772–794.

Politz, A. and Simmons, W. (1949). An Attempt to Get "Not at Homes" Into the Sample Without Callbacks. Journal of the American Statistical Association, 44, 9–31.

Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41–55.

Rubin, D.B. (1976). Inference and Missing Data (with discussion). Biometrika, 63, 581–592.

Särndal, C.-E. (2011a). Morris Hansen Lecture: Dealing With Survey Nonresponse in Data Collection, in Estimation. Journal of Official Statistics, 27, 1–21.

Särndal, C.-E. (2011b). Three Factors to Signal Non-Response Bias with Applications to Categorical Auxiliary Variables. International Statistical Review, 79, 233–254.

Särndal, C.-E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester, UK: Wiley.

Särndal, C.-E. and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. Journal of Official Statistics, 4, 251–260.

Särndal, C.-E. and Lundström, S. (2010). Design for Estimation: Identifying Auxiliary vectors to reduce nonresponse bias. Survey Methodology, 36, 131–144.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Schmeets, H. (2010). Increasing Response Rates and the Consequences in the Dutch Parliamentary Election Study 2006. Field Methods, 22, 391–412.

Schouten, B. (2007). A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption. Journal of Official Statistics, 23, 51–68.

Schouten, B., Calinescu, M., and Luiten, A. (2011a). Optimizing Quality of Response Through Adaptive Survey Designs. The Hague: Statistics Netherlands, Available at: http://www.cbs.nl/NR/rdonlyres/2D62BF4A-6783-4AC4-8E4512EF20C6675C/0/2011x1018.pdf. (Accessed May 24, 2013).

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Measures for the Representativeness of Survey Response. Survey Methodology, 35, 101–113.

Schouten, B., Schlomo, N., and Skinner, C. (2011b). Indicators for Monitoring and Improving Representativeness of Response. Journal of Official Statistics, 27, 231–253.

Singer, E. (2002). Use of Incentives to Reduce Nonresponse in Household Surveys. Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley, 163–177.

Singer, E. and Ye, C. (2013). The Use and Effects of Incentives in Surveys. The ANNALS of the American Academy of Political and Social Science, 645, 112–141.

Skinner, C.J. and D'Arrigo, J. (2011). Inverse Probability Weighting for Clustered Nonresponse. Biometrika, 98, 953–966.

Smith, T.W. (1995). Trends in Non-Response Rates. International Journal of Public Opinion Research, 7, 157–171.

Steeh, C., Kirgis, N., Cannon, B., and DeWitt, J. (2001). Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century. Journal of Official Statistics, 17, 227–247.

Steele, F. and Durrant, G.B. (2011). Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal. International Statistical Review, 79, 70–91.

Stoop, I.A.L. (2005). The Hunt for the Last Respondent: Nonresponse in Sample Surveys. The Hague: Social and Cultural Planning Office.

Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving Survey Response: Lessons Learned from the European Social Survey. Chichester: Wiley.

Synodinos, N.E. and Yamada, S. (2000). Response Rate Trends in Japanese Surveys. International Journal of Public Opinion Research, 12, 48–72.

Tanur, J. (1999). Looking Backwards and Forwards at the CASM Movement. Cognition and Survey Research, M. Sirken, D. Hermann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds). New York: Wiley, 13–20.

Thomsen, I. (1973). A Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Nonresponse When Analyzing Survey Data. Statistisk Tidskrift, 11, 278–285.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). The Psychology of Survey Response. New York: Cambridge University Press.

Wagner, J. (2010). The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. Public Opinion Quarterly, 74, 223–243.

Wetzels, W., Schmeets, H., Van den Brakel, J., and Feskens, R. (2008). Impact of Prepaid Incentives in Face-to-Face Surveys: A Large-Scale Experiment With Postage Stamps. International Journal of Public Opinion Research, 20, 507–516.

Yates, F. (1946). A Review of Recent Statistical Developments in Sampling and Sample Surveys. Journal of the Royal Statistical Society, 109, 12–43.

Accepted March 2013

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 355–358

Discussion

Olena Kaminska[1]

1. Introduction

The article by Michael Brick comes at a time when the survey methodology field is actively looking for solutions to constantly decreasing response rates. After a number of decades developing design features for achieving higher response rates, and therefore unintentionally educating our clients and funders that response rates are important, we are now struggling to explain the importance of nonresponse bias. But what is more challenging is to understand ourselves how we can deal with nonresponse bias in the best way.

I found the article to be a much needed reminder to the field of the gaps in our knowledge about nonresponse bias today, and how much is to be developed in order to identify best practice in dealing with nonresponse. The work is both comprehensive and current with a historical overview of research into nonresponse, and identification of areas with unanswered issues, and areas with the potential to answer pressing questions.

I enjoyed reading about recent developments in the field of nonresponse that directly refer to nonresponse bias, instead of response rate. Brick first reviews adaptive or responsive design that tailors data collection in order to decrease nonresponse bias. One attraction of such designs is the idea of tailoring fieldwork procedures in response to information obtained before or during the fieldwork. Yet to me the biggest value of such an approach is that for the first time we are developing design with an explicit aim of decreasing nonresponse bias. Adaptive and responsive designs do not have to be the only designs with such an aim; and as the author suggested, we should review already developed design features with respect to their influence on nonresponse bias. We know that incentives increase response rates (e.g., Singer et al. 2000; Singer et al. 1999), but do they also decrease nonresponse bias? We know that mentioning a salient topic of the survey may increase response rate (e.g., Groves et al. 2004), but does this decrease nonresponse bias? Questions like these require answers in order to tailor our practice to decreasing nonresponse bias directly, rather than through increasing response rate alone.

Another important development mentioned is the collection of new paradata which should give stronger predictors for the adjustment stage. While weighting for nonresponse is hoped to be a 'solution' to nonresponse bias, it largely depends on good correlates of nonresponse and of y-variables (more precisely, of estimates of substantive interest). Often little information is available on both respondents and nonrespondents; and gathering additional information that can be used in nonresponse adjustment models and that is

[1] ISER, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK. Email: olena@essex.ac.uk

tailored to important *y*-variables has direct impact on the quality of nonresponse adjustment. While little resources tend to be put into collecting paradata in comparison to large resources for converting reluctant respondents, it is possible that the reverse would be most beneficial for reducing nonresponse bias in final estimates.

A more complex development suggested by the author is an integration of three practices that largely have been developing autonomously so far: research into causes of nonresponse, development of design features to decrease nonresponse bias, and adjusting for nonresponse. For example, from a fieldwork perspective, responsive design is a very attractive set of procedures which in the end should result in minimal nonresponse bias on selected variables. Yet such a design, having differential selection and nonresponse probabilities, may lead to an increase in standard errors of estimates which can outweigh the gains from bias. While this is theoretically possible, little is known about such interaction at the moment. Thinking about both design features and nonresponse adjustment in this example would pose these questions earlier, and will challenge the development of designs that optimize collection and adjustment simultaneously.

With the above said, I feel that the literature on survey weighting is particularly in need of development in order to answer the questions being raised by the innovations in data collection procedures. Weighting has largely developed in the previous century for a one-time cross-sectional study of one population and for one survey protocol. Michael Brick's article is one of very few attempts today to develop the best weighting approach for a situation which differs from that above: a situation where the survey protocol changes during data collection. This includes two-stage design, where only some nonrespondents are attempted in the second stage, responsive design, or a design with increasing incentives in the later stages of the fieldwork. In my discussion, I comment on response probabilities in such situations and point out an alternative weighting procedure to account for selection probabilities and nonresponse.

2. Do Response Probabilities Change with Fieldwork?

This is one of the questions raised by Michael Brick in the article (Section 6). In my opinion, the answer to the above question is yes and no – and both perspectives are useful. When we think of fieldwork and design procedures to convert reluctant sample members, we aim to change reluctant sample members' probabilities conditional on not having yet participated. We do this either by sending reminders, issuing another call, offering higher incentive, sending more experienced interviewer and so on – each of these with one aim: to increase the chance of response of those who have not responded yet. The idea that *conditional* response probabilities are constant and cannot be changed over the fieldwork period is not practical in such a situation as it would imply that whatever we do – we cannot help bringing more respondents through design. In this situation researchers are interested in response probability at a particular call – and it is useful to treat such conditional probabilities as prone to manipulation via survey design features.

Nonetheless I share the opinion of the author (Section 6) that the above perspective of changing probabilities over time is not useful in all contexts; in particular, weighting adjustment should estimate *final* probabilities, that is, total, cumulative probabilities over all stages of survey fieldwork. This is because at the end of the fieldwork period we aim to

extrapolate the information from final respondents to the whole sample (or population). It is therefore important to know the final response probability for each sample member. From this perspective final probabilities under the same protocol and in the same survey situation (population, topic of the survey, etc.) are constant, and do not vary over fieldwork time (unlike the conditional probabilities discussed above). I understand that in the discussion of Figure 1 in the article when describing RHGs Michael Brick talks about final probabilities to respond.

3. Weighting for a Two-Stage Design

One of the contributions of Michael Brick's article is the discussion of weighting for a two-stage design, where some respondents participate in a survey in the first stage, and at the second stage all or a subsample of nonrespondents is followed, some of whom also provide interviews. The author suggests two ways of developing weights in this situation, Method A and Method B. I believe that both methods are unbiased under specific assumptions. Method B is unbiased under MAR assumption that all respondents are different from nonrespondents only on variables in the nonresponse adjustment model. Method A is unbiased under MAR assumption that Stage 2 respondents are different from nonrespondents only on variables in the nonresponse adjustment model. I agree with the author that Method A corrects for nonresponse bias better than Method B when Stage 2 respondents are more similar to final nonrespondents in comparison to Stage 1 respondents.

I would like to suggest Method C for weighting correction in a two-stage design, which not only recognizes the two stages of design, but also recognizes that each respondent has a chance to respond at either (but not both) of the two stages. The discussion from the previous section becomes useful here: at both stages of the design respondents have probabilities to respond – the probability of responding in the second stage is conditional on not responding in the first stage; the total probability is the combination of these two probabilities. Thus, the total response probability can be expressed as

$$p_{\text{total}} = p_1 + (1 - p_1)*p_2$$

where p_1 is the probability to respond at the first stage and p_2 is the conditional probability to respond at the second stage. $(1 - p_1)$ expression reflects a chance of a sample member being issued into Stage 2, which is conditional on nonresponse in Stage 1.

In the design where second stage nonrespondents are subsampled, a probability of selection (p_{sel}) should be included in the expression:

$$p_{\text{total}} = p_1 + (1 - p_1)*p_{\text{sel}}*p_2$$

The important point here is that every selected sampling unit has a value for each probability. In other words, respondents, who are observed to have responded in Stage 1, had a chance to not respond in Stage 1. In this situation they would have a chance to be selected into Stage 2, and a conditional chance to respond in Stage 2.

While the formulae make sense theoretically, estimating these probabilities in practice is challenging given that we do not observe a Stage 2 response outcome for those not selected into Stage 2 (either because of subselection or because they have responded in

Stage 1 already). Such calculation is nevertheless possible and can follow an approach similar to the one in Kaminska and Lynn (2012). First, p_1 is estimated in the usual way using predictors available for respondents and nonrespondents. Selection probability p_{sel} is known by design. Next, p_2 is estimated only for those who were issued into Stage 2, drawing upon the same pool of auxiliary variables as in the above model. Given the model for p_2, we can now estimate p_2 for all respondents, including respondents from Stage 1. This is possible because the same auxiliary variables are available for all respondents. This way we estimate response probability in Stage 1, p_1, and conditional response probability in Stage 2, p_2, for each respondent, regardless of the stage at which they participated. This provides us with all the components required for the nonresponse correction.

One advantage of this approach over methods A and B, described by Michael Brick, is that it estimates response probabilities at each stage empirically and independently of each other, thus avoiding the unnecessary assumptions.

4. Conclusion

It has been an honour to be a discussant of such an interesting, comprehensive, current and innovative article. There are many more thoughts and ideas in the article worth discussion and further development. I feel we are at the turning point of understanding nonresponse and I look forward to future developments in this field.

5. References

Groves, R.M., Presser, S., and Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. Public Opinion Quarterly, 68, 2–31.

Kaminska, O. and Lynn, P. (2012). Combining Refreshment or Boost Samples with an Existing Panel Sample: Challenges and Solutions. Proceedings of the International Panel Survey Methods Workshop.

Singer, E., van Hoewyk, J., and Maher, M.P. (2000). Experiments with Incentives in Telephone Surveys. Public Opinion Quarterly, 64, 171–188.

Singer, E., van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999). The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys. Journal of Official Statistics, 15, 217–230.

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 359–362

Discussion

Phillip S. Kott[1]

Let me start by thanking Mike Brick for his informative guide through the literature, both old and new, on unit nonresponse in household surveys, its impact on the quality of survey estimates, and some of the methods developed to reduce the negative consequences of that impact. Although I have a few quibbles, to which I will get shortly, I nonetheless found the article a treasure trove of useful ideas and references. Moreover, I agree wholeheartedly with many of its conclusions.

Now to those quibbles. We read in Section 3 that "modeling either the response propensity or the outcome variable can be effective for response bias." Nevertheless, "in our experience most cross-sectional household surveys produce multiple characteristics and there are few auxiliary variables that are related to *any* of these outcomes [italics added]. In that situation, response propensity modeling may be the only remaining tool to reduce nonresponse bias."

Had "any" been replaced by "every" I would be inclined to agree. We only need one model to explain unit response, but we need a separate model for each outcome variable. The single response model either (nearly) holds or does not, while the outcome models can hold for some variables and fail miserably for others.

The author, however, claims that we often can model unit response but do not have the auxiliary variables to model any outcome variable. In my experience, the variables we use to model response can usually be employed to model survey outcome. When we form response homogeneity groups (RHGs) or poststrata, we very often are also creating outcome model groups, the units within each group having a common mean for many survey variables of interest. Ironically, that is why we use the design weights when computing the implicit probability of response within each RHG/poststrata: it reduces or removes the bias of the resulting estimates under the outcome model and thereby reduces the overall mean squared error (see Kott 2012).

If I may belabor the point a bit, consider a variation of the example in the beginning of Section 7. A simple random sample is divided into initial responders and nonresponders. The latter group is subsampled and extra effort is given to elicit responses from the subsample. Suppose, unlike in the text, the extra effort had been successful. If the responding units from two groups (initial responders and nonresponders) have distinct outcome means, then we need to have separate weights for each. Otherwise, we do not.

In practice, a survey usually has many variables of interest. It is possible that some have a common mean across the two groups while others have distinct means. We do not know.

With full response, we do not have to assume a model of outcome behavior; we can use probability-sampling principles to produce unbiased estimates in some sense by weighting

[1] RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. Email: pkott@rti.org

the groups separately. We often accept additional variance to protect ourselves from potential bias with survey samples because the former can be measured while measuring the latter is illusive. Moreover, sample sizes tend to be large, so bias often dominates variance within overall mean squared error.

Now, suppose there remains some unit nonresponse in the initial-nonresponder group even after expending the added effort. It seems reasonable to treat the units from that group as if they had a common mean for a survey variable, a mean that may be distinct from that of the initial responders. Again, we are trying to protect ourselves from potential bias. We are not, of course, fully protected because we have assumed an outcome model that may be flawed. In particular, we have assumed that the survey variables of respondents and nonrespondents among the initial nonresponders have a common mean. In the very special case of this example, the outcome model is equivalent to the response model, namely that each initial nonresponder is equally likely to respond after the extra effort. This equivalence is not usually the case. Still, I argue that if there are variables on which to build a response model, there are variables on which to build outcome models. Moreover, it is precisely when that is the case that there is nonresponse bias to reduce (see Little and Vartivarian 2005).

Little and Vartivarian also point out that an increase in the variability of the weights does not always lead to higher variances. For my taste, there is an over-emphasis in the text − and in general practice − on weight variability. At the very least, a prudent statistician should conduct sensitivity analyses to assess differences in the resulting estimates for key survey variables caused by using alternative nonresponse models, one simpler and one more complex. Moreover, under the assumption that an estimator using the more complex model is bias free, one can test whether the simpler model leads to a systematically biased estimate for a particular survey variable using a procedure proposed by Fuller (1984) for determining whether survey weights matter in a linear regression. An analogous procedure using replication was suggested by Korn and Graubard (1993).

In the Fuller procedure (the replication version is trivial), each observation is duplicated with one version assigned to Domain A and weighted one way while the other is assigned to Domain B and weighted the alternative way. Recognizing that both versions are from the same primary sampling unit, one can then use design-based software to measure whether the difference in the domain means for the variable are larger than we would expect due to random chance alone. Since we are as much concerned with Type 2 error (ignoring a real bias) as Type 1 (finding a false one), I would argue against accepting the null hypothesis (that the observed difference in the estimates derives from random noise) when the absolute t value of the difference is much greater than 1.

I wish the text had provided a deeper discussion on using calibration weighting to adjust for unit nonresponse. As it correctly points out, calibration can be viewed as a generalization of reweighting using RHGs (also called "weighting classes") or poststratification depending on whether one is calibrating to the original weighted sample (RHGs) or the population.

With linear calibration weighting, the adjusted weight for respondent i has the form:

$$d_i^* = d_i \left\{ 1 + \left[\mathbf{X} - \sum_{k \in s_r} d_k \mathbf{x}_k \right]^T \left(\sum_{k \in s_r} d_k \mathbf{x}_k \mathbf{x}_k \right)^{-1} \mathbf{x}_i \right\} = d_i (1 + \mathbf{g}^T \mathbf{x}_i) = d_i \nu_i \qquad (1)$$

(by the way, I would not call ν_i the "linear regression estimate" as is done in the text). There is no reason why some components of $\mathbf{x}_i$ cannot be related to the efforts employed in eliciting response as advocated by the author. Moreover, as Mike correctly points out, the totals in $\mathbf{X}$ can contain both population totals and weighted sums from the original sample.

Using linear calibration weights, the estimator in Equation (10) of the text $\left(\hat{y}_{ca} = \sum_{i \in s_r} d_i^* y_i\right)$ is nearly unbiased under the response model:

$$\Pr(R_i = 1|\mathbf{x}_i) = \frac{1}{\boldsymbol{\gamma}^T \mathbf{x}_i}, \tag{2}$$

where the unknown parameter vector $\boldsymbol{\gamma}$ is estimated by $\mathbf{g}$ in Equation (1). Equation (2) is not a very plausible response model except when the components of $\mathbf{x}_i$ are group-membership indicators, the special case of RHGs/poststrata.

To my way of thinking, a better justification for linear calibration weighting in general is that the survey variable behaves roughly like a random variable with mean $\mathbf{x}_i^T \boldsymbol{\beta}$. In household surveys, many survey variables are binary and cannot be modeled precisely as a linear function of the components of $\mathbf{x}_i$. That is why I added the modifier "roughly." In some surveys, there may be variables for which the linear outcome model ($\mathrm{E}(y_i) = \mathbf{x}_i^T \boldsymbol{\beta}$) does not come close to holding. I suspect that the linear calibration estimates for the totals of these variables will often not be close to unbiased either.

Even though linear calibration can sometimes be effective in reducing or removing nonresponse bias when the response model in Equation (1) is clearly wrong (see D'Arrigo and Skinner 2010), I prefer using a back-link function that implicitly assumes a more plausible response model. One such is $d_i^* = d_i\left[1 + \exp\left(\mathbf{g}^T \mathbf{x}_i\right)\right]$, which assumes $\Pr(R_i = 1|\mathbf{x}_i) = \left[1 + \exp\left(\boldsymbol{\gamma}^T \mathbf{x}_i\right)\right]^{-1}$, a logistic response model with $\mathbf{g}$ in the back-link function being a consistent estimator for $\boldsymbol{\gamma}$ when the assumed response model holds. Calibration weighting using this back-link function provides double protection against nonresponse bias since the estimator is nearly unbiased in some sense when either the response model or the linear outcome model holds (or when both hold; see Kott and Liao 2012).

Let me also point out that the vector, let us now call it $\mathbf{q}_i$, of variables in the response model need not coincide with the calibration variables in $\mathbf{x}_i$. As a result, although the calibration-variable totals ($\mathbf{X}$ in $\sum_{i \in s_r} d_i^* \mathbf{x}_i = \mathbf{X}$) must be known for either the original weighted sample or the population, some of the component of $\mathbf{q}_i$ need not be.

The response model is now $\Pr(R_i = 1|\mathbf{q}_i) = f(\mathbf{g}^T \mathbf{q}_i)$ for some back-link function $f(.)$. Since nothing prevents components of $\mathbf{q}_i$ from being survey variables, this version of calibration weighting can be used to treat nonresponse that is not missing at random. To my knowledge, Deville (2000) was the first to point this out. In Deville's formulation, the number of components of $\mathbf{q}_i$ and $\mathbf{x}_i$ must coincide. Chang and Kott (2008) extend calibration weighting for nonresponse adjustment to allow more calibration variables than response-model variables. Many of the calibration weighting ideas in that article have been incorporated into SUDAAN 11 (RTI International 2012).

I have concentrated here on relatively small points related to areas of my research in an article that has a much broader sweep. Mike Brick is to be congratulated for producing a fine contribution to the literature.

References

Chang, T. and Kott, P.S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika, 95, 557–571.

D'Arrigo, J. and Skinner, C. (2010). Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. Survey Methodology, 36, 181–192.

Deville, J.-C. (2000). Generalized Calibration and Application to Weighting for Non-Response. Compstat: Proceedings in Computational Statistics; 14th Symposium Held in Utrecht, The Netherlands, J.G. Bethlehem and P.G.M. Van Der Heijden (eds). Physica Verlag: Heidelberg, 65–76.

Fuller, W.A. (1984). Least Squares and Related Analyses for Complex Survey Designs. Survey Methodology, 10, 97–118.

Korn, E.L. and Graubard, B.I. (1993). Response to "Bias in Weighted Versus Unweighted Estimates.". American Journal of Public Health, 83, 1351.

Kott, P.S. (2012). Why One Should Incorporate the Design Weights When Adjusting for Nonresponse Using Response Homogeneity Groups. Survey Methodology, 95–99.

Kott, P.S. and Liao, D. (2012). Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-weighting Routine. Survey Research Methods, 6, 105–111.

RTI International (2012). SUDAAN Language Manual, Release 11.0. Research Triangle Park. NC: RTI International.

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 363–366

Discussion

Roderick J. Little[1]

I appreciate the opportunity to comment on Mike Brick's review of unit nonresponse adjustments for household surveys. The topic of unit nonresponse in design and analysis has received increased attention with the recent literature on "responsive design", attempting to improve the quality/cost tradeoffs and the deployment of alternative data collection modes to help address the escalating problems of contacting and interviewing respondents. Brick's review assembles a substantial body of research, and provides a useful summary of the "design-based" perspective on unit nonresponse adjustment.

The "critical" part of his "critical review" is mainly directed at our lack of understanding of the processes that lead to nonresponse, and the fact that cognitive research methods are focused on improving questionnaire design, rather than on more general aspects of survey design and analysis. Design approaches to limit missing data are important – I recently chaired a National Research Council panel (National Research Council 2010) where a major focus was design and conduct of clinical trials to reduce the level of missing data. However, is the increase in nonresponse in surveys that much of a mystery? It seems to me clear that people are harder to reach, busier, increasingly inundated with requests to fill out surveys, many from self-serving sources, and just want to be left alone. Characteristics of nonrespondents are important, but as a modeler (Little 2004, 2012), I think the field is too focused on reasons for nonresponse and not enough on modeling the relationship between nonresponse and survey outcomes.

Brick's review embraces the design-based perspective. My "critical review" of the literature would focus more on the limitations of that perspective, both for responsive design and for developing improved nonresponse adjustments. Thus, I liked Brick's quote of Ferber (1949) that "the problem of response bias must be considered with specific reference to a particular question or characteristic", but Brick's review does not really address this key aspect. Models of survey outcomes are largely absent, the emphasis being on modeling the response propensity and associated weighting adjustments. This lack of explicit modeling is characteristic of the design-based perspective – models are implicit and buried in the estimating equations – but attempting to address unit nonresponse without modeling the outcome is for me like trying to tie a shoelace with one hand behind one's back.

I now offer some more specific comments, driven by this overall perspective.

[1] University of Michigan, School of Public Health, Dept of Biostatistics, 1420, Washington Heights, Ann Arbor, MI 48109-2029, USA. Email: rlittle@umich.edu

1. Limits of Weighting

Brick's emphasis on weighting, in the title and the equations he presents, is a reflection of the current state of the field, where unit nonresponse is nearly exclusively handled by weighting adjustments. However, from the modeling perspective, the goal is not to weight, but to predict values of survey variables for nonrespondents, with estimates of uncertainty that reflect imputation error. This philosophy applies whether the nonresponse is at the unit or item level. Some prediction estimators can be expressed using respondent weights, but suitable nonresponse weights often do not have the interpretation of a sampling weight as a sampled case representing a certain number of individuals in the population.

Weighting is limited – it can handle unit nonresponse in a cross-sectional survey, or a monotone pattern such as occurs with attrition in a panel survey, where the variables $Y_1, \ldots, Y_p$ can be arranged so that Y_j is observed for all the cases where Y_{j+1} is observed for $j = 1, \ldots, p - 1$. It does not handle nonmonotone patterns of missing data well, which is one reason why it is not the approach of choice for item nonresponse. Unit nonresponse in its basic form has a monotone or close to monotone pattern, but nonmonotone patterns can be expected to be more prominent in future, with increased inclusion of information from administrative sources that have their own patterns of missing data. Prediction approaches such as multiple imputation can handle both unit and item nonresponse, and place the emphasis where I believe it belongs, on modeling the survey outcomes. Applying multiple imputation to unit nonresponse is counter to the current orthodoxy, but I note an increasing interest in creating multiple versions of synthetic data sets, where *all* the data, not just nonrespondent data, are imputed (Rubin 1993; Kinney et al. 2011).

2. Near-exclusive Focus on Bias Over Variance

Brick's review mentions precision in a few places, but the emphasis is on bias. He states on page 2 without supporting evidence that "bias is the dominant component of the nonresponse-related error in the estimates". I find this almost exclusive focus on bias odd, particularly since precision is the predominant concern in sample design. A more balanced approach would also consider efficiency, mean squared error, and good confidence interval coverage, but that requires modeling the survey outcomes. The emphasis on bias leads naturally to response propensity weighting and associated R indicators, but

Table 1. Effect of weighting adjustments on bias and variance of a mean, by strength of association of the adjustment cell variables with response and outcome

	Association with outcome	
Association with nonresponse	Low	High
Low	Cell 1 Bias: - - - - Var: - - - -	Cell 3 Bias: - - - - Var: ↓
High	Cell 2 Bias: - - - - Var: ↑	Cell 4 Bias: ↓ Var: ↓

weighting on a very good predictor of the response propensity that is not related to the outcome is making things worse, leading to inefficient estimates with highly variable weights – see the bottom left cell of Table 1, from Little and Vartivarian (2005) – and confidence intervals with below nominal coverage. On the other hand, other approaches focused on mean squared error, like model-based shrinkage of the weights (e.g., Elliott and Little 2000), or stratifying or weighting based on the predicted mean of an outcome, address both bias and variance (Little 1986; Little and Vartivarian 2005).

Under missing at random (MAR), the response propensity is potentially an important predictor in a model for predicting the outcomes, because misspecification of the relationship between the outcome and the propensity leads to bias – this motivates penalized spline of propensity prediction (An and Little 2004; Zhang and Little 2009, 2011), which models the relationship between the outcome and the propensity as a flexible penalized spline.

3. Missing Not at Random (MNAR) Models

Contrary to Brick's discussion in Section 5, I do not think that tinkering with the weights is a fruitful approach to modeling deviations from missing at random. Some estimates under MNAR models can be constructed in a weighted form (an early example is Little 1985), but I think that the best way to address the problem is to explicitly model the joint distribution of the nonresponse indicators and the survey outcomes, as in the proxy pattern-mixture analysis of Andridge and Little (2011). One promising area for improving nonresponse adjustments is the inclusion of proxy and survey process variables, which are often subject to measurement error.

West and Little (2012) address measurement error in auxiliary variables using a pattern-mixture model. Another simple pattern-mixture approach to modeling deviations from MAR is to apply multiple imputation with offsets to reflect differences in the predictive distribution of outcomes for nonrespondents and respondents. Giusti and Little (2011) describe this approach on a rotating panel survey, with missing income values and a nonmonotonic pattern.

Deviations from MAR will always remain a hard problem, and I agree with Brick that finding good predictors of the outcomes is key. Incidentally, Brick mentions that the approach of Schouten (2007) does not assume MAR, but at a key point in the argument regression coefficients estimated from the respondents are substituted for coefficients defined for the whole sample. This substitution is only justified under the MAR assumption.

Nonresponse adjustments, unit or item, require modeling assumptions. It is a problem of prediction, not weighting, in my opinion.

4. References

Elliott, M.R. and Little, R.J.A. (2000). Model-Based Alternatives to Trimming Survey Weights. Journal of Official Statistics, 16, 191–209.

Giusti, C. and Little, R.J. (2011). A Sensitivity Analysis of Nonignorable Nonresponse to Income in a Survey With a Rotating Panel Design. Journal of Official Statistics, 27, 211–229.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, 79, 363–384.

Little, R.J. (1985). Nonresponse Adjustments in Longitudinal Surveys: Models for Categorical Data. Bulletin of the International Statistical Institute, Proceedings of the 45th Session: Invited Papers, Section 15.1, 1-18

Little, R.J. (2004). To Model or not to Model? Competing Modes of Inference for Finite Population Sampling. Journal of the American Statistical Association, 99, 546–556.

Little, R.J. (2012). Calibrated Bayes: An Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder). Journal of Official Statistics, 28, 309–372.

Little, R.J. and An, H. (2004). Robust Likelihood-Based Analysis of Multivariate Data with Missing Values. Statistica Sinica, 14, 949–968.

Little, R.J. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? Survey Methodology, 31, 161–168.

National Research Council (2010). The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Washington, D.C. National Academy Press.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 462–468.

West, B. and Little, R.J. (2012). Nonresponse Adjustment Based on Auxiliary Variables Subject to Error. Applied Statistics, early view. DOI: 10.1111/j.1467–9876.2012.01058.x

Zhang, G. and Little, R.J. (2009). Extensions of the Penalized Spline of Propensity Prediction Method of Imputation. Biometrics, 65, 911–918, DOI: 10.1111/j.1541-0420.2008.01155.x.

Zhang, G. and Little, R.J. (2011). A Comparative Study of Doubly-Robust Estimators of the Mean with Missing Data. Journal of Statistical Computation and Simulation, 81, 12, 2039–122058, DOI: 10.1080/00949655.2010.516750.

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 367–370

Discussion

Geert Loosveldt[1]

The article by Michael Brick about unit nonresponse and weighting adjustments presents an excellent overview of the concepts, trends, and strategies in unit nonresponse research. This overview clearly demonstrates that the conceptual and analytical framework of nonresponse research is highly evolved and has been much improved. The author mention that we have a better understanding of the correlates of nonresponse and the methods to reduce nonresponse due to noncontact. However, as the title suggests, it is a critical review. Perhaps as a result, the general undertone is rather pessimistic. According to the author, response rates are falling in most countries and most procedures to reduce nonresponse are not effective. As a consequence, weighting adjustment procedures are important, but the author state that we do not have a sufficiently thorough understanding of the impact of these procedures on the reduction of nonresponse bias. In the discussion Brick concludes that "even after decades of research on nonresponse we remain woefully ignorant of the causes of nonresponse at a profound level" and "Perhaps the time is ripe for new approaches to the vexing and important questions of why people do and do not respond to surveys." As always, a discussion involving statements such as these is an invitation to formulate some related considerations, comments, and suggestions. The starting point is a few observations about the trend in nonresponse rates in the European Social Survey (ESS).

The ESS is a biennial, face-to-face survey organized in as many European countries as possible and concerns changes in attitudes across Europe (*http://www.europeansocialsurvey.org/*). The first round of the survey was organized in 2002. Figure 1a presents the response rates in the ESS Rounds 1–4, and Figure 1b illustrates the refusal rates (Matsuo et al. 2010; similar results concerning Round 1–3 are presented in Stoop et al. 2010).

The results in Figure 1a clearly illustrate that there are differences between countries in terms of response rates. In Poland and Portugal, for example, the response rate is always near the target of 70 percent, whereas in France and Switzerland the response rate in each Round (1–4) is below 50 percent. In addition, for the refusal rates (the largest category of nonresponse in most countries) we observe clear differences (e.g., low refusal rates in Greece and high refusal rates in France and Switzerland, Figure 1b). There are also differences within countries. In some countries there is a systematic increase or decrease in the refusal rate across the ESS rounds (e.g., an increase in the Netherlands and a decrease in Spain). In a few countries there is an increase in response rates: the Czech Republic, Spain, France, and Portugal.

The observed differences between countries and differences within countries put the overall trend of increasing refusal rates and decreasing response rates, and the related pessimistic opinion of the author about survey participation, into perspective.

[1] Professor Dr Geert Loosveldt, Department Sociologie, Katholieke Universiteit Leuven, Parkstraat 45, Bus 3601 B-03000 Leuven, BELGIUM. Email: Geert.Loosveldt@soc.kuleuven.ac.be

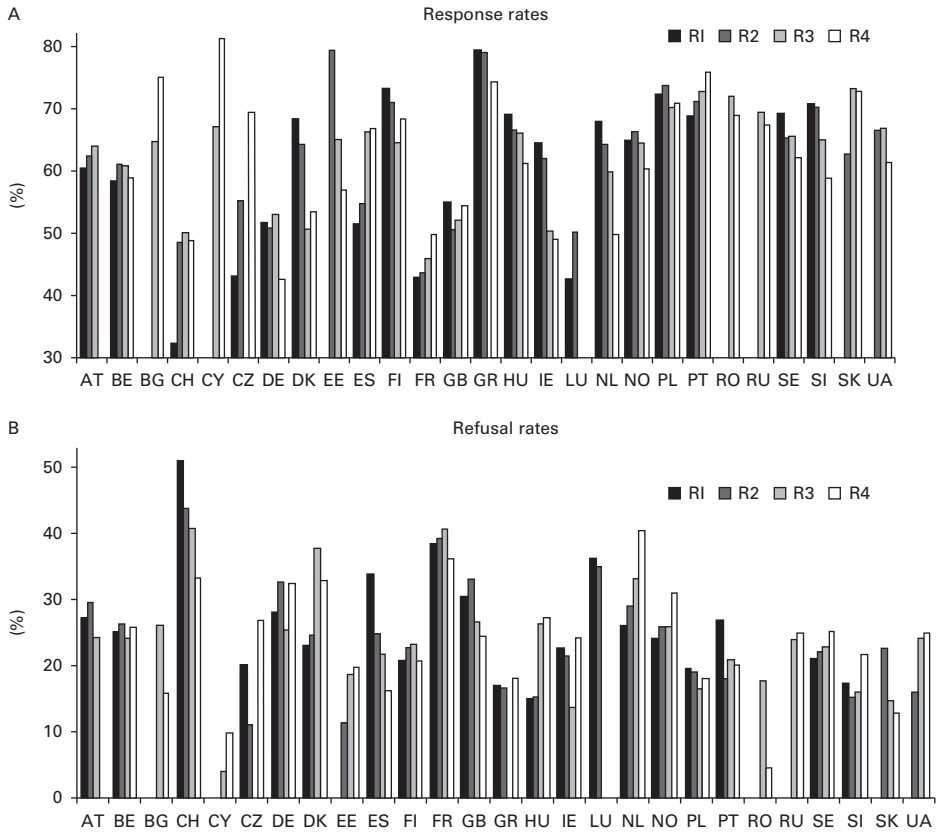


Fig. 1A. Response rates (%) in the ESS Rounds 1–4. (1B) Refusal rates (%) in the ESS Rounds 1–4

The differences also make it clear that a general theory about unit nonresponse must not only focus on the question of why people participate in a survey, but must also be able to identify the differences between and within countries. To understand the differences between countries and the deviations from the trend, characteristics at three different hierarchical levels seem to be relevant: the macro or country level, the intermediate or organizational level, and the micro or individual level. The authors' question of why people do not participate in surveys, and the discussed weighting adjustment procedures, are situated at the individual level. It is mainly a respondent-oriented approach that does not take into account the relevance of the other levels. I will argue that this restricted approach could be enriched by using information at the country level and organizational level that is relevant in explaining differences between and within countries. This additional information partially explains why people participate in surveys and is probably useful in optimizing weighting procedures.

The survey climate can be considered a relevant societal characteristic in explaining differences between countries. It relates to the public's willingness to cooperate and the extent to which people consider survey research, and thus survey interviews, to be useful and legitimate (Loosveldt and Storms 2008). The number of surveys in a society and the discussions in the media about the accuracy and utility of the results of various polls and

surveys all contribute to this climate. One can assume that a more positive survey climate stimulates individual participation. The individual subjective experience of the survey climate mediates the general survey climate (country level) and a respondent's decision to participate (individual level). This subjective experience manifests itself in an individual's opinions about different aspects of a survey (value, cost, enjoyment, reliability, and privacy). To answer the question of why people participate in surveys and to detect effective weighting variables, it is important to obtain information about the sample unit's opinion about surveys and the sample unit's characteristics that correlate with this opinion. In this context, the reasons for nonparticipation or refusals observed during the doorstep interaction with the sample unit can be informative. The doorstep interaction can also be used to ask respondents and nonrespondents a few basic questions about their opinions concerning surveys. This is a suggestion for comparative analysis as mentioned by the author in the discussion. It should be noted that fieldwork organizations such as National Statistical Institutes are partially responsible for the survey climate and can take initiatives to monitor and improve it (Lorenc et al. 2013). In this regard, unit nonresponse is not only the respondents' responsibility, as strongly suggested by the author's approach. To summarize the reflection on the survey climate, the survey climate can have an impact on the nonresponse rate and can be translated into characteristics at the individual level, which are correlated with substantial variables and discriminate between the group of respondents and that of nonrespondents.

The intermediate or organizational level refers to the capacity of the fieldwork organization and the way they organize and implement the survey. This level can be used to explain differences within countries with the same survey climate. The differences within countries illustrated in Figures 1a and 1b clearly demonstrate that the nonresponse profile within a country is not a fixed property. Characteristics of the survey design (e.g., use of incentives, selection and training of interviewers, quality and remuneration of interviewers) and paradata about the implementation of fieldwork procedures (e.g., efforts to contact respondents, refusal conversion procedure) is typical information at the organizational level that can be used to explain fluctuations in response rates within countries. As mentioned by Brick, paradata can be used to calculate response propensities based on the survey conditions. This refined definition of response propensity stresses the idea that it is not a fixed property of respondents. Paradata is available for all sample units and sometimes is the only information available with which to calculate response propensities. This is probably the reason why paradata is becoming popular. This kind of data meets the need of researchers to have information about both respondents and nonrespondents. However, available data is not always relevant data, and at the organizational level it is necessary to assess the relevance and meaning of data with regard to the respondent's decision to participate or not. Here also, it is necessary to translate the information at organizational level into relevant sample unit characteristics (e.g., number of contacts and ability to contact them) in order to answer the question of why respondents participate in or refuse an interview. These sample unit characteristics stem from the way in which the fieldwork is implemented and these kinds of characteristics are useful to calculate response propensities with as much exploratory power as possible. Similar comments can be formulated concerning register or sampling frame data and observational data. The latter is data about the sample unit's (respondents and nonrespondents) type

of dwelling and neighborhood characteristics such as litter and graffiti. This type of individual-level data is mostly collected by interviewers and can be used as proxy for socioeconomic status to calculate response propensities.

The current use of paradata, register or sampling frame data, and observable data to calculate response propensities illustrates the core problem of unit nonresponse analysis and weighting adjustment procedures: the need for sufficient and relevant information about nonrespondents. All the types of secondary data can only partially answer the question of why people refuse to participate in a survey. However, it is clear that this information deficit cannot be resolved by means of survey research. Therefore, it seems better not only to focus on the particular respondent participation question, but also to concentrate on what kind of information at each level can be used to decrease the nonresponse rate and to understand the differences between the group of respondents and of nonrespondents. The ultimate objective is to reduce bias and to improve survey estimates.

References

Loosveldt, G. and Storms, V. (2008). Measuring Public Opinions About Surveys. International Journal of Public Opinion Research, 20, 74–89.

Lorenc, B., Loosveldt, G., Mulry, M., and Wrighte, D. (2013). Understanding and Improving the External Survey Environment of Official Statistics. Survey Methods: Insights from the Field. Available at: http://surveyinsights.org/?p=161 (accessed June 21, 2013).

Matsuo, H., Billiet, J., Loosveldt, G., and Molnar, B. (2010). Response-Based Quality Assessment of ESS Round 4: Results for 30 Countries Based on Contact Files. Leuven: University of Leuven, Centre for Sociological Research, 77.

Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving Survey Response: Lessons Learned from the European Social Survey. Chichester, UK: Wiley.

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 371–374

Rejoinder

J. Michael Brick

I want to thank the editors of the *Journal of Official Statistics* for inviting me to prepare this article and for obtaining such a distinguished set of discussants for it. As I prepared the article, I quickly realized that reviewing the massive literature on nonresponse and nonresponse bias is a daunting exercise. It has given me even greater appreciation for those who have done such excellent research in this area.

I would also like to thank all the discussants. Their comments give valuable insights into nonresponse bias and I found their remarks very stimulating. I would also like to thank the discussants for pointing out issues in my initial draft; their suggestions helped me improve the quality of the article.

The diversity of the discussants' comments and concerns highlight some of the challenges we face dealing with nonresponse in surveys. Below, I briefly address some key similarities and differences I have with comments provided by each of the discussants.

Professor Loosveldt perceives my review as being pessimistic, and I understand this reaction. My review tried to paint the challenges of nonresponse as starkly as possible. However, I share his optimism about making progress, but only if we face the complex issues associated with nonresponse. Our field has many skilled and innovative methodologists and, if they work on nonresponse diligently, then I believe we will see significant improvements in our understanding and methodology.

Loosveldt notes that nonresponse has more levels and complications than are discussed in my review. I fully agree and would like to add to those he mentions factors such as the mode of data collection and 'house' or organizational effects. The effects of these factors can be substantial. He also mentions that the survey climate interacts with response propensities. I again agree, but note that thus far we have not done well in specifying exactly how the interaction works (Brick and Williams 2013). As he describes, the entire system, including the data collection process and other cultural factors, could be critical to nonresponse bias and we need to better understand these effects and how national statistical organizations can influence them.

Professor Loosveldt's extension to include cross-country comparisons provides a nice perspective on the nonresponse problem. His idea of capturing data in a doorstep interview seems to be an option worth pursuing. This idea appears to be related to those of Kulka et al. (1982) and Lynn (2003).

A final point of clarification is that I did not intend to suggest nonresponse was the respondent's responsibility. Rather, I was trying to urge those of us who mount surveys to take a more respondent-friendly orientation. In the early days, respondents may have been

more willing to do any survey, but those days (if they ever existed!) are gone. It is the survey's responsibility to be more respectful in asking respondents to spend their time on and give attention to a task they did not initiate.

Dr. Kaminska explains several of the complex issues related to nonresponse bias in ways that I found informative. I especially enjoyed her discussion of overall and conditional response propensities. She clarifies when and why each type of propensity is important to researchers. I agree with her that at the data collection stage the conditional propensities are essential, while at the weighting stage only the overall response propensities are important.

She also offers fresh ideas for weighting in a two-phase design. Her Method C seems very reasonable, and she clearly points out difficulties that might arise associated with computing the response propensities required for her method. I encourage her to pursue the research necessary to evaluate the statistical properties of her proposed method because I believe it has promise.

In describing her proposed method, she states that we should include the probability of selection for the second phase in the weights. I agree with her, but suspect others might not. As a design-based survey statistician, I would accept the inclusion of these selection probabilities in the weights as a default proposition and require strong evidence of their ineffectiveness before dropping them. This is related to a comment by Dr. Kott on accepting some increase in variance to reduce the potential for bias.

Dr. Kott notes that sometimes the same variables that affect response are related to the outcome variables and could be modeled for this purpose. I agree, and the paper by Micklewright et al. (2012) is an excellent example of this. His quibble about my use of the word 'any' is an important one. In the example I was trying to point out that if we had this information for any important outcome variables we should include it in the estimator regardless of whether it is related to response. Dr. Kott correctly points out that we would need this information for 'every' important outcome to support the robustness goal when modeling outcomes and using a design-based adjustment framework.

As mentioned above, I also agree with Dr. Kott on the importance of bias in the large sample sizes that are common in national statistical office surveys – and share the position that we should take actions to reduce the potential for bias even if it incurs some additional variance. Although Dr. Kott felt the text overemphasized variability of the weights, I did not intend that. I agree that there may be too much emphasis on this point in general, since with reasonable precautions adjustments for nonresponse rarely substantially increase the variance of the estimates.

Some of the differences Dr. Kott mentions may be a manifestation related to what Särndal (2007) referred to as a difference between calibration and GREG "thinking." Dr. Kott prefers a model justification for the linear calibration estimator, while I think in terms of restoring balance in the calibration variables. He prefers a logistic response model (Kott and Liao 2012), while I often choose raking. In practice, the differences are often not very substantial.

I am skeptical of the use of survey variables in calibration advocated in Chang and Kott (2008) and Kott and Chang (2010). Specifically, I worry that the procedure might induce substantial bias if the statistician makes a poor choice of survey variables for calibration.

I have not investigated this myself but look forward to more research to clarify the robustness of the procedure.

Professor Little states nonresponse requires modeling assumptions and I fully agree. The models I discussed differ from those he advocates – those I describe primarily model the response mechanism while he prefers modeling the relationship between response and the outcome – but modeling is essential. Furthermore, I agree that modeling the relationship between response and outcomes can be useful, especially if the result can be implemented in such a way to produce a general purpose nonresponse adjustment. As I noted in the article, I prefer that "powerful auxiliaries for key outcomes should be included in the estimator when they are available, irrespective of their relationship to response." The rationale is that such modeling reduces variance. If this procedure is followed, then residual nonresponse adjustment must be primarily based on a response propensity model. If the modeling of the outcomes is not done in advance, then modeling outcomes and response propensities is valuable.

Micklewright et al. (2012) is an example of where modeling the outcome led to adjustment related to response propensities. The adjustment they applied reduced the variance of the specific outcomes modeled as well as reducing nonresponse bias in the outcome and other statistics. If the auxiliary data they used had not been related to the specific outcome but was still related to response propensities, it is the type of general purpose nonresponse adjustment I would propose even though it would not reduce the bias for the specific outcome. Unfortunately, there may be no auxiliary data available that are strongly correlated with response propensities, and in this case response propensity adjustments are ineffective.

Professor Little restates his opinion that design-based inference is flawed and needs to be replaced by model-based approaches (Little 2012). I, on the other hand, find the design-based approach and nonresponse adjusted weights to be a valuable tool. Lohr (2007) gives some properties of weights that are desirable. Of those she describes, the properties of robustness, internal consistency of the estimates, and objectivity are critical in my assessment. Model-based estimates, as currently proposed, do not fully satisfy all of these properties. Related issues were raised by Brion, Smith, and Beaumont in their discussion of Little (2012).

The design-based procedures I described are general purpose, simple to use, and accessible for a wide variety of users. This means users can access the data set and produce an estimate without modeling a specific estimate. They can obtain the same estimate as the data set producer. The estimated totals they produce for subsets (e.g., males and females) equal the total for all persons when summed. These properties may sound trivial, but they are important to users. Over the years, national statistical offices have translated these user requirements into quality measures (e.g., Statistics Canada 2009). The quality measures include timeliness, accessibility, interpretability, and coherence; these measures are not statistical in the sense of producing minimal mean square error estimates.

I agree with Professor Little that, with sufficient effort, a model-based estimate may give a more efficient statistical solution for a particular estimate than a general purpose, design-based weighting procedure. If an important decision depended on one or a small set of estimates from a survey, it might be prudent to examine alternatives to the general

purpose approach to improve the accuracy of the estimates. My inclination would be to seek more efficient design-based alternatives for the specific estimates, but model-based alternatives are another reasonable approach. However, in my opinion the existing model-based methods do not sufficiently address user-oriented quality measures such as timeliness, accessibility, interpretability, and coherence for the vast majority of applications. I doubt model-based methods will be adopted in practice unless they do.

I also have a different perspective on whether "the field is too focused on reasons for nonresponse." Knowing the reasons for nonresponse is essential to design efforts to reduce nonresponse. Similarly, modeling nonresponse appropriately requires understanding the reasons for nonresponse. For example, Lin and Schaeffer (1995) provide compelling evidence that outcomes can be very dependent on the reason for nonresponse.

On the preferences for weighting and imputation, I consider both as methods of implementing an estimation scheme. In some cases, the two are equivalent; for example, hot-deck imputation can be rewritten in terms of item-specific weights for a particular estimate. The choice of whether to use weights or impute is based largely on usability considerations. Imputation is preferable when sufficient data, such as responses to other items by the same respondent, are available. Weighting is preferable when characteristics at the sampled unit level are limited. However, both weighting and imputation are just different tools for accomplishing the same goal.

Finally, I agree with Professor Little that multiple imputation can be valuable, even though it is not the best solution to all nonresponse adjustment problems. Multiple imputation is a form of replication and I, like many design-based statisticians, am fond of replication.

References

Kulka, R., McNeill, J., and Bonito, A. (1982). On the Manifest Designation of Key Items: a Cost Effective Procedure for Improving the Collection and Processing of Survey Data. Paper Presented at the 37th Annual Conference of the American Association for Public Opinion Research. Maryland, USA: Hunt Valley.

Lohr, S. (2007). Comment: Struggles with Survey Weighting and Regression Modeling. Statistical Science, 22, 175–178.

Lynn, P. (2003). PEDAKSI: Methodology for Collecting Data About Survey Non-Respondents. Quality & Quantity, 37, 239–261.

Särndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33, 99–119.

Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf (Accessed March 2013).

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 375–396

Incorporating User Input Into Optimal Constraining Procedures for Survey Estimates

Matthew Williams[1] *and Emily Berg*[2]

We examine the incorporation of analyst input into the constrained estimation process. In the calibration literature, there are numerous examples of estimators with "optimal" properties. We show that many of these can be derived from first principles. Furthermore, we provide mechanisms for injecting user input to create user-constrained optimal estimates. We include derivations for common deviance measures with linear and nonlinear constraints and we demonstrate these methods on a contingency table and a simulated survey data set. R code and examples are available at https://github.com/mwilli/Constrained-estimation.git.

Key words: Calibration; general deviance measures; nonlinear constraints; raking; user feedback.

1. Introduction

Constrained estimation has diverse applications in survey estimation. In the presence of auxiliary information, calibration of survey weights can improve the efficiency of a design consistent estimator. Deville and Särndal (1992) define calibrated weights as the weights that minimize a deviance function subject to the restriction that the weighted sum of a vector of auxiliary variables is equal to a known population total. They suggest a family of deviance functions and demonstrate that the resulting calibration estimators are asymptotically equivalent to a generalized regression estimator, a particular type of calibration estimator that arises from a quadratic deviance function. Chen and Sitter (1999) formulate the calibration problem using an empirical likelihood. Calibration can also be used to reduce a bias due to undercoverage of the sampling frame or nonresponse (for example, see Kott 2006, Chang and Kott 2008, and D'Arrigo and Skinner 2010). In a seminal paper, Deming and Stephan (1940) use iterative proportional fitting to enforce a restriction that the estimated marginal totals of a two-way table agree with census margins.

Whether the purpose of the calibration is to improve the efficiency of a design-unbiased estimator or reduce a bias due to nonsampling errors, care is often needed to avoid negative or extreme weights. Deville and Särndal (1992), Chen et al. (2002), and Singh and Mohl (1996) discuss methods for imposing range restrictions on calibrated weights.

[1] Research and Development Division, National Agricultural Statistics Service, U. S. Department of Agriculture, Fairfax, VA 22030, U.S.A. Email: matt.williams@nass.usda.gov
[2] Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. Email: emilyb@iastate.edu
Acknowledgments: Thanks to colleagues at NASS for their support. Thanks also to Malay Ghosh for feedback and encouragement.

These ensure that each sampled unit represents a reasonable positive number of units in the population.

Another application of constrained estimation is benchmarking of small area estimates to ensure that aggregated model-based estimates agree with a direct estimator or a previously published statistic for a larger region. Wang et al. (2008) review benchmarking methods in the context of a linear mixed model. They define a class of benchmarked estimators by minimizing a quadratic form subject to the benchmarking restriction. Nandram and Sayit (2011) incorporate linear constraints for small area probabilities using hierarchical Bayes and the standard beta-binomial model. In related work with shrinkage estimators, Ghosh (1992) imposes constraints on the mean and variance of Bayes estimates for a quadratic loss function. While variance constraints are quadratic (nonlinear), the use of a quadratic loss function leads to a closed form solution.

Many of the applications of constrained estimation discussed above apply linear constraints (see Särndal 2007; Estevao et al. 1995, who mention ratios of totals) to a set of initial estimates or initial weights by solving a constrained optimization problem. While the methods serve different purposes and have distinct interpretations, the functional forms are similar and derivations can be based on fundamental mathematical principles (such as the method of Lagrange multipliers). Because of the similarities between methods, constrained estimation in the survey world can seem like a tangle of overlapping terms and concepts. One of the objectives of this article is to clarify some of these associated concepts.

What is missing in the literature is a framework to create an interface between a user and the automated constraining procedure. Such a framework is essential for a statistical agency which is tasked with establishing estimates that are timely and accurate with the expectation of being compatible with subject or commodity knowledge and administrative data with partial coverage. Incorporating constraints into such a process must go beyond default settings and a choice of deviance measures. In addition to clarifying concepts, the purpose of this work is to establish such a framework.

1.1. Motivating Example

For statistical agencies, data often occur in triplets of *numerator* (**n**), *denominator* (**d**), and the *ratio* (**r**) of the two. Suppose we have a set of such triplets which must agree in aggregation with known targets (Table 1). Most methods in the literature would use linear constraints on the totals for **n** and **d**. But if **r** represents an agricultural rate of yield, which is production (**n**) per harvested area (**d**), then biological and industry knowledge would suggest adjusting the ratio directly (using nonlinear constraints) rather than the total production. The choice of which two of the three items in each triplet to adjust will often give distinct solutions. Figure 1 compares the relative adjustments made to each initial estimate when applying equivalent methods for constrained estimation to **n** and **d** versus **d** and **r**. The linear approach applies a constant proportional adjustment (decreasing for **n** and increasing for **d**). The nonlinear approach decreases **r** and increases **d**, but not at the same rates across all rows.

Constrained estimation provides a way for an analyst to incorporate external knowledge of the process that generated the basic estimators (either the direct survey estimators or

Table 1. Simulated Survey Data (rounded). Targets increase (light) and decrease (dark)

	Num (**n**)	Den (**d**)	Ratio (**r**)
1	2,586.20	56.55	45.73
2	30,491.31	913.17	33.39
3	4,141.68	78.83	52.54
4	1,975.41	68.59	28.80
5	18,827.87	362.00	52.01
6	6,280.19	137.20	45.77
7	8,597.05	182.03	47.23
8	4,995.37	242.78	20.58
9	7,402.01	216.61	34.17
10	1,168.46	52.52	22.25
11	5,455.36	243.30	22.42
12	1,778.24	60.79	29.25
13	3,208.09	195.24	16.43
14	2,249.00	56.44	39.85
15	2,215.65	72.80	30.44
16	14,297.99	454.96	31.43
17	3,948.72	190.49	20.73
18	1,653.01	77.39	21.36
19	2,545.01	86.12	29.55
20	2,749.02	72.91	37.70
Total	126,565.63	3,820.71	33.13
Target	120,237.35	3,935.33	30.55

estimators based on a subsequent model). For instance, contributions of large operators in establishment surveys, sizes of nonresponse and bias adjustments, administrative records, historical data, and qualitative information about the data-generating mechanism can be difficult to integrate into the basic estimation procedure, but might factor into an analyst's decision to set some values and reweight the adjustments on others. The analyst would then need a procedure to enforce these additional "user" constraints. For example, we can use analyst knowledge to fix entire rows in Table 1 *and* fix individual ratios r_i, *and*

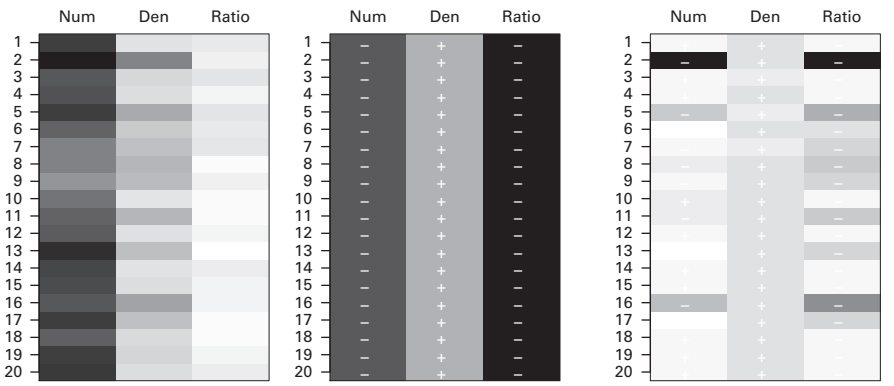


Fig. 1. Heat Map for Default Constraint of Triplets: (left to right) Log(Data), % Change (Linear), % Change (Nonlinear). White to black increases counts or size of change. Signs (−/+) show direction of change

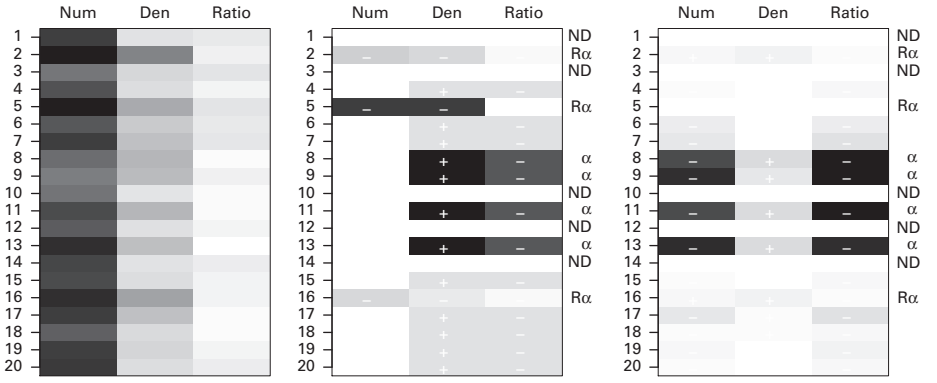


Fig. 2. Heat Map for User Constraint of Triplets: (left to right) Log(Data), % Change (Linear), % Change (Nonlinear). White to black increases counts or size of change. Signs (−/+) show direction of change. Num and Den fixed (ND), Ratio set (R), rows reweighted (α)

reweight to redistribute the amount of change absorbed by some rows. In heat map representation (Figure 2), these adjustments take the form of white cells (no change) and increases in intensity (darker up-weighted cells).

In the next section, we review the relationship between constraints and deviance measures. We introduce the concept of user interaction with an optimal procedure and explore several examples that might occur. Section 3 contains the details for a Newton-type method to generate solutions. In Section 4, we revisit the data set from Deming and Stephan (1940), applying our framework to incorporate user interaction. In Section 5 we elaborate on the example of linear and nonlinear constraints for triplets described in Subsection 1.1. Finally, we conclude in Section 6 with a summary and implications for further research. R code for methods and examples is available at https://github.com/m-willi/Constrained-estimation.

2. Constrained Estimation

We consider the vector of observations (or unrestricted estimates) $\mathbf{y}$ of length n. We may wish to impose $k < n$ linear constraints $\mathbf{Ax} = \mathbf{q}$, where $\mathbf{x}$ is a constrained version of $\mathbf{y}$. Linear constraints take the form of weighted sums $\mathbf{a}_i\mathbf{x} = \sum_j a_{ij}x_j = q_i$ for $i \in 1, \ldots, k$ where $\mathbf{a}_i$ is the *ith* row of the $k \times n$ coefficient matrix $\mathbf{A}$. We restrict $\mathbf{A}$ to have full row rank k. Otherwise at least one $\mathbf{a}_i$ leads to a redundant constraint or creates a conflicting constraint. Consider an example in which constraints are imposed on all marginal totals of a two-way table with R rows and C columns. Because both row and column margins sum to the total for the table, a coefficient matrix $\mathbf{A}$ containing $R + C$ rows, one for each column and row sum, will contain one redundant row. This creates a deficient row rank for $\mathbf{A}$ of $R + C - 1$. A row associated with one of the row or column sums can be removed to produce a coefficient matrix with $R + C - 1$ rows and thus full row rank. (See Section 4 for further discussion of restrictions on the marginal totals of a two-way table).

We also consider $k < n$ *nonlinear* constraints $g(\mathbf{x}) = \mathbf{q}$. While the general class of nonlinear functions (all functions which are *not* necessarily linear) is extremely broad, we limit consideration to those that are well defined and have $n \times k$ continuous derivatives

$\mathbf{D}_g(\mathbf{x})$. In practice we consider polynomial, rational, and transcendental functions and compositions of them. These are generally wellknown and wellbehaved nonlinear functions. For example, the ratio of numerator and denominator from our motivating example (Subsection 1.1) is simple and wellbehaved when the denominator is nonzero. The variance constraint imposed by Ghosh (1992) for shrinkage estimates is a basic quadratic function used to counteract the overshrinking which occurs commonly in applications such as small area estimation.

We can no longer appeal to matrix rank to ensure that we do not have any conflicting constraints. However, it is clear that equations such as $x_1 + x_2 = q_1$, $x_1 - x_2 = q_2$, and $x_1/x_2 = q_3$ produce a conflict. Many methods for solving nonlinear systems of equations use linearization techniques involving derivatives (see Section 3). For these methods to find solutions, further restrictions may be placed on $\mathbf{D}_g(\mathbf{x})$. For our purposes we will assume $\mathbf{D}_g(\mathbf{x})$ has full column rank k for each value of $\mathbf{x}$.

2.1. Deviance Measures

Since $n > k$, the constraints by themselves do not imply a unique solution $\mathbf{x}$, but instead a family of solutions. A reasonable criteria to select a member of this family is to choose the $\mathbf{x}$ "closest" to $\mathbf{y}$. This concept of closeness implies minimizing a scalar deviance between $\mathbf{x}$ and $\mathbf{y}$. We will generally restrict these deviances to be rather simple and interpretable. From the calibration literature (for example, see Deville and Särndal 1992), there are several deviance functions used. We highlight the three most popular (for example, see D'Arrigo and Skinner 2010): the quadratic deviance $\chi^2(\mathbf{x}|\mathbf{y})$, the Poisson deviance $I(\mathbf{x}|\mathbf{y})$, and the discrimination information $D(\mathbf{x}|\mathbf{y})$. Each of these measures falls within the framework developed. Practitioners may use their current preferred deviance measure and still take advantage of the results and ideas presented here. Alternatively, one can change the deviance measure while still maintaining the other structures described below, such as the weighting matrix and the form of the constraints.

We express these deviances in matrix formulation and provide a weighting structure ($\mathbf{W}$) which allows for user input from an analyst or another model (see Subsection 2.3). The matrix formulations for some of the deviance measures may seem unnecessary, but the key insights come from the matrix formulation of the *constraints*. Expressing both in terms of matrix operations makes them more directly compatible (Subsection 2.2). We assume the base $\mathbf{W}$ is symmetric and invertible (although often the case, $\mathbf{W}$ need not be positive definite). We define $\langle \mathbf{v} \rangle$ as a square diagonal matrix with vector v on the diagonal and 0s elsewhere. We use the notation $[\cdot]$ to denote elementwise operations in two ways: First we use $[\mathbf{ab} + \mathbf{c}]$ for vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ of the same dimension to produce a vector with the *ith* element equal to $a_i b_i + c_i$. Second we denote $f[\mathbf{v}]$ as yielding a vector with *ith* element equal to $f(v_i)$. In other words, $f[\mathbf{v}]$ applies the scalar function $f(\cdot)$ elementwise to each v_i.

The Quadratic Deviance

$$\chi^2(\mathbf{x}|\mathbf{y}) = (\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$$

Examples include the Pearson chi-squared distance ($\mathbf{W} = \langle \mathbf{y} \rangle^{-1}$) and the Least Squares distance ($\mathbf{W} = Var(\mathbf{y})^{-1}$). Use is often motivated by a regression-based approach (Fuller 2002).

The Poisson Deviance

$$l(\mathbf{x}|\mathbf{y}) = 1'\mathbf{W}_d\left[\mathbf{y}\log\left[\frac{\mathbf{y}}{\mathbf{x}}\right] - \mathbf{y} + \mathbf{x}\right]$$

where $\mathbf{W}_d$ is a diagonal matrix of full rank. Motivation comes from the deviance measure of a log-linear model comparing a restricted model of means to the saturated model (Agresti 2002). In this case, $\mathbf{y}$ is the data (or saturated model) and $\mathbf{x}$ is the restricted model for $\mathbf{y}$. We will show in Subsection 2.2 that the Poisson deviance leads to the *pseudo-empirical maximum likelihood estimator* of Chen and Sitter (1999).

The Discrimination Information

$$D(\mathbf{x}|\mathbf{y}) = l(\mathbf{y}|\mathbf{x}) = 1'\mathbf{W}_d\left[\mathbf{x}\log\left[\frac{\mathbf{x}}{\mathbf{y}}\right] - \mathbf{x} + \mathbf{y}\right]$$

The name seems to come from the application of the principal of minimum discriminability to cell probabilities (see Ireland and Kullback 1968, who attribute this to Good and Kullback). So-called *raking* methods such as iterative proportional fitting (IPF; Deming and Stephan 1940) are readily available to minimize this deviance for specific settings.

Of the three, $\chi^2(\mathbf{x}|\mathbf{y})$ is the simplest to implement and will often lead to closed-form solutions (Subsection 3.1). However, when $\mathbf{y}$ are positive survey weights, some of the resulting $\mathbf{x}$ may be negative. $D(\mathbf{x}|\mathbf{y})$ and $l(\mathbf{x}|\mathbf{y})$ are often preferred in this context, because $\mathbf{x}$ will remain positive for both methods. We can see that $D(\mathbf{x}|\mathbf{y})$ and $l(\mathbf{x}|\mathbf{y})$ are closely related and easy to confuse. However, the estimating equations for each are clearly different (see Table 2), so the emphasis is often placed here rather than on the original measures. To further add to the confusion, when $\mathbf{y}$ is already close to satisfying the constraints ($\mathbf{A}\mathbf{y} \approx \mathbf{q}$ or $g(\mathbf{y}) \approx \mathbf{q}$), the three deviance criteria give very similar results, thus explaining the error in Deming and Stephan (1940) (see Section 4).

2.2. *Solving for Linear and Nonlinear Constraints*

Suppose we are given a vector $\mathbf{y}$ and wish to find the $\mathbf{x}$ satisfying a possibly nonlinear constraint $g(\mathbf{x}) = \mathbf{q}$ for some vector-valued function $g(\mathbf{x})$ with derivative matrix $\mathbf{D}_g(\mathbf{x})$ (the linear form $\mathbf{A}\mathbf{x} = \mathbf{q}$ is a special case with $\mathbf{D}_g(\mathbf{x}) = \mathbf{A}'$). Since such an $\mathbf{x}$ will generally not be unique, let $\mathbf{x}$ minimize the deviance $d(\mathbf{x}|\mathbf{y})$. Assume $d(\mathbf{x}|\mathbf{x}) = 0$ for all appropriate $\mathbf{x}$. However $d(\mathbf{x}|\mathbf{y})$ need not be symmetric $d(\mathbf{x}|\mathbf{y}) \neq d(\mathbf{y}|\mathbf{x})$. We assume the derivative

Table 2. Five common deviance measures (Deville and Särndal 1992) and the corresponding functions needed for estimation

Name	Deviance	$d^{(1)}$	$h[\mathbf{u}]$	$h^{(1)}[\mathbf{u}]$
Quadratic	$(\mathbf{x}-\mathbf{y})'\mathbf{W}(\mathbf{x}-\mathbf{y})$	$(\mathbf{x}-\mathbf{y})$	$\mathbf{y}+\mathbf{u}$	1
Discrimination	$1'\mathbf{W}_d\left[\mathbf{x}\log\left[\frac{\mathbf{x}}{\mathbf{y}}\right] - \mathbf{x} + \mathbf{y}\right]$	$\log\left[\frac{\mathbf{x}}{\mathbf{y}}\right]$	$[\mathbf{y}\exp[\mathbf{u}]]$	$[\mathbf{y}\exp[\mathbf{u}]]$
Hellinger	$(\sqrt{\mathbf{x}}-\sqrt{\mathbf{y}})'\mathbf{W}_d(\sqrt{\mathbf{x}}-\sqrt{\mathbf{y}})$	$1-\left[\frac{\mathbf{y}}{\mathbf{x}}\right]^{\frac{1}{2}}$	$[\mathbf{y}[1-\mathbf{u}]^{-2}]$	$2[\mathbf{y}[1-\mathbf{u}]^{-3}]$
Poisson	$1'\mathbf{W}_d\left[\mathbf{y}\log\left[\frac{\mathbf{y}}{\mathbf{x}}\right] - \mathbf{y} + \mathbf{x}\right]$	$1-\left[\frac{\mathbf{y}}{\mathbf{x}}\right]$	$[\mathbf{y}[1-\mathbf{u}]^{-1}]$	$[\mathbf{y}[1-\mathbf{u}]^{-2}]$
Alternative Quadratic	$(\mathbf{x}-\mathbf{y})'\mathbf{W}_d\langle\mathbf{x}\rangle^{-1}(\mathbf{x}-\mathbf{y})$	$1-\left[\frac{\mathbf{y}}{\mathbf{x}}\right]^{2}$	$[\mathbf{y}[1-\mathbf{u}]^{-\frac{1}{2}}]$	$\frac{1}{2}[\mathbf{y}[1-\mathbf{u}]^{-\frac{3}{2}}]$

$\partial d(\mathbf{x}|\mathbf{y})/\partial \mathbf{x} = \mathbf{W}d^{(1)}(\mathbf{x}|\mathbf{y})$ is composed of well-defined elementwise invertible functions on the $\mathbf{x}$ vector. In other words, the *ith* element $d^{(1)}(\mathbf{x}|\mathbf{y})_i$ only contains information from $\mathbf{x}_i$ and $\mathbf{y}_i$ not $\mathbf{x}_j$ or $\mathbf{y}_j$. We also assume $d^{(1)}(\mathbf{x}, \mathbf{x}) = \mathbf{0}$. It's clear that $\chi^2(\mathbf{x}|\mathbf{y})$, $l(\mathbf{x}|\mathbf{y})$, and $D(\mathbf{x}|\mathbf{y})$ are each examples of $d(\mathbf{x}|\mathbf{y})$ (see Table 2 for these and two more from Deville and Särndal 1992). We take $\mathbf{W}$ to be a symmetric and invertible weight matrix.

In order to minimize $d(\mathbf{x}|\mathbf{y})$, subject to constraints $g(\mathbf{x}) = \mathbf{q}$, we use the method of Lagrange multipliers (see, for example Stewart 2011). When such a solution $\mathbf{x}$ exists, the derivatives $\mathbf{W}d^{(1)}(\mathbf{x}|\mathbf{y})$ are parallel to the columns in $\mathbf{D}_g(\mathbf{x})$, the derivatives of each of the k constraints. The $k \times 1$ vector $\boldsymbol{\lambda}$ scales for the differences in magnitude of these parallel vectors. Symbolically,

$$\mathbf{W}d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}, \tag{1}$$

or equivalently,

$$d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}.$$

Since $d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{u}$ is an elementwise invertible operation on $\mathbf{x}$ producing the vector $\mathbf{u}$, the inverse function $h(\mathbf{u}) = \mathbf{x}$ exists and is also elementwise. Together with $g(\mathbf{x}) = \mathbf{q}$, we obtain the following estimating equations

$$\begin{aligned} \mathbf{x} &= h[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}] \\ \mathbf{q} &= g(h[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}]). \end{aligned} \tag{2}$$

Two properties become apparent from (1) and (2):

Lemma 1. *The solution* $\mathbf{x}$ *to (2) is invariant to the choice of the scalar* $\alpha \neq 0$ *in* $\mathbf{W}_{new} = \alpha\mathbf{W}_{old}$.

Example: If we are using $(\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$ or $100(\mathbf{x} - \mathbf{y})'\mathbf{W}(\mathbf{x} - \mathbf{y})$ as the deviance $d(\mathbf{x}|\mathbf{y})$, we will get the same solution $\mathbf{x}$.

Lemma 2. *The solution* $\mathbf{x}$ *to (2) is invariant to the rotation of constraints* $\mathbf{L}g(\mathbf{x}) = \mathbf{L}\mathbf{q}$ *for full rank square rotation matrix* L.

Example: In a two-way table, if we constrain all row and column totals, we have one redundant constraint. Ignoring any one row or column total gives the same solution $\mathbf{x}$.

Proof. Both properties come from $\boldsymbol{\lambda}$ being a dummy variable, an intermediate value used to solve for $\mathbf{x}$. This property implies an invariance to one-to-one transformations. In Equation (1), using $\alpha\mathbf{W}$ is equivalent to using $\boldsymbol{\eta} = \alpha^{-1}\boldsymbol{\lambda}$ as the multiplier. Likewise, rotating $\mathbf{L}g(\mathbf{x})$ will lead to the derivative $\mathbf{D}_g(\mathbf{x})\mathbf{L}'$, which is equivalent to using the rotated multiplier $\boldsymbol{\eta} = \mathbf{L}'\boldsymbol{\lambda}$. □

Now consider partitioning $\mathbf{y}' = [\mathbf{y}'_{-s}, \mathbf{y}'_s]$ and $\mathbf{x}' = [\mathbf{x}'_{-s}, \mathbf{x}'_s]$ indexed by the set s of size n_s and its complementary set $-s$ of size n_{-s}. Define the selection operator $\boldsymbol{\delta}(s) = [\mathbf{0}_{n_s \times n_{-s}}, \boldsymbol{I}_{n_s}]'$ such that $\mathbf{y}_s = \boldsymbol{\delta}'\mathbf{y}$. The weight matrix $\mathbf{W}$ is also partitioned corresponding to s and $-s$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_a & \mathbf{W}_b \\ \mathbf{W}'_b & \mathbf{W}_c \end{bmatrix}$$

Set values in $\mathbf{W}$ corresponding to the set s equal to 0:

$$\mathbf{W}_0 = \begin{bmatrix} \mathbf{W}_a & \mathbf{0}_{n_{-s}\times n_s} \\ \mathbf{0}_{n_s\times n_{-s}} & \mathbf{0}_{n_s\times n_s} \end{bmatrix}.$$

Then the Moore-Penrose generalized inverse of $\mathbf{W}_0$ is

$$\mathbf{W}_\mathbf{0}^- = \begin{bmatrix} \mathbf{W}_a^{-1} & \mathbf{0}_{n_{-s}\times n_s} \\ \mathbf{0}_{n_s\times n_{-s}} & \mathbf{0}_{n_s\times n_s} \end{bmatrix}.$$

Two more properties of the estimating equations (1) and (2) are now available:

Lemma 3. *For the estimating equations in (2) we add additional equality constraints of the form* $\mathbf{x}_s = \mathbf{y}_s$. *The following implementations give equivalent solutions for* $\mathbf{x}_{-s}$:

- *Augment the constraint targets* $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{y}_s']$ *and the corresponding equations* $g^{*\prime} = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$.
- *Keep the original* $\mathbf{q}$ *and* $g(\mathbf{x})$ *and substitute* $\mathbf{W} = \mathbf{W}_0$ *and* $\mathbf{W}^{-1} = \mathbf{W}_0^-$.

Lemma 4. *For the estimating equations in (2) we add additional equality constraints of the form* $\mathbf{x}_s = \mathbf{z}_s$ *for arbitrary values* $\mathbf{z}_s \neq \mathbf{y}_s$. *If* $\mathbf{W}$ *is diagonal* ($\mathbf{W}_d$) *or block-diagonal* ($\mathbf{W}_b = \mathbf{0}_{n_{-s}\times n_s}$), *the following implementations give equivalent solutions for* $\mathbf{x}_{-s}$:

- *Augment the constraint targets* $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{z}_s']$ *and the corresponding equations* $g^{*\prime}(\mathbf{x}) = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$.
- *Keep the original* $\mathbf{q}$ *and* $g(\mathbf{x})$ *and substitute* $\mathbf{y}_s = \mathbf{z}_s$, $\mathbf{W} = \mathbf{W}_0$, *and* $\mathbf{W}^{-1} = \mathbf{W}_0^-$.

Proof. See Appendix A for details. □

2.3. User Interaction

The goal of this proposed framework is to provide an interface between an informed analyst (or metamodel) and an automated "optimal" procedure which minimizes a deviance measure as described above. The choice of deviance measure will likely be made based on the application area and current conventions (i.e., the discrimination information for *raking* problems). A default weight matrix $\mathbf{W}$ may be a function of estimated variances based on a sample design or a specified model. As we have mentioned in our example in Subsection 1.1, knowledge of the process may be more difficult to fully and directly incorporate into the initial estimation procedures, thus motivating the need for an analyst to make adjustments.

We may expect the user to have limited control over the original $\mathbf{y}$ and the necessary constraints $\mathbf{q}$, leaving only the $\mathbf{W}$ to be adjusted. However, the user is free to provide additional constraints by augmenting $\mathbf{A}$ and $g(\mathbf{x})$. From Lemmas 3 and 4, several of these augmentations can be implemented by changing $\mathbf{y}$ and $\mathbf{W}$, thus preventing an increase in the dimension of the estimating equations (2).

- The user wishes to protect some $\mathbf{y}_s$ from changing. Some values may be the result of previously published data and are therefore ineligible for adjustment.

- The user sets $\mathbf{x}_s = \mathbf{z}_s \neq \mathbf{y}_s$. The user may wish to replace some dubious values or force changes in a direction opposite of the default procedure.
- The user reduces the changes to $\mathbf{y}_s$ without fixing the $\mathbf{x}_s$ values. From Lemma 1, we know that choice of scalar $\alpha \neq 0$ in $\alpha\mathbf{W}$ has no impact on $\mathbf{x}$. However, multiplying subsets of $\mathbf{W}$ by α *will* affect $\mathbf{x}$:

$$\mathbf{W}_\alpha = \begin{bmatrix} \frac{1}{\alpha}\mathbf{W}_a & \frac{1}{\sqrt{\alpha}}\mathbf{W}_b \\ \frac{1}{\sqrt{\alpha}}\mathbf{W}_b' & \mathbf{W}_c \end{bmatrix}.$$

  For $\alpha > 1$, the values of $\mathbf{x}_{-s}$ stray further from $\mathbf{y}_{-s}$, thus absorbing more change.

The sets $\{s\}$ and $\{-s\}$ must be chosen carefully when using $\mathbf{W}_0$ to avoid singularities. Since they are equivalent to adding more constraints to $\mathbf{q}$, we may inadvertently create a constraint on $\mathbf{x}$ which conflicts with $g(\mathbf{x}) = \mathbf{q}$. A finite choice of α, which provides weaker protection, can be used without this problem. Furthermore, $\mathbf{W}_0$ and α may be used together by establishing more than one partitioning set $\{s\}$.

This system provides a good compromise between an automated approach which ignores important expert knowledge for a specific subset $\mathbf{x}_s$, and a completely manual process which may use ad hoc methods to fill in the complementary $\mathbf{x}_{-s}$ values where knowledge is limited.

3. Implementation with Newton's Method

Given $\mathbf{x}$, we can use Newton's method to iteratively solve for the $\boldsymbol{\lambda}$ satisfying the second line of (2). We then update $\mathbf{x}$ and iterate the process until convergence. Denote $h^{(1)}(\mathbf{u}) = \partial h(\mathbf{u})/\partial \mathbf{u}$ as the *matrix* of derivatives. Since $h(\mathbf{u})$ is an elementwise function on $\mathbf{u}$, $h^{(1)}(\mathbf{u}) = \langle h^{(1)}[\mathbf{u}] \rangle$ where $h^{(1)}[\mathbf{u}]$ is a *vector* of elementwise derivatives of $\partial h(u_i)/\partial u_i$. Applying one chain rule for nested function gives:

$$\partial g(h[\mathbf{u}])/\partial \mathbf{u} = \langle h^{(1)}[\mathbf{u}] \rangle \mathbf{D}_g(h[\mathbf{u}]).$$

Then applying another chain rule for a change of variables:

$$\partial g(h[\mathbf{B}\boldsymbol{\lambda}])/\partial \boldsymbol{\lambda} = \mathbf{B}' \langle h^{(1)}[\mathbf{B}\boldsymbol{\lambda}] \rangle \mathbf{D}_g(h[\mathbf{B}\boldsymbol{\lambda}]),$$

where $\mathbf{B}$ is an arbitrary matrix. Let $\mathbf{B} = \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})$. For a given $\mathbf{x}^i$, Newton's method becomes:

$$\begin{aligned} \boldsymbol{\lambda}_i^{j+1} = \boldsymbol{\lambda}_i^j + &\left[\mathbf{D}_g'(\mathbf{x}^i)\mathbf{W}^{-1}\left\langle h^{(1)}\left[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right]\right\rangle \mathbf{D}_g\left(h\left[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right]\right)\right]^{-1} \\ &\times \left(\mathbf{q} - g\left(h\left[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right]\right)\right). \end{aligned} \tag{3}$$

We update $\mathbf{x}_i$ in an outer loop to satisfy the first line of (2):

$$\mathbf{x}^{i+1} = h\left[\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right]. \tag{4}$$

After convergence, the estimate $\mathbf{x}$ will be the same regardless of rotation (Lemma 2). However, rotations of the constraints may lead to different intermediate values for

(3) and (4). By updating $\boldsymbol{\eta} = \mathbf{L}'\boldsymbol{\lambda}$ using $\mathbf{L}g(\cdot)$, $\mathbf{Lq}$, and $\mathbf{D}_g(\cdot)\mathbf{L}'$, we will get the same final solution, but different $\mathbf{x}^i$ and $\boldsymbol{\lambda}_i^j$ before convergence.

3.1. Linear Constraints

For linear constraints, $h(\mathbf{u})$ will eliminate the need to iterate (4) for every $d(\mathbf{x}|\mathbf{y})$. For the quadratic deviance $\chi^2(\mathbf{x}|\mathbf{y})$, the inner loop (3) is one step, thus leading to the closed form solution

$$\mathbf{x} = \mathbf{y} + \mathbf{W}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}')^{-1}(\mathbf{q} - \mathbf{A}\mathbf{y}). \tag{5}$$

This result is common in the econometrics and engineering literature (Green 2000; Pizzinga 2010), where the $\mathbf{y}$ are least squares estimates of regression coefficients and $\mathbf{W}$ is their covariance matrix. The linear case with diagonal $\mathbf{W}_d$ for the quadratic and discrimination information deviances is available in the survey literature discussed above.

3.2. Alternatives to Newton's Method

We have presented a Newton method above to provide a general approach that can utilize different deviance measures and can accommodate both linear and nonlinear constraints. Many alternatives to Newton methods exist for specific optimization problems. In the survey literature, alternatives for solving the estimating equations (2) tend to be specific to *one* deviance measure and *linear* constraints. For example, iterative proportional fitting (IPF) is a popular way to impose linear restrictions on the cells of a multiway table using the discrimination information deviance. Software for IPF include the R function "loglin" and the SAS subroutine "ipf". The function "apop_rake" in the Apophenia library (http://apophenia.info/) implements IPF in a low-level programming language.

We emphasize Newton's method for two main reasons. Firstly, Newton's method is applicable to more general classes of constraints, deviance functions, and data structures. It is not limited to linear constraints on multiway tables, but can apply simple (i.e., continuously differentiable) nonlinear constraints to any data set that can be represented as an array. Secondly, the procedures of Subsection 2.3 for incorporating user input via the modification of $\mathbf{W}$ and $g(\mathbf{x})$ are readily implemented with Newton's method. Choosing between methods may depend on software availability, the experience of the user, and the size and nature of the data set. However, the properties of the solutions (see the Lemmas above) and the use of a user framework come from the estimating equations (2) and therefore hold regardless of the manner in which a solution was obtained (IPF, Newton's method, stochastic search, etc.).

Within our proposed Newton method there are alternatives to using $h(\mathbf{u})$. Let $h_{\mathbf{x}}(\mathbf{u})$ be a function of $\mathbf{u}$ given $\mathbf{x}$ satisfying $h_{\mathbf{x}}(d^{(1)}(\mathbf{x})) = \mathbf{x}$. The inverse need not be true: $d^{(1)}(h_{\mathbf{x}}(\mathbf{u})) \neq \mathrm{u}$. Obviously $h(\mathbf{u})$ is a special case of $h_{\mathbf{x}}(\mathbf{u})$. For the Poisson deviance, we choose $h_{\mathbf{x}}(\mathbf{u}) = [\mathbf{x}\mathbf{u} + \mathbf{y}]$ with $h_{\mathbf{x}}^{(1)}[\mathbf{u}] = \mathbf{x}$, which is *not* a function of $\mathbf{u}$. For this choice of $h_{\mathbf{x}}(\mathbf{u})$, for linear constraints, we need an outer loop (4), but not an inner (3) loop for $\boldsymbol{\lambda}$. See Appendix B for details. Since both the $h(\mathbf{u})$ and the $h_{\mathbf{x}}(\mathbf{u})$ approaches lead to the same solution at convergence, preference between the two methods may lie in interpretability of

the intermediate steps. For example, we prefer $h_{\mathbf{x}}(\mathbf{u})$ for the Poisson deviance, because the steps are the same as for the quadratic deviance, but with $\langle \mathbf{x} \rangle \mathbf{W}_d^{-1}$ replacing the $\mathbf{W}^{-1}$. Thus we can minimize the Poisson deviance by iteratively using the methods for the quadratic deviance. Using $h(\mathbf{u})$ for the Poisson deviance recreates the *pseudo-empirical maximum likelihood estimator* (Chen and Sitter 1999) since the solution $\mathbf{x}$ is invariant to choosing $\boldsymbol{\eta} = -\boldsymbol{\lambda}$. Chen et al. (2002) give an alternative iterative Newton method for linear constraints using this estimator.

4. Deming and Stephan (1940) Revisited

We revisit a classic example of *raking* by using our generalized techniques on the data set from Deming and Stephan (1940). The observed data (Table 3) are cell counts N_{ij} in a two-way table with margins $N_{i.}$ and $N_{.j}$ for rows $i \in 1, \ldots, 6$ and columns $j \in 1, \ldots, 4$. The constrained margins $M_{i.}$ and $M_{.j}$ are the targets q. The grand totals $N_{..} = M_{..}$ by coincidence and need not be true in general. The objective is to find cell counts M_{ij} that are closest to N_{ij} in terms of deviance, while satisfying the marginal constraints.

Although Deming and Stephan assert that their method of iterative proportional fitting (IPF) minimizes the least squares deviance $\chi^2(\mathbf{x}|\mathbf{y})$, IPF actually minimizes the discrimination deviance $D(\mathbf{x}|\mathbf{y})$ (Deville and Särndal 1992). We can obtain the estimates that minimize $\chi^2(\mathbf{x}|\mathbf{y})$ in one step (5) and use iteration (3) to obtain estimates minimizing $l(\mathbf{x}|\mathbf{y})$ and $D(\mathbf{x}|\mathbf{y})$. As it turns out, the estimates are quite close across the three deviance measures.

Using the discrimination deviance, we wish to compare the original results to those from two hypothetical user actions (Figure 3). For the default choice, it seems that the row margins are dominant (rows are all + or all −) and that most change occurs in the first column (darkest).

- The user specifies two cells ($M_{3,1} = 1{,}516$ and $M_{5,4} = 160$) and prevents these from changing. These are changes in the opposite direction from the default. Therefore the rest of the values in those rows and columns must take on more change (darker) and may switch direction (− to + or + to −).
- The user down-weights columns 3 and 4 by a factor of $\alpha = 5$. This allows the values in these columns to absorb more change and thus provides a weak protection for

Table 3. Data from Deming and Stephan (1940). Margin targets increase (light) and decrease (dark)

$i \backslash j$	1	2	3	4	$N_{i.}$	$M_{i.}$
1	3,623	781	557	313	5,274	5,252
2	1,570	395	251	155	2,371	2,395
3	1,553	419	264	116	2,352	2,432
4	10,538	2,455	1,706	1,160	15,859	15,766
5	1,681	353	171	154	2,359	2,330
6	3,882	857	544	339	5,622	5,662
$N_{.j}$	22,847	5,260	3,493	2,237	33,837	
$M_{.j}$	22,877	5,285	3,462	2,213		33,837

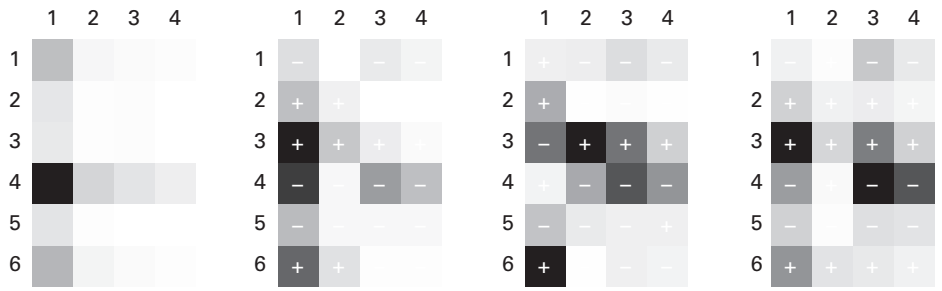


Fig. 3. Heat Maps using the Discrimination Measure: (left to right) Original Data, Default Changes, User Set Changes ($M_{3,1}$ and $M_{5,4}$), and Reweighted Changes ($M_{.3}$ and $M_{.4}$). White to black increases counts or size of change. Signs (−/+) show direction of change

columns 1 and 2. We notice that the general pattern of changes is similar to the default, but columns 3 and 4 are darker and columns 1 and 2 lighter compared to the original solution. This confirms that we have indeed shifted more change onto the last two columns.

4.1. Implementation

First we stack $\mathbf{y}$ by rows:

$$\mathbf{y} = [N_{i=1}, N_{i=2}, N_{i=3}, N_{i=4}, N_{i=5}, N_{i=6}]'.$$

Then we formulate $\mathbf{q}$, remembering to remove one redundant constraint $M_{.1}$ (Lemma 2 assures us that any choice of row or column margin will do):

$$\mathbf{q} = [M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{5.}, M_{6.}, M_{.2}, M_{.3}, M_{.4}]'.$$

Next we construct $\mathbf{A}$, which is simply a table of 1s and 0s. (Table 4 shows the transpose $\mathbf{A}'$). For example, the 2*nd* column of $\mathbf{A}$, (row of $\mathbf{A}'$) corresponds to $x_2 = M_{1,2}$. This cell is involved in the first constraint $M_{1.} = \sum_j M_{1,j}$ and the seventh constraint $M_{.2} = \sum_i M_{i,2}$, so the corresponding values in $\mathbf{A}$ have 1s. The rest of the entries for x_2 are 0s. We use $\mathbf{W} = \langle \mathbf{y} \rangle^{-1}$ for $\chi^2(\mathbf{x}|\mathbf{y})$ and $\mathbf{W}_d = \mathbf{I}_{24}$, the identity matrix, for $l(\mathbf{x}|\mathbf{y})$ and $D(\mathbf{x}|\mathbf{y})$. For fixing values, we construct $\mathbf{W}_0^-$ and $\mathbf{W}_{d0}^-$ with zeros for setting $M_{3,1} = 1{,}516$ and $M_{5,4} = 160$. For down-weighting values, we pre- and postmultiply $\mathbf{W}$ by a diagonal matrix with 1 for columns 1 and 2 and $\sqrt{1/5}$ for columns 3 and 4.

5. Simulated Survey Example

We now provide more detail for the motivating example in Subsection 1.1, in which data occur in triplets of *numerator* (**n**), *denominator* (**d**), and the *ratio* (**r**) of the two. At each level of aggregation (individual, regional, national), we only need to focus on two of the three. It is often the case that we have already set (and published) triplets at a higher level of aggregation (national totals) and now wish to set triplets at lower levels constrained to be consistent when aggregated. For example, we would need the totals for **n** and **d** to sum to the higher level totals. In the context of the methods discussed above, we can do this in at least two ways:

Table 4. Value of $\mathbf{A}'$ for Deming and Stephan (1940) data

#	$x_{\#}$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$j=2$	$j=3$	$j=4$
1	$M_{(1,1)}$	1	0	0	0	0	0	0	0	0
2	$M_{(1,2)}$	1	0	0	0	0	0	1	0	0
3	$M_{(1,3)}$	1	0	0	0	0	0	0	1	0
4	$M_{(1,4)}$	1	0	0	0	0	0	0	0	1
5	$M_{(2,1)}$	0	1	0	0	0	0	0	0	0
6	$M_{(2,2)}$	0	1	0	0	0	0	1	0	0
7	$M_{(2,3)}$	0	1	0	0	0	0	0	1	0
8	$M_{(2,4)}$	0	1	0	0	0	0	0	0	1
9	$M_{(3,1)}$	0	0	1	0	0	0	0	0	0
10	$M_{(3,2)}$	0	0	1	0	0	0	1	0	0
11	$M_{(3,3)}$	0	0	1	0	0	0	0	1	0
12	$M_{(3,4)}$	0	0	1	0	0	0	0	0	1
13	$M_{(4,1)}$	0	0	0	1	0	0	0	0	0
14	$M_{(4,2)}$	0	0	0	1	0	0	1	0	0
15	$M_{(4,3)}$	0	0	0	1	0	0	0	1	0
16	$M_{(4,4)}$	0	0	0	1	0	0	0	0	1
17	$M_{(5,1)}$	0	0	0	0	1	0	0	0	0
18	$M_{(5,2)}$	0	0	0	0	1	0	1	0	0
19	$M_{(5,3)}$	0	0	0	0	1	0	0	1	0
20	$M_{(5,4)}$	0	0	0	0	1	0	0	0	1
21	$M_{(6,1)}$	0	0	0	0	0	1	0	0	0
22	$M_{(6,2)}$	0	0	0	0	0	1	1	0	0
23	$M_{(6,3)}$	0	0	0	0	0	1	0	1	0
24	$M_{(6,4)}$	0	0	0	0	0	1	0	0	1

- We can focus on adjusting **n** and **d** leading to linear constraints

$$\mathbf{q} = \left[\sum_i n_i, \sum_i d_i\right]' = \mathbf{A}\mathbf{x}.$$

- We can focus on adjusting **d** and **r** leading to nonlinear constraints

$$\mathbf{q} = \left[\sum_i r_i d_i, \sum_i d_i\right]' = g(\mathbf{x}).$$

We can motivate the first method based on simplicity. However, the second method appeals to us if there is more intuition for **r** than **n**. It may also be the case that **r** is more independent of **d** than **n** is. For example, agricultural agencies publish total production (**n**), harvested area (**d**), and yield per area (**r**) for major crops. Focusing on production **n** may be overemphasizing constraints on area **d**. In addition, there is much scientific and commodity knowledge about the values for yield **r**.

We consider a simulated set of triplets for $i = 1, \ldots, 20$ artificial regions which grow soybeans. Based on published values for the U.S. (www.nass.usda.gov), we choose a symmetric distribution of soybean yields ranging between 15–55 bu/acre and a skewed

distribution for harvested area over a 20-fold range with units in either the 100s (county) or 1,000s (state) of acres.

- Simulate 20 values for $d_i \sim \left[Unif\left(\frac{1}{1{,}000}, \frac{1}{50}\right)\right]^{-1}$.
- Simulate 20 values for $r_i \sim N(\mu = 35,\ \sigma = 10)$ independently of d_i.
- Calculate $n_i = d_i \times r_i$ for each value.

The resulting data were shown in Table 1. The target values for constraints were arbitrarily chosen such that the target total for **n** and **d** were 95% and 103% respectively of the observed totals.

5.1. A Hypothetical User Experience

A hypothetical user wants to constrain the triplets data from Table 1 with the option of imposing adjustments based on experience and judgment. The user has experience with *raking*, so decides to use the discrimination deviance. The nonlinear formulation is new to the user, so both it and the linear approach are run in parallel to compare the results. The actions of the user are summarized as a flowchart in Figure 4.

The user begins with the default solutions from our motivating example above (Figure 1), but then realizes that regions 1, 3, 10, 12, and 14 are only sampled annually and

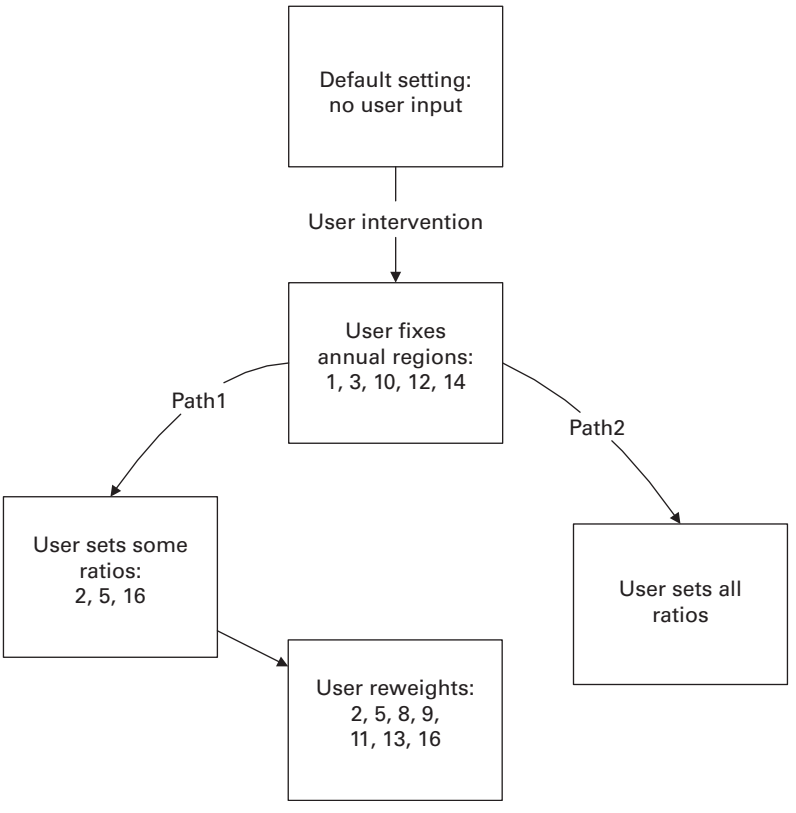


Fig. 4. Process flow of user decisions and estimates for the triplets data set

have not been sampled in the current survey. Instead, the most recent valid values have been passed forward. These have already been published and are therefore not eligible to be changed. To protect these values, the user adds 0s into the corresponding entries of the weight matrix **W**. The procedure is run again and new values are produced.

Now the user looks at the yield ratios **r** more carefully and compares them to the survey estimates. Historically, the survey gives high quality estimates for this ratio. If possible, the user would like to keep these ratios fixed. The user decides to take two different paths and explore their impact (Figures 5 and 6):

- Path 1: The user sets ratios (rounded to the integer) for regions 2, 5, and 16. Several regions change more than the user can comfortably justify. Regions 2, 5, 8, 9, 11, 13, and 16 are reweighted ($\alpha = 5$) to absorb more change from the other regions. The heat maps confirm that adjustments are now more concentrated (darker) in these regions.
- Path 2: The user fixes all ratios and is surprised to see that the linear and nonlinear approaches give identical results. By setting all **r**, both **r** and **n** are eliminated from adjustment, producing two linear constraints on **d**. Regions with higher yield **r** have harvested area **d** decreased, whereas those with lower yield have harvested area

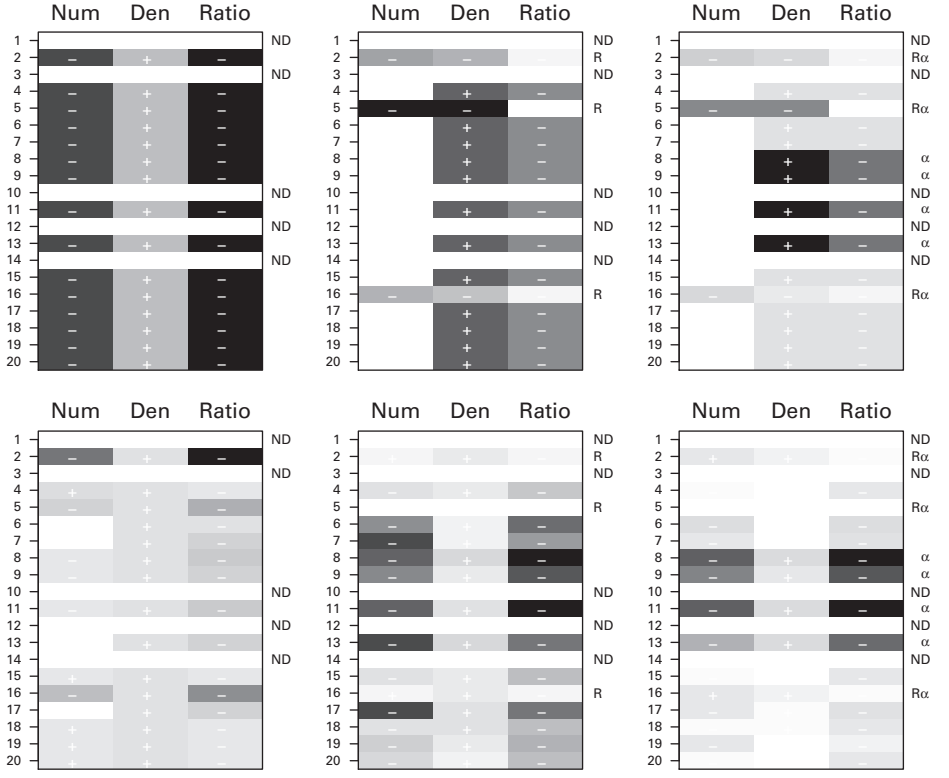


Fig. 5. Heat Maps for Path 1 (using discrimination deviance) for linear (top row) and nonlinear (bottom row) approaches. User successively adds constraints (left to right): Fixing annual regions (ND), setting yield ratios (R), reweighting to redistribute (α)

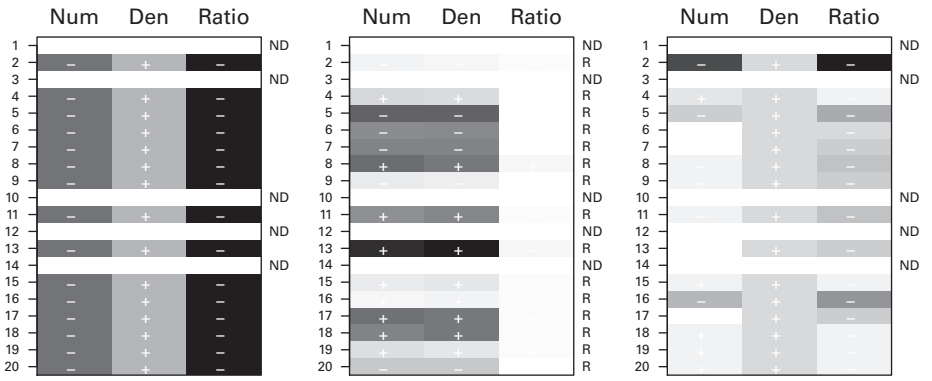


Fig. 6. Heat Maps for Path 2 (using discrimination deviance). Linear (left) and nonlinear (right) approaches converge when the user sets all yield ratios (center). Num and Den fixed (ND), Ratio set (R)

increased. Thus the overall production **n** has been decreased, but the overall harvested area **d** has been increased to simultaneously meet both aggregate targets.

5.2. Implementation Details

To implement the process above, we define $\mathbf{y}' = [\mathbf{n}'_{\mathbf{y}}, \mathbf{d}'_{\mathbf{y}}, \mathbf{r}'_{\mathbf{y}}]$ as the stacked set of unconstrained values. We seek the corresponding stacked constrained values $\mathbf{x}' = [\mathbf{n}'_{\mathbf{x}}, \mathbf{d}'_{\mathbf{x}}, \mathbf{r}'_{\mathbf{x}}]$ whose aggregate values satisfy the target $\mathbf{q}' = [120{,}237.35,\ 3{,}935.33,\ 30.55]$. For the linear approach, we use the numerator and denominator directly: $\mathbf{y}'_l = [\mathbf{n}'_{\mathbf{y}}, \mathbf{d}'_{\mathbf{y}}]$, $\mathbf{x}'_l = [\mathbf{n}'_{\mathbf{x}}, \mathbf{d}'_{\mathbf{x}}]$, and $\mathbf{q}'_l = [120{,}237.35,\ 3{,}935.33]$. Depending on the user's choices, the **A** matrix varies (Table 5). For the default settings, **A** is simply two rows of indicators, with 1 where an element of $\mathbf{x}_l$ is present in the sums $\left\{\sum n_i, \sum d_i\right\}$ and 0 otherwise. For Path 1, the user sets some r_i. Then for the total $\sum n_i$, the term $r_i d_i$ replaces some n_i. Thus the corresponding entries in **A** are 0 for n_i and r_i for d_i. Path 2 has a similar **A** matrix except with more r_i present. No adjustment to **A** is needed for jointly fixing the pairs $\{n_i, d_i\}$ for $i = 1, 3, 10, 12, 14$. For these cases, we construct $\mathbf{W}_0^-$ and $\mathbf{W}_{d0}^-$ with corresponding zeros. We suggest using $\mathbf{W} = \langle \mathbf{y} \rangle^{-1}$ for $\chi^2(\mathbf{x}|\mathbf{y})$ and $\mathbf{W}_d = \mathbf{I}_{40}$ for $l(\mathbf{x}|\mathbf{y})$ and $D(\mathbf{x}|\mathbf{y})$.

For the nonlinear approach, we directly adjust denominator and ratio: $\mathbf{y}'_{nl} = [\mathbf{d}'_{\mathbf{y}}, \mathbf{r}'_{\mathbf{y}}]$, $\mathbf{x}'_{nl} = [\mathbf{d}'_{\mathbf{x}}, \mathbf{r}'_{\mathbf{x}}]$, but still use the total production and harvested area as the targets $\mathbf{q}'_{nl} = [120{,}237.35,\ 3{,}935.33]$. We define $g(\mathbf{x}) = [\mathbf{r}'_{\mathbf{x}}\mathbf{d}_{\mathbf{x}}, 1'\mathbf{d}_{\mathbf{x}}]'$. Then

$$\mathbf{D}_g(\mathbf{x}) = \begin{bmatrix} \mathbf{r}_{\mathbf{x}} & 1 \\ \mathbf{d}_{\mathbf{x}} & 0 \end{bmatrix}.$$

For the nonlinear case, fixing ratios r_i introduces more zeros into $\mathbf{W}_0^-$ and $\mathbf{W}_{d0}^-$. Setting values for r_i will change the initial $\mathbf{y}_{nl}$ (as opposed to changing **A** for the linear approach).

6. Conclusions and Future Work

In this work, we have provided an overview of constrained estimation and solutions for several common deviance measures based on first principles. While these tools are useful, our main goal was to use them to motivate a framework in which an analyst and a default

Table 5. Values of $\mathbf{A}'$ for Triplets Example

		Default		Path 1		Path 2	
#	$x_{\#}$	$\sum n_i$	$\sum d_i$	$\sum n_i$	$\sum d_i$	$\sum n_i$	$\sum d_i$
1	n_1	1	0	1	0	1	0
2	n_2	1	0	0	0	0	0
3	n_3	1	0	1	0	1	0
4	n_4	1	0	1	0	0	0
5	n_5	1	0	0	0	0	0
6	n_6	1	0	1	0	0	0
7	n_7	1	0	1	0	0	0
8	n_8	1	0	1	0	0	0
9	n_9	1	0	1	0	0	0
10	n_{10}	1	0	1	0	1	0
11	n_{11}	1	0	1	0	0	0
12	n_{12}	1	0	1	0	1	0
13	n_{13}	1	0	1	0	0	0
14	n_{14}	1	0	1	0	1	0
15	n_{15}	1	0	1	0	0	0
16	n_{16}	1	0	0	0	0	0
17	n_{17}	1	0	1	0	0	0
18	n_{18}	1	0	1	0	0	0
19	n_{19}	1	0	1	0	0	0
20	n_{20}	1	0	1	0	0	0
21	d_1	0	1	0	1	0	1
22	d_2	0	1	r_2	1	r_2	1
23	d_3	0	1	0	1	0	1
24	d_4	0	1	0	1	r_4	1
25	d_5	0	1	r_5	1	r_5	1
26	d_6	0	1	0	1	r_6	1
27	d_7	0	1	0	1	r_7	1
28	d_8	0	1	0	1	r_8	1
29	d_9	0	1	0	1	r_9	1
30	d_{10}	0	1	0	1	0	1
31	d_{11}	0	1	0	1	r_{11}	1
32	d_{12}	0	1	0	1	0	1
33	d_{13}	0	1	0	1	r_{13}	1
34	d_{14}	0	1	0	1	0	1
35	d_{15}	0	1	0	1	r_{15}	1
36	d_{16}	0	1	r_{16}	1	r_{16}	1
37	d_{17}	0	1	0	1	r_{17}	1
38	d_{18}	0	1	0	1	r_{18}	1
39	d_{19}	0	1	0	1	r_{19}	1
40	d_{20}	0	1	0	1	r_{20}	1

optimal procedure interact, allowing the user to input extra knowledge to create optimal "user-constrained" results. We demonstrated this framework on a classic *raking* example with linear constraints in the form of margins. We then examined two different approaches to a standard survey problem of constraining aggregate totals and ratios, one implying linear constraints and the other nonlinear ones. Overall, these methods provide a

framework from which to build an interface between automated model processes and expert knowledge via an analyst or metamodel.

We have deliberately avoided discussion of expectations and variances. It should be clear that $\mathbf{y}$ is often stochastic with an estimated distribution (perhaps just a mean and variance). However, $\mathbf{W}$ can be a function of $\mathbf{y}$ (as is often the case for the quadratic measure). More importantly, the user's choice of $\mathbf{q}$, α, and $\mathbf{z}_s$ (and whether or not to use the default settings) is undoubtedly related to both $\mathbf{y}$ and external information. Thus the distribution of $\mathbf{x}$ has connections to both $\mathbf{y}$ and the decision process of the analyst. Tools for finding asymptotic variance estimates when the $\mathbf{y}$ are sampling weights are already available in the literature for calibration (Deville and Särndal 1992; D'Arrigo and Skinner 2010) and would require minor modifications to apply to our setting. However, modeling the uncertainty associated with the decision process of the analyst might first involve implementing these methods and capturing and exploring their behavior. We feel that the framework here is sufficient to begin this process. Data-mining and decision science methods may then be able to construct larger metamodels which incorporate more of these sources of variability.

Appendix A. Justification of $\mathbf{W}_0^-$

Instead of focusing on solving for $\boldsymbol{\lambda}$ and $\mathbf{x}$ (see Subsection 2.2), we assume this is possible and consider the question of whether to modify $\mathbf{q}$ or $\mathbf{W}$ to enforce additional equality constraints $\mathbf{x}_s = \mathbf{y}_s$ or $\mathbf{x}_s = \mathbf{z}_s$ where $\mathbf{x}' = [\mathbf{x}'_{-s}, \mathbf{x}'_s]$ and $\mathbf{y}' = [\mathbf{y}'_{-s}, \mathbf{y}'_s]$ are partitioned and $\mathbf{z}_s \neq \mathbf{y}_s$ is arbitrary.

One option is to augment the $\mathbf{q}$ vector: $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{y}'_s]$ (or $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{z}'_s]$). The corresponding $g(\mathbf{x})$ is augmented $g^{*\prime}(\mathbf{x}) = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$. Then $\mathbf{D}_g(\mathbf{x})$ is also augmented $D_g^*(\mathbf{x}) = [\mathbf{D}_g(\mathbf{x}), \boldsymbol{\delta}]$. We would then use $d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$ (with $\boldsymbol{\lambda}^{*\prime} = [\boldsymbol{\lambda}', \boldsymbol{\eta}']$) and $\mathbf{q}^*$ to solve for $\mathbf{x}$.

Another option is to change the $\mathbf{W}$ or $\mathbf{W}^{-1}$ matrices. Since setting equalities for $\mathbf{x}_s$ should reduce the dimensions of the problem, introducing 0s into $\mathbf{W}$ may also work. We partition $\mathbf{W}$ accordingly and use $\mathbf{W}_0$ and $\mathbf{W}_0^-$ as defined in Subsection 2.3. Note that $\mathbf{W}_0^-$ and $\boldsymbol{\delta}$ are related by the following:

$$\mathrm{W}_0^- = \mathbf{W}^{-1} - \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\mathbf{W}^{-1}.$$

This can be verified using the block inverse formulas to confirm

$$\mathrm{W}_a^{-1} = \{\mathbf{W}^{-1}\}_a - \{\mathbf{W}^{-1}\}_b\{\{\mathbf{W}^{-1}\}_c\}^{-1}\{\mathbf{W}^{-1}\}'_b,$$

where

$$\mathbf{W}^{-1} = \begin{bmatrix} \{\mathbf{W}^{-1}\}_a & \{\mathbf{W}^{-1}\}_b \\ \{\mathbf{W}^{-1}\}'_b & \{\mathbf{W}^{-1}\}_c \end{bmatrix}.$$

A.1. Proof of Lemma 3

This scenario occurs when a user decides that some of the $\mathbf{y}_s$ need to be protected and are kept unchanged during the constraint process. We will show that the $\mathbf{W}_0^-$ and the $\mathbf{D}_g^*(\mathbf{x})$ methods lead to equivalent solutions for $\mathbf{x}_s = \mathbf{y}_s$.

Starting with the $\mathbf{D}_g^*(\mathbf{x})$ equations:

$$\begin{aligned} d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^* \\ d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta} \\ \boldsymbol{\delta}' d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{0}_{n_s} = \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta} \\ \boldsymbol{\eta} &= -(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}. \end{aligned}$$

Then plugging $\boldsymbol{\eta}$ back in:

$$\begin{aligned} d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^* \\ &= \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} - \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}\boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} \\ &= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} \end{aligned}$$

This is the same as substituting $\mathbf{W}_0^-$ for $\mathbf{W}^{-1}$ in the default (no user input) setting.

Not only does the $\mathbf{W}_0^-$ approach give the same results as the $\mathbf{D}_g^*(\mathbf{x})$ approach, it also reduces dimensions instead of increasing them:

$$\begin{aligned} d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} \\ \begin{bmatrix} \{d^{(1)}(\mathbf{x}|\mathbf{y})\}_{-s} \\ \mathbf{0}_{n_s} \end{bmatrix} &= \begin{bmatrix} \{\mathbf{W}_\alpha^{-1}\mathbf{D}_g(\mathbf{x})\}_{-s}\boldsymbol{\lambda} \\ \mathbf{0}_{n_s} \end{bmatrix}. \end{aligned}$$

Thus we only need to keep track of n_{-s} equations and k constraints for the $\mathbf{W}_0^-$ approach instead of n equations and $k + n_s$ constraints with the $\mathbf{D}_g^*(\mathbf{x})$ approach.

A.2. Proof of Lemma 4

Now let us consider the case that $\mathbf{x}_s = \mathbf{z}_s$ for some arbitrary $\mathbf{z}_s \neq \mathbf{y}_s$. There are at least two ways to proceed:

- Create $\mathbf{y}^{*\prime} = [\mathbf{y}'_{-s}, \mathbf{z}'_s]$ and use the $\mathbf{W}_0^-$ approach as in the previous section.
- Keep $\mathbf{y}$ and set $\mathbf{q}^{*\prime} = [\mathbf{q}', \mathbf{z}'_s]$ with $g^{*\prime}(\mathbf{x}) = [g(\mathbf{x})', \mathbf{x}'\boldsymbol{\delta}]$.

We begin with the second option and explore the conditions under which the two are equivalent. For convenience, define $d^{(1)}(\mathbf{x}_s|\mathbf{y}_s) = \{d^{(1)}(\mathbf{x}|\mathbf{y})\}_s$. Also note that $d^{(1)}(\mathbf{x}_s|\mathbf{z}_s) = \mathbf{0}_{n_s}$.

Starting with the $\mathbf{D}_g^*(\mathbf{x})$ equations:

$$\begin{aligned} d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^* \\ d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta} \\ \boldsymbol{\delta}' d^{(1)}(\mathbf{x}|\mathbf{y}) &= d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) = \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta}\boldsymbol{\eta} \\ \boldsymbol{\eta} &= (\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}[d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) - \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}]. \end{aligned}$$

Then

$$
\begin{aligned}
d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^* \\
&= \mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}[d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) - \boldsymbol{\delta}'\mathbf{W}^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}] \\
&= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \mathbf{W}^{-1}\boldsymbol{\delta}(\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta})^{-1}d^{(1)}(\mathbf{z}_s|\mathbf{y}_s) \\
&= \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} + \begin{bmatrix} \{\mathbf{W}^{-1}\}_b\{\{\mathbf{W}^{-1}\}_c\}^{-1} \\ \mathbf{I}_{n_s} \end{bmatrix} d^{(1)}(\mathbf{z}_s|\mathbf{y}_s).
\end{aligned}
$$

When $\mathbf{W}$ is diagonal $\mathbf{W}_d$ (or block-diagonal with $\mathbf{W}_b = \mathbf{0}_{n_{-s}\times n_s}$): $\{\mathbf{W}^{-1}\}_b = \mathbf{0}_{n_{-s}\times n_s}$. Then $\{d^{(1)}(\mathbf{x}|\mathbf{y})\}_{-s} = \mathbf{W}_\alpha^{-1}\{\mathbf{D}_g(\mathbf{x})\}_{-s}\boldsymbol{\lambda}$. So solving $d^{(1)}(\mathbf{x}|\mathbf{y}) = \mathbf{W}^{-1}\mathbf{D}_g^*(\mathbf{x})\boldsymbol{\lambda}^*$ for $\mathbf{x}_{-s}$ is equivalent to solving $d^{(1)}(\mathbf{x}|\mathbf{y}^*) = \mathbf{W}_0^-\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda}$. We would prefer the $\mathbf{y}^*$ method because it allows us to use $\mathbf{W}_0^-$ to reduce dimensions.

When $\mathbf{W}$ is more generally symmetric and invertible (as for the quadratic deviance), we may get two distinct estimates for $\mathbf{x}_{-s}$ from the $\mathbf{D}_g^*(\mathbf{x})$ and $\mathbf{W}_0^-$ approaches. Each approach gives an optimal solution to a set of constraints and slightly different deviance functions. The $\mathbf{W}_0^-$ approach ignores $d^{(1)}(\mathbf{z}_s|\mathbf{y}_s)$, the discrepancy between $\mathbf{y}_s$ and $\mathbf{z}_s$. Whereas the $\mathbf{D}_g^*(\mathbf{x})$ method uses the off-diagonal blocks of $\mathbf{W}$ to incorporate this term.

Appendix B. Justification of $h_{\mathrm{x}}(\mathbf{u})$ for Poisson Deviance

To obtain the $\mathbf{x}$ which minimizes $l(\mathbf{x}|\mathbf{y})$ subject to the constraint $g(\mathbf{x}) = \mathbf{q}$, we derive alternate estimation equations:

$$
\begin{aligned}
d^{(1)}(\mathbf{x}|\mathbf{y}) &= \mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x})\boldsymbol{\lambda} \\
1 - \left[\frac{\mathbf{y}}{\mathbf{x}}\right] &= \mathbf{u} \\
\mathbf{x} - \mathbf{y} &= \langle\mathbf{x}\rangle\mathbf{u} \\
\mathbf{x} &= \mathbf{y} + \langle\mathbf{x}\rangle\mathbf{u} \\
\mathbf{q} &= g(\mathbf{y} + \langle\mathbf{x}\rangle\mathbf{u}).
\end{aligned}
$$

Then $h_{\mathbf{x}}(\mathbf{u}) = [\mathbf{y} + \mathbf{x}\mathbf{u}]$ with $h_{\mathbf{x}}^{(\mathbf{1})}(\mathbf{u}) = \langle\mathbf{x}\rangle$.

Substituting $h_{\mathbf{x}}(\mathbf{u})$ and $h_{\mathbf{x}}^{(\mathbf{1})}(\mathbf{u})$ into (3) and (4), we get an inner iteration

$$
\begin{aligned}
\boldsymbol{\lambda}_i^{j+1} = \boldsymbol{\lambda}_i^j + &\left[\mathbf{D}_g'(\mathbf{x}^i)\mathbf{W}_d^{-1}\langle\mathbf{x}^i\rangle\mathbf{D}_g\left(\mathbf{y} + \langle\mathbf{x}^i\rangle\mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right)\right]^{-1} \\
&\times\left(\mathbf{q} - g\left(\mathbf{y} + \langle\mathbf{x}^i\rangle\mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i^j\right)\right)
\end{aligned}
$$

and an outer iteration

$$
\mathbf{x}^{i+1} = \mathbf{y} + \langle\mathbf{x}^i\rangle\mathbf{W}_d^{-1}\mathbf{D}_g(\mathbf{x}^i)\boldsymbol{\lambda}_i.
$$

We suggest $\mathbf{x}^0 = \mathbf{y}$ and $\boldsymbol{\lambda}_0^0 = 0$ as good initial values, with $\boldsymbol{\lambda}_i^0 = \boldsymbol{\lambda}_{i-1}$ from the previous iteration of $\mathbf{x}^i$.

For the linear case $g(\mathbf{x}) = \mathbf{Ax}$, the inner loop (3) is one step, eliminating $\boldsymbol{\lambda}$:

$$\mathbf{x}^{j+1} = \mathbf{y} + \langle \mathbf{x}^j \rangle \mathbf{W}_d^{-1} \mathbf{A}' \left(\mathbf{A} \mathbf{W}_d^{-1} \langle \mathbf{x}^j \rangle \mathbf{A}' \right)^{-1} (\mathbf{q} - \mathbf{Ay}).$$

We suggest starting with $\mathbf{x}^0 = \mathbf{y}$ since that will give an $\mathbf{x}^1$ which minimizes $\chi^2(\mathbf{x}|\mathbf{y})$ when $\mathbf{W} = \langle \mathbf{y} \rangle^{-1} \mathbf{W}_d$.

7. References

Agresti, A. (2002). Categorical Data Analysis, (2nd edition). New York: Wiley.

Chang, T. and Kott, P. S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika, 95, 555–571.

Chen, J. and Sitter, R. (1999). A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys. Statistica Sinica, 9, 385–406.

Chen, J., Sitter, R., and Wu, C. (2002). Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys. Biometrika, 89, 230–237.

D'Arrigo, J. and Skinner, C. (2010). Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. Survey Methodology, 36, 181–192.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known. Annals of Mathematical Statistics, 11, 427–444.

Deville, J. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Estevao, V., Hidiroglou, M., and Särndal, C.-E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11, 181–204.

Fuller, W. (2002). Regression Estimation for Survey Samples. Survey Methodology, 28, 5–23.

Ghosh, M. (1992). Constrained Bayes Estimation with Applications. Journal of the American Statistical Association, 87, 533–540.

Green, W.H. (2000). Econometric Analysis, (4th edition). New Jersey: Prentice-Hall.

Ireland, C.T. and Kullback, S. (1968). Contingency Tables with Given Marginals. Biometrika, 55, 179–188.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. Survey Methodology, 32, 133–142.

Nandram, B. and Sayit, H. (2011). A Bayesian Analysis of Small Area Probabilities Under a Constraint. Survey Methodology, 37, 137–152.

Pizzinga, A. (2010). Constrained Kalman Filtering: Additional Results. International Statistical Review, 78, 189–208.

Singh, A. and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. Survey Methodology, 22, 107–115.

Stewart, J. (2011). Multivariate Calculus, (7th edition). Belmont, California: Brooks Cole.

Särndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33, 99–119.

Wang, J., Fuller, W., and Qu, Y. (2008). Small Area Estimation Under a Restriction. Survey Methodology, 34, 29–36.

Received September 2012
Revised February 2013
Accepted June 2013

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 397–423

Rapid Estimates of Mexico's Quarterly GDP

Víctor M. Guerrero[1,2]*, Andrea C. García*[1]*, and Esperanza Sainz*[1]

This work presents a procedure for creating a timely estimation of Mexico's quarterly GDP with the aid of Vector Auto-Regressive models. The estimates consider historical GDP data up to the previous quarter as well as the most recent figures available for two relevant indices of Mexican economic activity and other potential predictors of GDP. We obtain two timely estimates of the Grand Economic Activities and Total GDP. Their corresponding delays are at most 15 days and 30 days respectively from the end of the reference quarter, while the first official GDP figure is delayed 52 days. We follow a bottom-up approach that imitates the official calculation procedure applied in Mexico. Empirical validation is carried out with both in-sample simulations and in real time. The mean error of the 30-day delayed estimate of total GDP is 0.13% and its root mean square error is 0.67%. These figures compare favorably with those of no-change models.

Key words: Flash estimates; macroeconomic forecasts; mean square error; timely estimates; time series forecasts; VAR models.

1. Introduction

The National Institute of Statistics and Geography, Statistics Mexico (SM) for short, releases quarterly figures of Mexico's Gross Domestic Product or GDP (referred to as PIBT in Spanish) 50–52 days after the end of the reference quarter. In order to analyze the state of the economy in a timely fashion, we propose an estimate delayed no more than 30 days. Our proposal combines the three most important official sources of information: a) the historical record of subsectors of PIBT from the quarterly System of National Accounts (SNA); b) the most recent monthly figures in the databases of the Index of Global Economic Activity (IGAE in Spanish) and the Monthly Index of Industrial Activity (IMAI in Spanish); and c) some general exogenous indicators, mostly from official sources. Section 2 provides more detailed information on IMAI and IGAE.

Our procedure comes as a response to users' demand of timely data for decision making, a need evidenced by the 2008 world financial crisis. In fact, most users prefer timely

[1] Dirección General del Servicio Público de Información, Instituto Nacional de Estadística y Geografía (INEGI). Av. Patriotismo 711, Torre "A", Piso 9, Col. San Juan Mixcoac, México 03730, D. F., México. Emails: guerrero@itam.mx, andy18cel@gmail.com and maria.sainz@inegi.org.mx

[2] Departamento de Estadística, Instituto Tecnológico Autónomo de México (ITAM). Río Hondo 1, Col. Progreso Tizapán, México 01080, D. F., México.

Acknowledgments: We thank the Associate Editor Christophe Planas and two anonymous referees for detailed and helpful comments that helped us to improve the presentation of this article. This work was facilitated by the useful input and advice of Lourdes Mosqueda from INEGI, who made the databases for the GDP calculation available. Similarly, Blanca R. Sainz, Yuriko Yabuta, Enrique Ordaz and the President of INEGI, Eduardo Sojo, provided the support necessary to carry out this research. Comments and suggestions by Lars-Erik Öller and Gyorgi Gyomai are gratefully acknowledged. V. M. Guerrero participated in this project thanks to a sabbatical year granted by Instituto Tecnológico Autónomo de México (ITAM), to a professorship granted by Asociación Mexicana de Cultura, A. C. and to the generous hospitality of INEGI.

estimates, even at the expense of precision. Rapid estimates are also called "flash estimates" or "timely estimates" and many international meetings have taken place in order to discuss different issues and technicalities related to this topic, the trade-off between timeliness and precision being of utmost relevance. These meetings have been organized in Ottawa (May 2009), Scheveningen (December 2009) and at the Eurostat headquarters in Luxembourg (September 2010). Some of the most important recommendations that came out of those meetings can be summarized as follows: 1) national statistical agencies should provide rapid estimated figures that make use of official information; 2) such figures should be released at the latest with a 30-day delay; 3) to gain credibility with the users, the estimates should be obtained without relying on a specific economic theory; and 4) the estimation procedure should follow essentially the same approach that is used to calculate the final official figures (see, for instance, Kuzin et al. 2010, Mazzi and Montana 2009, Mazzi et al. 2009, Mustapha and Djolov 2010 and UNECE Secretariat 2009).

The following methods have been used to carry out timely estimation:

i) Bridge equations that relate high frequency data (say monthly) with low frequency (say quarterly) data; for example, Klein and Sojo (1989) predicted quarterly US GDP data from monthly indicators and from disaggregated forecasts of demand components, thus obtaining the total GDP forecast by aggregation. Some other applications of bridge equations appear in Rünstler and Sédillot (2003), Baffigi et al. (2004), Zheng and Rossiter (2006), and Diron (2006).
ii) MIDAS (Mixed Data-frequency Sampling) models that use data with different frequencies of observation, as in Ghysels et al. (2004) and Clements and Galvao (2008), or as in Zadrozny (1990).
iii) Diffusion indices that capture the information of a large number of variables by means of a small number of unobserved common factors, as in Klein and Sojo (1989), who used this technique to obtain a single indicator from a set of 25 monthly indicators. Some other examples are those of Forni and Reichlin (1998) and Stock and Watson (2002). An explanation of this methodology can be found in Armah and Swanson (2008).
iv) Dynamic factor models proposed originally by Geweke (1977) and employed recently by Forni et al. (2005) and Aruoba et al. (2009).
v) Forecast combination that averages forecasts of GDP growth obtained with different regression models, as in Kitchen and Monaco (2003).

We decided not to use method (ii) due to the decisions the analyst has to make when applying it, such as parameterization of the polynomial coefficients involved, appropriate choice of the number of lags and whether or not an autoregressive structure is required (e.g., Clements and Galvao 2008). Besides, the nonlinear estimation procedure involved also imposes a computational burden, since we require a method to be applied to a large number of variables in just one day.

Similarly, methods (iii), (iv) and (v) were discarded because we need an estimate of growth for the three Grand Activities, not just for Total PIBT, in order to enhance the possibilities of analysis. Further, the behaviors of these activities differ markedly, as was verified by Mexican data, and therefore have to be estimated separately. Thus we have chosen bridge equations with a bottom-up approach. This is in accordance with the SNA and

approaches the estimation from the side of the use of goods and services, thus contrasting with the demand side approach used in the US to calculate flash estimates (see Katz 2006). Moreover, the bridge equations are not used here to link high frequency with low frequency data; instead we propose to use them to link databases with less coverage (IGAE and IMAI) to another one with more coverage (PIBT), though both contain monthly data. The fact that these three databases contain monthly data will be discussed further in Section 2. Since the original databases lack timely information, we resort to time series models to forecast the unobserved variables at the subsector level. Model adequacy is checked using standard econometric tests and predictive ability is analyzed by way of simulations with real time data vintages, as indicated by Koenig et al. (2003). The simulations are carried out with the estimates derived by aggregation to Grand Economic Activities and Total PIBT.

Section 2 presents the decisions made to solve the modeling and forecasting problems that arise because of the large number of subsectors under consideration. We also consider some features of the databases, timeliness and coverage being essential. Section 3 describes the statistical methods employed, particularly the VAR models. In Section 4 we illustrate the application of our method to a group of sectors of tertiary activities. Here, the databases contain the vintage available as of April 2010. We also show some results of the historical simulations and briefly analyze the estimates of the three Grand Activities and Total PIBT. This section also provides an update of the results currently obtained in real time. Section 5 contains some comments and conclusions that focus on the logistics of routine application of the method. The main conclusion of this work is that it is feasible to use reliable and rapid estimates of Mexico's PIBT, one with a 15-day delay and another one delayed at most 30 days, as recommended by the international statistical community. Comparing these estimates to naïve no-change forecasts, we found the former significantly more accurate. The estimation procedure is relatively easy to use and we consider it applicable in other countries that also need rapid GDP estimates.

2. Grouping of Subsectors and Data Availability

In Mexico, PIBT is calculated by aggregating the monthly Gross Value Added (GVA) of all classes of economic activity into the GVA of sub-branches, then going up from sub-branches to branches, to subsectors, to sectors, to Grand Activities and finally to total GVA. Then the monthly GVA values are added to the quarter to obtain PIBT. Our approach attempts at mimicking the official calculation of PIBT as closely as possible, as recommended in international seminars. However, we start at the subsector level and use a set of decision criteria that allows us to group subsectors as objectively as possible. The classification of economic activities corresponds to production of final goods and services in the country and covers all economic, productive and nonproductive activities, regardless of their profit motives. From here on, we use PIBT and quarterly GVA interchangeably.

According to the North American Industrial Classification System (NAICS) there are 1,051 different classes of economic activity, but only 737 of them are present in Mexico. These classes are grouped into 500 subbranches, 256 branches, 79 subsectors, 20 sectors and three Grand Activities. Due mainly to data availability, at the outset of this study it was decided to start the estimation at the subsector level, that is, estimating the data for groups

of subsectors (the grouping employed is shown in Appendix A). Three groups correspond to primary activities, nine to secondary activities and 17 to tertiary activities. Those groups of activities will be considered as variables whose outcome will be estimated using statistical models. Instead of the word "estimate" we could have used "forecast", but we retain "estimate" as this is the word preferred by the statistical community and it reflects the fact that our estimates are not only based on historical data.

The following criteria are used to group the subsectors:

(a) Subsector share of total value of the sector (or Grand Activity in some cases), for example the livestock subsector was considered as an individual variable because it represents about 35% of the GVA of primary activities, although less than 2% of total GVA.
(b) Impact that the subsector may have on other subsectors; a case in point is mining services. This was taken as a separate variable because it comprises the drilling of wells, an activity that has a direct impact on the subsectors "oil and gas" and "construction of civil engineering works".
(c) Availability of information useful to estimate the subsectors. Several manufacturing subsectors were grouped into one because they lack timely information individually.
(d) Existing relations between different subsectors, such as in the tertiary activities "corporation management and firms" and "businesses support, waste management and remediation services", which are fundamentally related to business activities.

PIBT covers 94% of annual GDP; exceptions are only series reported annually. PIBT differs from IGAE and IMAI in that it is expressed in monetary units (constant pesos at 2003 prices), whereas IGAE and IMAI are released as indices with the base year 2003. For internal purposes, SM generates the IGAE and IMAI databases expressed as GVA at constant prices. We use such monthly disaggregated information as well as some other monthly variables described below. The IMAI database includes industrial activities of sectors 21 to 33 of the NAICS (2007), that is, all secondary activities. Since there is a 42-day gap between the release of information and the month being reported, we can anticipate the figure of PIBT with a 12-day delay using data on two out of the three months of the quarter, estimating month three using time series models.

The IGAE database complements that of IMAI to achieve almost total coverage of PIBT, since it covers all the subsectors that appear in Appendix A except for the few subsectors indicated there. Besides this, IGAE comprises either one or two months of a quarter and its figure is released 57 days after the end of the month of reference. Its coverage is close to 90% of that of PIBT and it provides timely figures before the end of every quarter. Hence, its database can be used to predict PIBT with a 27-day delay when two months of IGAE are available for a quarter. The models that use these data are known as c2 models, while i2 models refer to the use of only one month of IGAE and one month of IMAI (or equivalently two months of IGAE, one of which is incomplete). Figure 1 shows the coverage of the databases and the release dates for a given year "a"; there we see that IMAI has nearly 30% coverage of PIBT, while IGAE's coverage fluctuates around 90%. The IMAI data appears 42 days after the end of a month, for example the figure of November(a-1) is published in January of year "a" and that of October(a) is published in December of year "a". Similarly the IGAE figures are released 57 days after the end of the reference month, except for October whose figure is released in January.

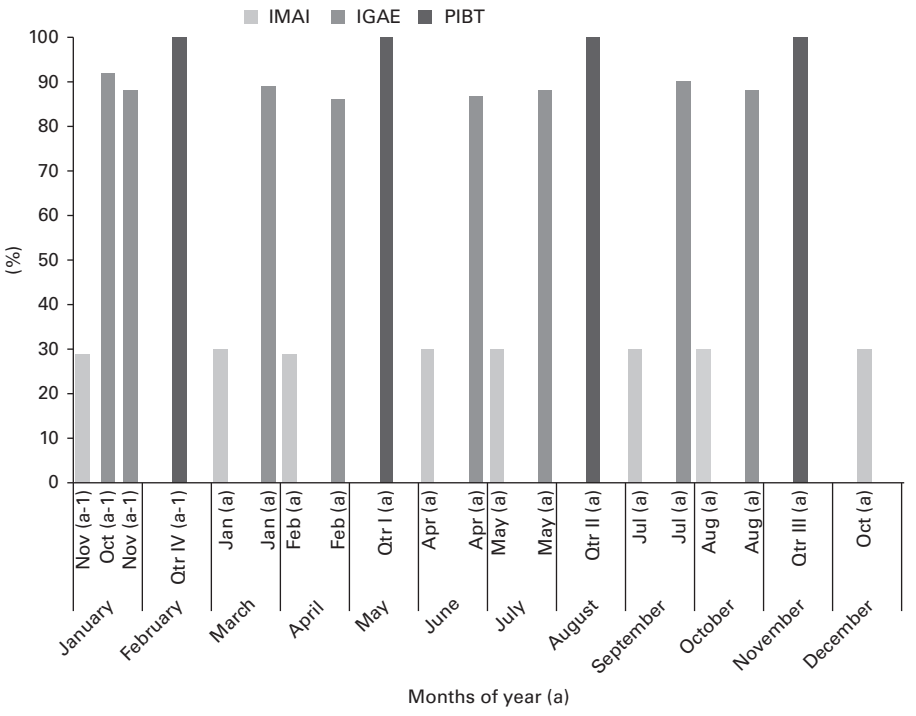


Fig. 1. Months of publication of IMAI, IGAE and PIBT data for a given year "a"

An i2 estimate makes use of 40% of the basic information available on PIBT (30% coming from IGAE and 10% from IMAI), so that we actually have to estimate 60% of the Total PIBT unavailable 12 days after the end of the quarter. Similarly, a c2 estimate uses 2 months of IGAE, that is, 60% of the basic information on PIBT, and therefore we only need to estimate the remaining 40% unavailable 27 days after the end of the quarter. This of course makes a c2 estimate more reliable than the corresponding i2 estimate. Some other official databases and information systems provide potential predictors of the variables leading to the PIBT estimate. They are: Monthly Business Opinion Survey; System of Composite Coincident and Leading Indicators; Consumer Confidence Survey; Trade Balance; and National Occupation and Employment Survey. Another source of information employed is that of the Central Bank of Mexico, as well as some other domestic sources. Finally, the models included dummy variables to capture the effect of such events as Easter, the 2009 swine flu epidemics (AH1N1), a leap year, and level shifts due to annual revisions and benchmarking, as recommended by the International Monetary Fund (see Bloem et al. 2001). A schematic view of the steps followed each quarter to obtain the estimates from both Models i2 and c2 can be seen in Appendix B.

3. Statistical Models and Analysis

The basic tool that we used to generate forecasts is a VAR model, which can be deemed a reduced form representation of a structural equation system without assuming that an economic theory underlies it. Thus we use these models to capture the empirical

regularities in the historical record of the multiple time series under consideration, as well as the interdependencies of the endogenous variables it comprises. Moreover, we emphasise here the well-known predictive ability of a VAR model (see Lütkepohl 2005).

A finite order VAR model can be written as

$$\Pi(\mathrm{B})\mathbf{Z}_t = \Lambda_0\mathbf{D}_t + \Lambda_1\mathbf{X}_t + \ldots + \Lambda_q\mathbf{X}_{t-q} + \mathbf{a}_t \tag{3.1}$$

where $\mathbf{Z}_t = (Z_{1,t}, \ldots, Z_{k,t})'$ is a column vector of k endogenous variables observed at time $t = 1, \ldots, N$, $\Pi(\mathrm{B}) = \mathrm{I}_k - \Pi_1\mathrm{B} - \ldots - \Pi_p\mathrm{B}^p$ is a matrix polynomial of order $p < \infty$, I_k is the identity matrix of order k and $\Pi_1, \ldots, \Pi_p$ are constant parameter matrices, defined as

$$\Pi_j = \begin{pmatrix} \pi_{j,11} & \pi_{j,12} & \ldots & \pi_{j,1k} \\ \pi_{j,21} & \pi_{j,22} & \ldots & \pi_{j,2k} \\ \ldots & \ldots & \ldots & \ldots \\ \pi_{j,k1} & \pi_{j,k2} & \ldots & \pi_{j,kk} \end{pmatrix} \text{for} j = 1, \ldots, p\,. \tag{3.2}$$

The vector $\mathbf{D}_t = (\mathrm{D}_{1,t}, \ldots, \mathrm{D}_{k,t})'$ contains the deterministic elements, such as the constant and dummy variables for events with potential predictive ability on $\mathbf{Z}_t$, while $\mathbf{X}_t, \ldots, \mathbf{X}_{t-q}$ are vectors of lagged ($q \geq 0$) exogenous variables and $\Lambda_0, \ldots, \Lambda_q$ are constant matrices. Finally, $\{\mathbf{a}_t\}$ is assumed to follow a white noise vector process distributed as $\mathbf{a}_t \sim \mathrm{N}_k(\mathbf{0}_k, \Sigma_a)$, where Σ_a is a symmetric matrix with diagonal elements $\mathrm{Var}(a_{it}) = \sigma_i^2$ and off-diagonal elements $\mathrm{Cov}(a_{it}, a_{jt}) = \sigma_{i,j}$, with $i, j = 1, \ldots, k$ and $j \neq i$.

We assume the process is second order stationary and estimate the model by Ordinary Least Squares. We use it to generate optimal, in the sense of minimum Mean Square Error (MSE), linear forecasts conditional on the historical information $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_N)'$, that is,

$$\begin{aligned} \mathrm{E}(\mathbf{Z}_{N+1}|\mathbf{Z}) = {} & \Pi_1\mathbf{Z}_N + \ldots + \Pi_p\mathbf{Z}_{N+1-p} + \\ & \Lambda_0\mathbf{D}_{N+1} + \Lambda_1\mathbf{X}_{N+1} + \ldots + \Lambda_q\mathbf{X}_{N+1-q} \end{aligned} \tag{3.3}$$

where the observations of the exogenous variables are assumed to be known. Thus the MSE matrix of the one-step-ahead forecast is

$$\mathrm{MSE}[\mathrm{E}(\mathbf{Z}_{N+1}|\mathbf{Z})] = \mathrm{Var}(\mathbf{a}_{N+1}|\mathbf{Z}) = \Sigma_a. \tag{3.4}$$

Building VAR models in practice requires first deciding the expression of the variables that will enter the model, bearing in mind that they must be stationary. In our context, the data is seasonally unadjusted, since that is the type of data used to calculate PIBT and it was decided beforehand that a natural expression for the variables had to be like annual (month on month) relative variations, since that is how economic growth is usually interpreted in Mexico. Besides, using seasonally adjusted data would have prevented us from using VAR models, since seasonal adjustment procedures are known to induce noninvertibility of the theoretical models to be employed (see, for instance, Maravall 1993). Hence, it only remained to check whether that transformation produced stationary variables or whether an additional monthly difference had to be used.

Rather than using unit root tests, we decided to apply the monthly difference to all the variables already expressed as annual variations. This decision was taken because the outcomes of these tests are affected by the presence of deterministic effects and structural changes, as indicated by Enders (2003, ch. 4). In our case it was unclear which effects had to be considered, and such effects change as time goes by. Furthermore, the size and power of individual unit root tests are sensible to the presence of error autocorrelation in the model employed by the test (since the first order autoregressive coefficient and its standard error cannot be estimated appropriately in that case).

Thus, rather than performing unit root tests before building the VAR model, we decided to apply the same degree of differencing to all the variables in the system to be modeled. It is clear that this procedure may produce over-differencing, but this is not as serious a problem as that of under-differencing when the model is built for forecasting purposes. In fact, Sánchez and Peña (2001) argue in favor of over-differencing rather than under-differencing when using autoregressive models to generate forecasts. Thus, once the model was estimated we checked that the roots of the corresponding determinantal equation were outside the unit circle. A final and very important argument to support our decision is that we were looking for a generic transformation to be applied to all the variables in the different VAR models, because the process is required to be easy to use in routine applications (every quarter) by the personnel at SM.

Therefore, the variables enter the VAR model expressed in general as

$$Z_t = DO^{IGAE}V_t = \frac{O_t^{IGAE}}{O_{t-12}^{\mathrm{IGAE}}} - \frac{O_{t-1}^{IGAE}}{O_{t-13}^{IGAE}} \tag{3.5}$$

where O_t^{IGAE} is the originally observed variable at time t, coming from the IGAE database, $O^{IGAE}V_t$ is its annual variation and $DO^{IGAE}V_t$ is the monthly difference of the annual variation. It should be clear that we need to apply this transformation to the data in order to build the model, but once the required forecast is obtained we can go back to the original scale with ease by simply applying the inverse transformation. To determine the value p of Model (3.1) we applied sequential likelihood ratio tests. Thus we tested H_0: the order is p *vs*. H_A: the order is p-1, with $p = 4$ as the initial value. We discarded those variables whose estimated coefficient was not significant at the 5% level and checked for no error autocorrelation with the Ljung-Box multivariate statistic $Q*$.

We considered a univariate equation for AGRIC, because this sector follows a pattern completely different from the other economic activities. Data for this sector refers to an agricultural period that starts in October, while the previous agricultural period ends in March of the following year, so that an overlap of six months occurs between two consecutive agricultural periods. This feature is explained by the fact that the Autumn-Winter cycle begins in October and finishes in March of the next year. Harvest usually begins in December and ends the next September. The sowing of the Spring-Summer cycle begins in April and ends in September of the same year, while the first harvest starts in June and finishes in March of the next year.

The model employed is given by

$$\begin{aligned} DAGRICV_t^{IGAE} = {} & \varphi_0 + \varphi_1 DAGRICV_{t-1}^{IGAE} + \ldots + \varphi_m DAGRICV_{t-m}^{IGAE} \\ & + \beta_1 D_{1,t} + \ldots + \beta_r D_{r,t} + \gamma_1 X_{1,t} + \ldots + \gamma_s X_{s,t} + \varepsilon_t \end{aligned} \tag{3.6}$$

with $DAGRICV^{IGAE}$ being the change of the annual variation of AGRIC, with data from the *IGAE* database. We employed bridge equations to link variables coming from the *IGAE* database with the monthly GVA for the subsectors with missing data (see Appendix A). The typical form of a bridge equation is

$$O_t^{GVA} = \alpha_0 + \alpha_1 \hat{O}_t^{IGAE} + \beta_1 D_{1,t} + \ldots + \beta_r D_{r,t} + \gamma_1 X_{1,t} + \ldots + \gamma_s X_{s,t} + \varepsilon_t \qquad (3.7)$$

where O_t^{GVA} is the monthly GVA variable and $\hat{O}_t^{\mathrm{IGAE}}$ is the predicted IGAE variable from the VAR model; the αs, βs and γs are parameters to be estimated and r is the number of deterministic variables (D) such as trend, seasonality and dummies for calendar effects and interventions. Moreover, s is the number of exogenous or predetermined variables (X) with respect to O_t^{GVA}, such as indicator variables of annual level shifts, as well as autoregressive (AR) and moving average (MA) terms. Furthermore, $\{\varepsilon_t\}$ is a sequence of zero-mean non-autocorrelated random errors, in order for Ordinary Least Squares to apply. By using bridge equations we imply that the data for the three months of each quarter have to be estimated.

The statistical models produce forecasts that are considered optimal if they are unbiased and the h-period ahead forecast error behaves as an MA(h-1) model, with $h = 1, 2, \ldots$ (see Diebold 2001, ch. 11). For the VAR models we first obtained the optimal linear forecast with Expression (3.3) and applied the inverse transformation of (3.5) to obtain the forecast in the original scale. The expression used for c2 models is

$$\hat{O}_{N+1}^{IGAE} = O_{N-11}^{IGAE}\left(\hat{Z}_{N+1} + O_N^{IGAE}/O_{N-12}^{IGAE}\right) \qquad (3.8)$$

in which case only one month has to be predicted. For i2 models, two months must be predicted and the corresponding expressions are

$$\begin{aligned} \hat{O}_{N+1}^{IGAE} &= O_{N-11}^{IGAE}\left(\hat{Z}_{N+1} + O_N^{IGAE}/O_{N-12}^{IGAE}\right) \text{ and} \\ \hat{O}_{N+2}^{IGAE} &= O_{N-10}^{\mathrm{IGAE}}\left(\hat{Z}_{N+2} + \hat{O}_{N+1}^{\mathrm{IGAE}}/O_{N-11}^{\mathrm{IGAE}}\right). \end{aligned} \qquad (3.9)$$

The forecast is valid for the original variable from the IGAE database in which case $\hat{O}_{N+h}^{\mathrm{GVA}} = \hat{O}_{N+h}^{IGAE}$ for $h = 1, 2$, when the IGAE database does not lack information on any subsectors. Otherwise, the forecasts from (3.8) and (3.9) are used in the bridge equation (3.7) to obtain the monthly GVA forecast for each month of the quarter. Appendix B provides a schematic view of the estimation procedure employed.

To validate the forecasting ability of our procedure, we carried out nine in-sample simulations (called historical in Appendix C) as well as one out-of-sample (in real time) simulation and analyzed their forecast errors. These were the only possible simulations that could be performed due to data availability. We decided to use a rolling rather than a recursive procedure and produced "the actual forecasts one could make with the model as time progresses" as recommended by Fair and Shiller (1990, p. 376). Thus a six-year rolling window of data was used to estimate the VAR models, because in Mexico there is an approximate six-year cycle in the economy induced by the Presidential elections. Based on this decision we assigned relevance to the most recent information, while still using a sufficiently long stream of data for large sample results to be applicable.

SM provided data only from the year 2003 onwards, because there was a change of base in that year (PIBT data before 2003 had the base year 1993) and this change of base year involved a new classification of products and activities. There was also an update in concepts and procedures, particularly in the information and communication technology sector. These facts ruled out the possibility of joining the old and new PIBT series (we should recall that we required a complete database, including all subsectors). Appendix C shows the dates associated with the data vintages employed and the type of estimates obtained with those databases. We should also stress that the VAR models and bridge equations generate forecasts of the monthly variables, while the purpose of our procedure is to obtain forecasts of PIBT. Thus what really matters is to evaluate the quarterly forecasts, not the monthly ones.

The following forecast errors refer to the estimated PIBT (that is, O^{PIBT}) obtained as the average of the monthly GVA figures of the quarter, including the monthly forecasts. In simulation j, the one-quarter-ahead forecast error with origin in quarter T_j is defined as

$$e_{T_j+1} = O^{PIBT}_{T_j+1} - \hat{O}^{PIBT}_{T_j+1} \text{ for } j = 1, \ldots, J. \tag{3.10}$$

Note that T_j is applicable to quarters, while the subindex t applies to months. We used the following summary measures of forecast errors:

$$\text{Mean Error (ME)}: \; ME(e_1) = \sum\nolimits_{j=1}^{J} e_{T_j+1}/J \tag{3.11}$$

$$\text{Root Mean Square Error (RMSE)}: \; RMSE(e_1) = \sqrt{\sum\nolimits_{j=1}^{J} e^2_{T_j+1}/J} \tag{3.12}$$

$$\text{Theil's U statistic}: \; \text{U} = \frac{\sum_{j=1}^{J} e^2_{T_j+1}}{\sum_{j=1}^{J}\left(O^{\text{PIBT}}_{T_j+1} - O^{\text{PIBT}}_{T_j+1,nc}\right)^2} \tag{3.13}$$

where the alternative naïve forecast involved, $O^{\text{PIBT}}_{T_j+1,nc}$, is obtained on the assumption of no-change in the monthly difference of its annual variation, so that it consists of the average of its three monthly values, each of which is calculated as

$$O^{GVA}_{t_j+k,nc} = O^{GVA}_{t_j+k-12}\left(D\bar{O}^{IGAE}V_k + O^{IGAE}_{t_j+k-1}/O^{IGAE}_{t_j+k-13}\right) \text{for } k = 1, 2, 3. \tag{3.14}$$

This expression serves to calculate the no-change one-month-ahead forecast with origin in month t_j for $j = 1, \ldots, J$ and it is similar to that in (3.8) except that $\hat{Z}$ is now assumed to fluctuate about its mean and is therefore replaced by its average for the corresponding six-year period, $D\overline{O}^{IGAE}V_k$. The ratio of variables from the IGAE database available before the end of the quarter indicates the annual change, while the 12-period lagged GVA variable signals the level of the series. In summary, the no-change forecast of PIBT is obtained as

$$O^{PIBT}_{T_j+1,nc} = \sum\nolimits_{k=1}^{3} O^{GVA}_{t_j+k,nc}/3. \tag{3.15}$$

We do not report the Mean Absolute Error because it provides essentially the same information as the RMSE, as indicated by Granger (1996). A check of predictive ability

can be done with the Mincer-Zarnowitz regression (see Diebold 2001, ch. 11) to verify that all the information in the dataset employed to obtain the forecast was employed efficiently, that is,

$$e_{T_j+1} = \eta_0 + \eta_1 \hat{O}^{PIBT}_{T_j+1} + u_{T_j+1}, \text{ for } j = 1, \ldots, J, \tag{3.16}$$

with u_{T_j+1} a non-autocorrelated random error with mean zero and constant variance for all T_j. Forecast optimality is fulfilled when $\eta_0 = \eta_1 = 0$.

Another check that can be applied when an alternative forecast exists, as in the present case with the no-change forecast, can be obtained using the regression

$$O^{PIBT}_{T_j+1} = \nu_1 \hat{O}^{PIBT}_{T_j+1} + \nu_2 \hat{O}^{PIBT}_{T_j+1,nc} + u_{T_j+1} \text{ for } j = 1, \ldots, J, \tag{3.17}$$

with u_{T_j+1} a random error term, possibly heteroscedastic and autocorrelated. Thus we employed Newey and West's (1987) correction to obtain robust estimates of the standard errors. Now, a forecast-encompassing test is useful to determine whether one of the two forecasts incorporates all the relevant information, as suggested by Fair and Shiller (1990), although Equation (3.17) corresponds to Diebold's (2001, ch. 11) model specification. Thus, if $\nu_I = 1$ and $\nu_2 = 0$, the proposed forecast incorporates the information of the no-change forecast, and the opposite occurs when $\nu_1 = 0$ and $\nu_2 = 1$. For other values of ν_I and ν_2 it is sensible to combine the two forecasts because they both add information.

4. Numerical Illustration

To illustrate the results obtained with the proposed methodology, in what follows we describe its application to a group of subsectors of Tertiary Activities, with the database available on April 27, 2010 that includes two sets of monthly data on IGAE (January and February 2010) so that the sample size covers data from 2004:03 to 2010:02 ($N = 72$).

4.1 Model Estimation Results

The estimation results shown in Table 1 pertain to the c2 model VAR31 that includes four endogenous variables of the tertiary sector: COMER (Trade, including sectors 43–46 of NAICS), TRANS (Transportation, with subsectors 481–488), MENS (Messaging, subsectors 491–492) and ALMAC (Warehousing services, subsector 493). Model estimation was carried out using the computer package EViews7 (Econometric Views version 7, Quantitative Micro Software). Due to the large number of estimated parameters appearing in the VAR models (e.g., in the VAR31 model there are 14 coefficients in each of the four equations, eight of which are associated with the lagged endogenous variables, plus the constant and five coefficients associated with the exogenous variables) we summarize the estimation results in Table 1. Here we can see the order of the VAR model (p) as well as the significance achieved by the (transformed) variables in the left column that explain the variability of the (transformed) variables in the upper row.

In Table 1 we see that COMER explains MENS (at the 5% significance level) and ALMAC (at the 10% level), but it is not explained by any endogenous variable in the system. The significance levels of the endogenous variables come from F tests for all the lags of the variable under consideration. TRANS explains TRANS, MENS and ALMAC

Table 1. Estimation results of model VAR31 (with the Apr10c2 database)

$p = 2$	COMER	TRANS	MENS	ALMAC
COMER	–-	–-	**	*
TRANS	–-	**	**	**
MENS	–-	*	***	–-
ALMAC	–-	–-	–-	**
ITDEMD	***	***	**	–-
ICPFPD(-3)	***	***	**	***
BCEV(-1)	–-	–-	*	***
SEPUGV(-1)	***	***	*	–-
R^2(%)	71.3	69.6	50.8	41.7
$\hat{\sigma}_\varepsilon$	0.04	0.02	0.07	0.05
Q*: Lags (p-value)	12 (0.07)	16 (0.13)	20 (0.29)	24 (0.39)

Notes: *** indicates significant at the 1% level, ** at the 5% level, * at the 10% level and –- non-significant at the 10% level.

(with the indicated significance levels) and is explained by itself and MENS; MENS explains TRANS and MENS, and is explained by COMER, TRANS and MENS; ALMAC serves only to explain its own behavior, and is also explained by COMER and TRANS. The exogenous variables are: ITDEMD (annual difference of the Tendency Indicator of Domestic Demand, coming from the Monthly Business Opinion Survey), which explains all the endogenous variables except ALMAC; ICPFPD(-3) (annual difference of the Producer Confidence Indicator for the Future Economic Situation of the Country, also coming from the Business Opinion Survey), which explains all the endogenous variables with its lag of order 3; BCEV(-1) (annual variation of the Trade Balance Exports lagged one period), which explains MENS and ALMAC; and SEPUGV(-1) (annual variation of the Public Sector Budget Expenditures), which explains three of the four endogenous variables with its first lag.

The lower part of Table 1 shows the percent determination coefficients (lying between 41.7% and 71.3%), the residual standard error for each equation (lying between 0.02 and 0.07), and the last row presents the joint Ljung-Box Q* statistics for different lags, together with their p-values, indicating no residual autocorrelation at the 5% significance level. We remark that timely data coming from opinion surveys were found very useful to explain the endogenous variables in the VAR models employed. In this illustration, the exogenous variables ITDEMD and ICPFPD come from the Business Opinion Survey.

Figure 2 shows time series plots of the transformed series (DCOMERV, DTRANSV, DMENSV and DALMACV) together with their corresponding forecasts for March 2010. These plots allow us to visualize a reasonably stationary behavior of the transformed series.

The corresponding plots in the original scale appear in Figure 3. Data for months 2004:03 through 2009:12 come from the monthly GVA database. COMER_GVA, MENS_GVA and ALMAC_GVA are estimated directly with model VAR31 and their corresponding data from the IGAE database is shown for the period 2010:01–2010:02, while the value for 2010:03 is estimated. On the other hand, for TRANS_GVA we show the estimated values obtained by way of a bridge equation for 2010:01–2010:03. These plots allow us to see that the series do not have a constant level and therefore are in need of

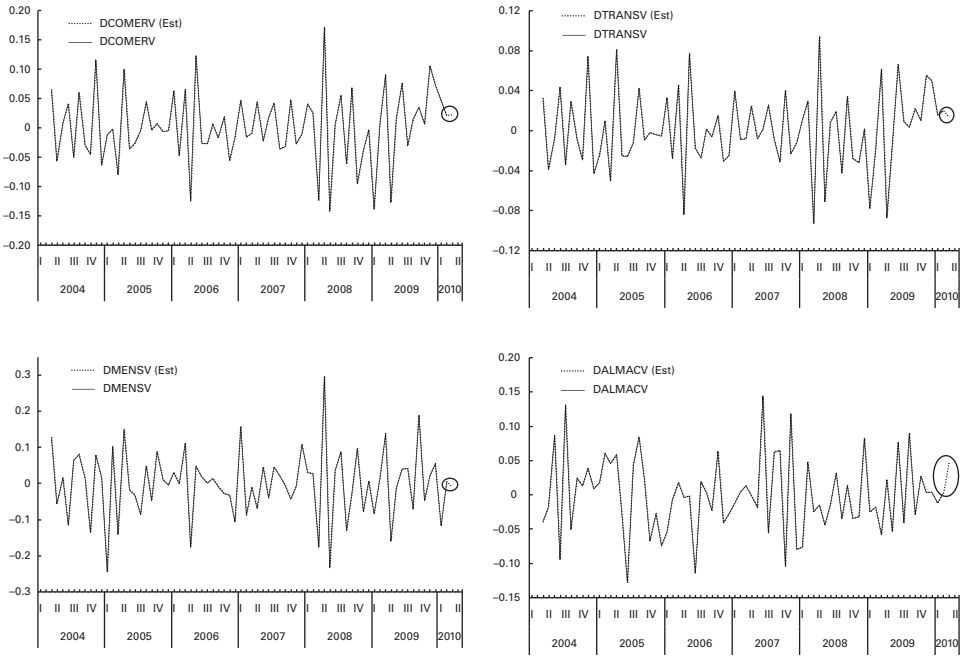


Fig. 2. Monthly transformed variables of model VAR31 from 2004:01 to 2010:02 and estimate for 2010:03

the suggested transformation (the monthly difference of the annual variation) to become approximately stationary. A fall in the level is clearly seen in the upper panels during the last months of 2008 and is less pronounced in the lower panels.

For the VAR31 model, only TRANS requires a bridge equation because subsectors 485 and 488 lack data in the IGAE database, as seen in Appendix A. Figure 4 is useful for appreciating the difference between the series coming from the IGAE and PIBT databases. TRANS_IGAE is the series estimated by the VAR model and contains data up to February 2010, while TRANS_GVA has data up to December 2009 only. Thus, it is necessary to transfer the forecast information from the former to the latter with the aid of a bridge equation that includes a constant, the estimated variable TRANS_IGAE, a dummy variable to account for a level change in year 2005 (*A2005*) and a moving average term of order 12,

$$\begin{aligned} \hat{TRANS}_t^{GVA} &= 60,989,103 + 1.18\hat{TRANS}_t^{IGAE} - 7,844,128A2005_t + 0.85\text{MA}(12) \\ &\quad (4.79) \qquad (40.43) \qquad (-4.52) \qquad (25.15) \end{aligned} \tag{4.1}$$

t statistics appear in parenthesis and indicate significance at the 1% level. Moreover, we obtained $R^2 = 97.6\%$, $\hat{\sigma}_\varepsilon = 5{,}109{,}454$ and the Ljung-Box statistic Q^*: Lags (p-value) 12(0.43), 16(0.30), 20(0.06) and 24(0.09), so that there is no evidence of inadequacy.

In the same way as for the VAR31 model, we estimated the VAR11 model with its bridge equation and the autoregressive equation for the variable AGRIC, the VAR21 and VAR22 models that do not need bridge equations, and the VAR32, VAR33 and VAR34 models with their respective bridge equations.

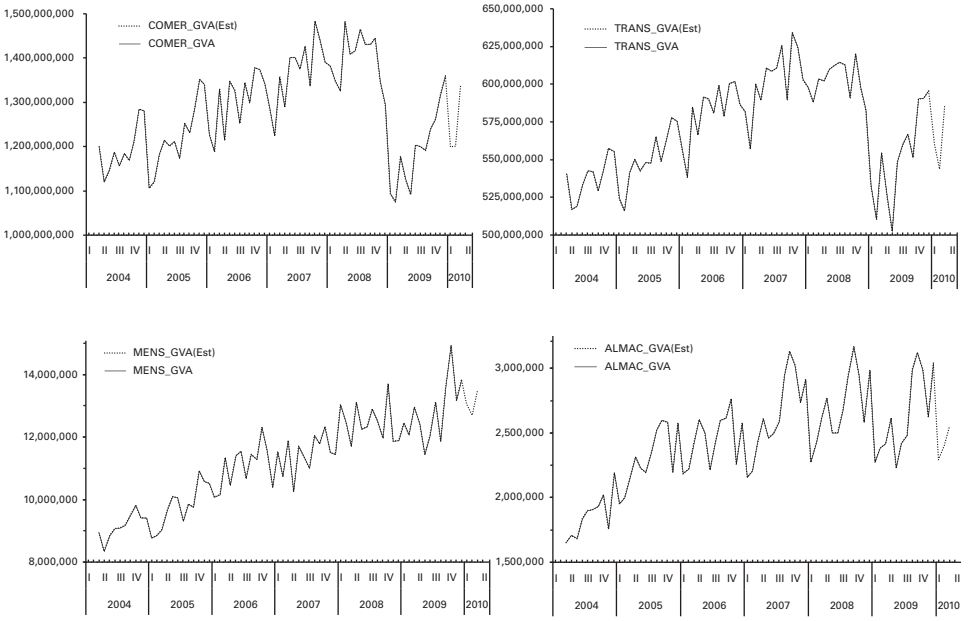


Fig. 3. Variables of the VAR31 model and estimated values in the original scale

4.2 Forecast Evaluation

Evaluation of forecast ability of our procedure was done by simulating using the databases available at the time of reference and using the two models, i2 and c2. Thus nine historical simulations were carried out for quarters 2008:I through 2010:I, as well as one further simulation in real time for quarter 2010:II. Appendix C shows the estimation schedule of

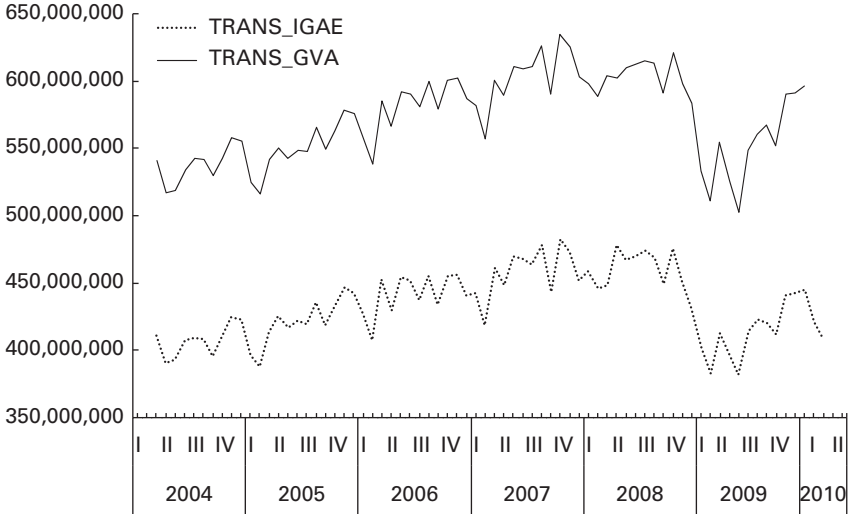


Fig. 4. Original variables TRANS_IGAE from the IGAE database and TRANS_GVA from PIBT

the simulations and the applicable models. The simulation results are shown in Tables 2 and 3 for the three Grand Economic Activities and Total PIBT.

The original data was expressed in thousands of pesos, but the data appearing in the tables is expressed in millions of pesos for clarity of exposition. In Table 2 we see that the ME of Model c2 for Primary Activities is slightly lower than that for Model i2. By looking at the RMSE we can also state that precision is better for Model c2, but the percent estimation errors in Table 3 show that the RMSEs are too high for both models. For Secondary Activities we see in Table 2 that the ME is slightly lower for Model i2 than for Model c2 and the RMSE is also slightly better for Model i2, but the percent estimation errors are essentially the same for both models. This is to be expected, since the IMAI and IGAE databases contain basically the same information for Secondary Activities. What should be emphasized is that the RMSEs for Secondary Activities are substantially lower than those

Table 2. Simulation results for each of the Grand Economic Activities and Total PIBT. Millions of pesos at 2003 value

Quarter	Primary			Secondary		
	Observed	Errors		Observed	Errors	
	data	i2 model	c2 model	data	i2 model	c2 model
2008:I	285,391	−16,980	−19,271	2,653,576	−41,547	−45,492
2008:II	338,570	7,830	−1,500	2,729,747	−24,411	−11,889
2008:III	295,822	−7,491	9,803	2,672,789	−3,965	−3,295
2008:IV	360,094	18,874	6,317	2,624,089	36,581	52,829
2009:I	301,210	−24,260	−4,812	2,427,509	16,123	22,026
2009:II	360,655	−13,316	−3,593	2,457,649	21,659	17,450
2009:III	301,831	7,230	909	2,532,108	−42,667	−42,667
2009:IV	370,113	28,222	29,415	2,591,980	−6,816	−6,816
2010:I	282,657	10,121	5,923	2,547,287	8,149	7,733
2010:II	365,391	12,528	−1,330	2,664,219	33,376	33,362
ME	–-	2,276	2,186	–-	− 352	2,324
RMSE	–-	16,236	12,036	–-	27,299	29,735
	Tertiary			Total PIBT		
2008:I	5,269,578	−29,332	−35,628	8,208,545	−87,859	−100,390
2008:II	5,448,525	−51,630	36,656	8,516,842	−68,211	23,268
2008:III	5,527,957	27,345	18,389	8,496,567	15,888	24,897
2008:IV	5,496,849	−111,122	−49,285	8,481,031	−55,666	9,861
2009:I	4,861,519	124,571	70,840	7,590,238	116,434	88,054
2009:II	4,894,911	29,160	−75,585	7,713,215	37,503	−61,727
2009:III	5,285,423	128,653	82,286	8,119,362	93,217	40,529
2009:IV	5,373,928	85,390	34,587	8,336,021	106,796	57,186
2010:I	5,093,032	32,344	67,047	7,922,976	50,613	80,703
2010:II	5,288,196	−48,668	−67,993	8,297,805	−22,764	−55,961
ME	–-	18,671	8,131	–-	18,595	10,642
RMSE	–-	77,619	57,617	–-	73,390	61,203

Table 3. Simulation results for the Grand Economic Activities. Percent estimation errors

Quarter	Primary		Secondary		Tertiary		Total PIBT	
	i2 model	c2 model	i2 model	c2 model	i2 model	c2 model	i2 model	c2 model
2008:I	−5.95	−6.75	−1.57	−1.71	−0.56	−0.68	−1.07	−1.22
2008:II	2.31	−0.44	−0.89	−0.44	−0.95	0.67	−0.80	0.27
2008:III	−2.53	3.31	−0.15	−0.12	0.49	0.33	0.19	0.29
2008:IV	5.24	1.75	1.39	2.01	−2.02	−0.90	−0.66	0.12
2009:I	−8.05	−1.60	0.66	0.91	2.56	1.46	1.53	1.16
2009:II	−3.69	−1.00	0.88	0.71	0.60	−1.54	0.49	−0.80
2009:III	2.40	0.30	−1.69	−1.69	2.43	1.56	1.15	0.50
2009:IV	7.63	7.95	−0.26	−0.26	1.59	0.64	1.28	0.69
2010:I	3.58	2.10	0.32	0.30	0.64	1.32	0.64	1.02
2010:II	3.43	−0.36	1.25	1.25	−0.92	−1.29	−0.27	−0.67
ME	0.44	0.53	0.00	0.10	0.39	0.16	0.25	0.13
RMSE	4.92	3.62	1.05	1.14	1.49	1.12	0.91	0.77

for Primary Activities and there is no appreciable estimation bias. For Tertiary Activities, both the ME and the RMSE are higher for Model i2 than for Model c2, because the latter model includes more timely information than the former. Again, there does not seem to be any estimation bias (an appropriate statistical test is applied below), and the RMSE of Model c2 is reasonably low and comparable with that obtained for Secondary Activities.

Finally, both ME and RMSE for Total PIBT are larger for Model i2 than for Model c2. Precision and lack of bias are better for this variable than for each of the Grand Activities considered separately in both absolute and relative terms. Furthermore, by looking at the MEs we conclude that Primary Activities is the variable with highest estimation bias although nonsignificant at the 5% level, as shown by the test applied below. Moreover, the RMSEs allow us to appreciate that the Primary Activities estimate has a much lower precision than the other two activities. By contrast, the Total PIBT results are deemed successful because the RMSE for Model c2 is relatively low (0.77%) and there is no estimation bias (0.13%) as compared with each of the Grand Activities.

Some other comparisons of the estimation results are made in the following section. In order to test for significant estimation bias we used Equation (3.16) and obtained the results

Table 4. Checking for the absence of bias with the Mincer-Zarnowitz equation applied to each of the Grand Economic Activities (in millions of pesos at 2003 value)

Model	Statistic	Primary	Secondary	Tertiary	Total PIBT
i2	$\hat{\eta}_0$	−10,611	330,100	1,029,767	1,141,379
	t	−0.17	1.55	2.72	2.77
	$\hat{\eta}_1$	0.0398	−0.1276	−0.1931	−0.1378
	t	0.20	−1.55	−2.67	−2.73
c2	$\hat{\eta}_0$	−5,766	329,555	392,617	493,897
	t	−0.13	1.37	0.91	0.98
	$\hat{\eta}_1$	0.0245	−0.1265	−0.0733	−0.0592
	t	0.18	−1.36	−0.89	−0.96

in Table 4 for each of the Grand Activities and Total PIBT. Model i2 estimates are significantly biased (at the 5% level, since the critical point of a student's t distribution with 8 degrees of freedom is 2.31) for Tertiary Activities and Total PIBT, so that the i2 model underestimates these two variables (about 0.39% and 0.25%, respectively). It should be stressed that Model c2 does not produce significant bias for any economic activity.

4.3 Comparison with the Forecasts from the No-Change Model

In order to validate the precision results empirically, we consider an alternative estimation procedure based on a very simple competing model. In fact, we consider a no-change model for the monthly differences of the annual rates of growth. The IGAE database employed for this very simple model contains two complete months of data, and hence they are comparable only with the results provided by the c2 model. In Table 5 we show the results for the three Grand Economic Activities and Total PIBT with the no-change model.

Table 5. Simulation results with the no-change model for each of the Grand Economic Activities. Millions of pesos at 2003 value

	Primary Activities			Secondary Activities		
Quarter	Observed	Error	Error %	Observed	Error	Error %
2008:I	285,391	3,767	1.32	2,653,576	−78,393	−2.95
2008:II	338,570	4,035	1.19	2,729,747	38,839	1.42
2008:III	295,822	12,915	4.37	2,672,789	−8,215	−0.31
2008:IV	360,094	44,103	12.25	2,624,089	140,673	5.36
2009:I	301,210	−8,154	−2.71	2,427,509	5,603	0.23
2009:II	360,655	−9,269	−2.57	2,457,649	−32,914	−1.34
2009:III	301,831	−4,985	−1.65	2,532,108	40,818	1.61
2009:IV	370,113	12,420	3.36	2,591,980	77,218	2.98
2010:I	282,657	−9,728	−3.44	2,547,287	54,376	2.13
2010:II	365,391	3,049	0.83	2,664,219	24,444	0.92
ME	–-	4,815	1.29	–-	26,245	1.01
RMSE	–-	16,055	4.61	–-	63,094	2.41
	Tertiary Activities			Total PIBT		
2008:I	5,269,578	−81,622	−1.55	8,208,545	−156,248	−1.90
2008:II	5,448,525	14,957	0.27	8,516,842	57,831	0.68
2008:III	5,527,957	25,033	0.45	8,496,567	29,733	0.35
2008:IV	5,496,849	361,597	6.58	8,481,031	546,373	6.44
2009:I	4,861,519	−49,445	−1.02	7,590,238	−51,996	−0.69
2009:II	4,894,911	−10,121	−0.21	7,713,215	−52,304	−0.68
2009:III	5,285,423	165,520	3.13	8,119,362	201,353	2.48
2009:IV	5,373,928	113,003	2.10	8,336,021	202,641	2.43
2010:I	5,093,032	113,771	2.23	7,922,976	158,418	2.00
2010:II	5,288,196	−25,151	−0.48	8,297,805	−17,659	−0.21
ME	–-	62,754	1.15	–-	91,814	1.09
RMSE	–-	139,483	2.58	–-	209,671	2.50

By comparing the MEs of Table 5 with those of Table 2 we see that the no-change model yields higher ME values, indicating a tendency to underestimate PIBT. Moreover, the RMSEs are also higher for the no-change model than for the proposed procedure, lending empirical support to the latter in terms of statistical efficiency. These conclusions are more clearly seen when the errors are expressed as percentages. The no-change estimates for Primary Activities are particularly bad for quarters 2008:III, 2008:IV (with a 12.25% error that was considered inadmissible), 2009:IV and 2010:I. For Secondary Activities, the particularly bad estimates (those with errors greater than 2%) correspond to quarters 2008:I, 2008:IV, 2009:IV and 2010:I, with 5.36% being the highest error. Similarly, for Tertiary Activities the estimation errors greater than 2% appeared in quarters 2008:IV, 2009:III, 2009:IV and 2010:I, with 6.58% as an extremely large error.

We again considered the 2% threshold for Total PIBT and obtained larger estimation errors in the same quarters as before, the largest being 6.44%. The worst estimate provided by the no-change model is that for quarter 2008:IV, which may be due to the worldwide financial crisis. In Table 6 we can see the Theil's U statistics of our procedure against the no-change model. All these statistics are less than unity, indicating a preference for our procedure as being better for Total PIBT than for each of the Grand Economic Activities. Thus, in terms of precision our proposed procedure is better than the no-change model.

Even though Table 6 shows a clear superiority of our procedure, it was deemed convenient to verify that all the relevant information was employed, otherwise we would be able to improve on the estimation by combining the two estimates at hand. To that end we used the encompassing test based on Equation (3.17). Table 7 shows the estimation results of that equation for each of the Grand Economic Activities. There, we confirm that the proposed procedure contains the information provided by the no-change model, since the corresponding calculated t statistics with eight degrees of freedom for that model are smaller than the critical point at the 5% significance level (2.31), except for tertiary activities. On the contrary, the t statistics for the c2 model are all significant at the 5% level. Thus, the naïve model does not contribute any useful information to the estimation in our procedure and there is no reason to combine the two estimates. Notice that the $\hat{\nu}_1$ values for Secondary Activities and Total PIBT are very close to unity, which is to be expected for a good estimate; in fact, when we tested the hypothesis H_0: $\nu_1 = 1$, we did not reject it in any of the four cases (even in the extreme case of Primary Activities the t statistic took on the value 1.33).

4.4 Comparing the Estimation Errors Against PIBT Revisions

In order to judge the magnitude of the estimation errors we compare them with the revisions of PIBT carried out each subsequent quarter at SM. In Tables 8 to 11 we show the

Table 6. Root mean square errors and Theil's U statistics to compare the proposed procedure with the no-change model. Grand Economic Activities in millions of pesos at 2003 value

Method	Primary	Secondary	Tertiary	Total PIBT
Proposal	12,035.8	29,735.5	57,617.0	61,203.0
No-change	16,055.4	63,093.6	139,483.0	211,247.3
Theil's U	0.56	0.22	0.17	0.09

Table 7. Validating the predictive ability of the proposed procedure. Grand Economic Activities

Model	Statistic	Primary	Secondary	Tertiary	Total PIBT
c2	$\hat{\nu}_1$	0.69	0.99	0.85	1.03
vs.	t	2.09	4.02	14.74	10.51
No-	$\hat{\nu}_2$	0.32	0.01	0.16	−0.03
change	t	0.98	0.03	2.67	−0.32

revisions as well as its difference in percentage terms (Revision %). In Mexico, PIBT is also subjected to other revisions (e.g., every year), but the quarterly revisions are the most important for an analysis of the current state of the economy. Hence, we compare those revisions with the estimates coming from the c2 model.

Tables 8 to 11 show a systematic pattern in which the first revision is smaller than the second one and the second revision in turn is smaller than the third one, except in quarter 2009:II for Primary Activities and quarter 2009:I for Total PIBT. In these tables we see that in a given year the following revisions are made:

Quarter I: $I_1 = Rev_1(I)$, $I_2 = Rev_1(I_1) = Rev_2(I)$, $I_3 = Rev_1(I_2) = Rev_2(I_1) = Rev_3(I)$;

Quarter II: $II_1 = Rev_1(II)$, $II_2 = Rev_1(II_1) = Rev_2(II)$; and Quarter III: $III_1 = Rev_1(III)$.

Thus, we have six one quarter behind revisions (revisions of type $Rev_1(X)$, with X a given quarter), three two quarter behind revisions (revision of type $Rev_2(X)$) and one three quarter behind revision (revision of type $Rev_3(X)$). This way, for the years and quarters in our sample we have 13 type $Rev_1(X)$ revisions, six type $Rev_2(X)$ and two type Rev_3 (X), from which we obtain the summary of results shown in Table 12. The differences attributable to revisions are expressed as percentages in order to compare them with the estimation errors of our procedure.

In Table 12 we see that all the MEs are positive, indicating that revisions tend to increase the GVA for all the economic activities. A similar pattern was seen for the estimation errors for both i2 and c2 models (see Tables 2 and 3). We also see that higher percentage revisions occur for Primary Activities and for Secondary Activities, both in

Table 8. PIBT revisions in subsequent quarters after publication. Primary Economic Activities. Millions of pesos at 2003 value

Quarter	Observed data	First revision	Revision %	Second revision	Revision %	Third revision	Revision %
2008:I	285,391	285,915	0.18	286,298	0.32	297,083	4.10
2008:II	338,570	342,337	1.11	356,568	5.32	––	––
2008:III	295,822	298,967	1.06	––	––	––	––
2008:IV	360,094	––	––	––	––	––	––
2009:I	301,210	301,451	0.08	299,714	−0.50	297,247	−1.32
2009:II	360,655	366,265	1.56	362,506	0.51	––	––
2009:III	295,419	296,961	0.52	––	––	––	––
2009:IV	370,113	––	––	––	––	––	––
2010:I	282,657	281,669	−0.35	––	––	––	––

Table 9. PIBT revisions in subsequent quarters after publication. Secondary Economic Activities. Millions of pesos at 2003 value

Quarter	Observed data	First revision	Revision %	Second revision	Revision %	Third revision	Revision %
2008:I	2,653,576	2,654,331	0.03	2,658,227	0.18	2,694,726	1.55
2008:II	2,729,747	2,730,294	0.02	2,778,339	1.78	–-	–-
2008:III	2,672,789	2,712,285	1.48	–-	–-	–-	–-
2008:IV	2,624,089	–-	–-	–-	–-	–-	–-
2009:I	2,427,509	2,429,546	0.08	2,429,901	0.10	2,416,358	−0.46
2009:II	2,457,649	2,459,517	0.08	2,453,219	−0.18	–-	–-
2009:III	2,532,108	2,522,487	−0.38	–-	–-	–-	–-
2009:IV	2,591,980	–-	–-	–-	–-	–-	–-
2010:I	2,547,287	2,547,909	0.02	–-	–-	–-	–-

terms of MEs or RMSEs. However, the reasons for such revisions are different: for Primary Activities there is a lack of data and any new piece of information may substantially change what was already published, while for Secondary Activities there is a great deal of timely data and the database is continually updated.

We can also observe an increase in the percentages by going from one quarter behind revisions to two quarter behind and three quarter behind revisions. Nevertheless, since there are more one quarter behind revisions than other types of revisions, we cannot trust all of them equally and thus we prefer to look at the present results only as indicative of what should be studied more deeply in future work focusing on revisions of PIBT. By looking at the RMSEs in Table 12 we appreciate a decrease in magnitude from Primary Activities to Total PIBT as in Tables 2 and 3. Moreover, the proportion of the third revision with respect to the estimation error of our procedure is 0.8 for Primary Activities, 1.0 for Secondary Activities, 0.4 for Tertiary Activities and 0.7 for Total PIBT, so that our estimates are as precise as the third revision for Secondary Activities. Similarly, our estimates for Primary Activities are slightly less precise than the third revision; the same thing happens with Total PIBT, and the lowest precision occurs when estimating Tertiary Activities.

Table 10. PIBT revisions in subsequent quarters after publication. Tertiary Economic Activities. Millions of pesos at 2003 value

Quarter	Observed data	First revision	Revision %	Second revision	Revision %	Third revision	Revision %
2008:I	5,269,578	5,268,424	−0.02	5,259,868	−0.18	5,277,294	0.15
2008:II	5,448,525	5,441,312	−0.13	5,458,024	0.17	–-	–-
2008:III	5,527,957	5,537,215	0.17	–-	–-	–-	–-
2008:IV	5,496,849	–-	–-	–-	–-	–-	–-
2009:I	4,861,519	4,874,842	0.27	4,885,200	0.49	4,892,965	0.65
2009:II	4,894,911	4,900,607	0.12	4,913,207	0.37	–-	–-
2009:III	5,192,144	5,198,930	0.13	–-	–-	–-	–-
2009:IV	5,373,928	–-	–-	–-	–-	–-	–-
2010:I	5,093,032	5,093,048	0.00	–-	–-	–-	–-

Table 11. PIBT revisions in subsequent quarters after publication. Total PIBT. Millions of pesos at 2003 value

Quarter	Observed data	First revision	Revision %	Second revision	Revision %	Third revision	Revision %
2008:I	8,208,545	8,208,671	0.00	8,204,393	−0.05	8,269,103	0.74
2008:II	8,516,842	8,513,943	−0.03	8,592,930	0.89	--	--
2008:III	8,496,567	8,548,467	0.61	--	--	--	--
2008:IV	8,481,031	--	--	--	--	--	--
2009:I	7,590,238	7,605,840	0.21	7,614,814	0.32	7,606,570	0.22
2009:II	7,713,215	7,726,389	0.17	7,7289,320	0.20	--	--
2009:III	8,019,672	8,018,378	−0.02	--	--	--	--
2009:IV	8,336,021	--	--	--	--	--	--
2010:I	7,922,976	7,922,626	−0.00	--	--	--	--

4.5 An Update for Quarters 2010:III to 2011:IV

Since the procedure has been applied in a routinely manner, the results in Tables 13 and 14 complement those of Tables 2 and 3. The ME and RMSE measures in the new tables were obtained with data from 2008:I to 2011:IV and show a decrease of the RMSE for Model c2, especially for Total PIBT (from 0.77% in Table 3 to 0.67% in Table 14). These results lend further empirical support to our suggested procedure.

5. Final Comments

The proposed estimation procedure starts every quarter as soon as the IMAI and IGAE data is released, 12 and 27 days after the end of the reference quarter respectively. In order to do this, the exogenous variables already have to be updated in the databases and once the data is in the form required by the models it is possible to estimate them with a six-year rolling window of data. The underlying assumptions of the models have to be verified and their specifications changed if necessary. The first models to be estimated for a given quarter are of type i2 and their most recent specifications are those of the c2 models for the previous quarter. Therefore, the i2 specification incorporates three additional months of data, during which time the economic system may have undergone abrupt changes, whereas the c2 specification is simpler because it is carried out only 15 days after the most recent i2 estimation and only a few data updates occur.

Table 12. Summary of the quarterly percent revisions for the Grand Economic Activities

Revision type	Primary %	Secondary %	Tertiary %	Total PIBT %
		ME		
$Rev_1(X)$	0.59	0.19	0.08	0.13
$Rev_2(X)$	1.36	0.48	0.23	0.35
$Rev_3(X)$	1.39	0.55	0.40	0.48
		RMSE		
$Rev_1(X)$	0.86	0.58	0.15	0.25
$Rev_2(X)$	2.77	0.99	0.32	0.50
$Rev_3(X)$	3.04	1.14	0.47	0.54

Table 13. Simulation results for each of the Grand Economic Activities and Total PIBT. Millions of pesos at 2003 value

Quarter	Primary			Secondary		
	Observed	Errors		Observed	Errors	
	data	i2 model	c2 model	data	i2 model	c2 model
2010:III	298,073	449	−8,796	2,688,324	−7,349	−7,060
2010:IV	371,926	15,708	7,261	2,718,258	946	114
2011:I	287,045	−13,557	191	2,684,995	32,940	32,905
2011:II	328,311	−19,067	−17,529	2,750,829	−892	−837
2011:III	311,353	9,631	−4,593	2,772,088	11,652	11,651
2011:IV	337,429	−36,673	−6,579	2,799,920	23,620	23,623
ME	–-	−1,297	−511	–-	3,587	5,227
RMSE	–-	17,443	11,041	–-	24,092	25,823
	Tertiary			Total PIBT		
2010:III	5,507,938	112,567	44,867	8,494,335	105,667	29,011
2010:IV	5,666,809	7,913	−21,707	8,756,994	24,567	−14,332
2011:I	5,362,853	68,021	29,238	8,334,892	87,404	62,335
2011:II	5,507,979	−12,689	−42,741	8,587,119	−32,649	−61,107
2011:III	5,746,740	11,340	21,791	8,830,181	32,623	28,850
2011:IV	5,880,205	33,607	2,001	9,017,554	20,554	19,046
ME	–-	25,467	7,173	–-	26,507	10,639
RMSE	–-	70,280	49,271	–-	68,840	54,383

The procedure does not allow calculation of variances for the estimates, because model estimation is not carried out simultaneously but for separate groups of variables. An important line of future work would consider solving this deficiency. Another possibility for future methodological research that may improve the forecasting ability of the models lies in recognizing that the transformations applied to stationarize the series are monotonic

Table 14. Simulation results for the Grand Economic Activities. Percent estimation errors

Quarter	Primary		Secondary		Tertiary		Total PIBT	
	i2 model	c2 model	i2 model	c2 model	i2 model	c2 model	i2 model	c2 model
2010:III	0.15	−2.95	−0.27	−0.26	2.04	0.81	1.24	0.34
2010:IV	4.22	1.95	0.03	0.00	0.14	−0.38	0.28	−0.16
2011:I	−4.72	0.07	1.23	1.23	1.27	0.55	1.05	0.75
2011:II	−5.81	−5.34	−0.03	−0.03	−0.23	−0.78	−0.38	−0.71
2011:III	3.09	−1.48	0.42	0.42	0.20	0.38	0.37	0.33
2011:IV	−10.87	−1.95	0.84	0.84	0.57	0.03	0.23	0.21
ME	−0.60	−0.28	0.14	0.20	0.49	0.14	0.33	0.13
RMSE	5.26	3.34	0.92	0.99	1.33	0.95	0.84	0.67

and nonlinear. Thus, by back-transforming to the original scale we induce some bias in the estimation that may be corrected, at least approximately, as in Guerrero (1993). Recently, Ghysels (2012) generalized the MIDAS approach to a Vector Auto-Regressive (VAR) setting and since such an approach is in line with ours, we should try it in future work.

The main conclusion of this work is that not only can we obtain timely estimates of Mexico's PIBT, but the resulting estimates are reasonably precise, as indicated by the comparison criteria employed. It is also clear that the 15-day delay estimate of Secondary Economic Activities PIBT is more precise than the estimates of the other two Grand Economic Activities. With a 30-day delay, the estimate of Secondary Activities remains reasonably precise and we can also obtain a good estimate of Tertiary Economic Activities. However, there is room for improvement in the Primary and Tertiary Activities estimates and some additional effort has to be made to obtain more useful and timely information for the sectors involved in those activities. Thus, we advise SM to make some extra effort to improve the data collection in the agriculture sector and design opinion surveys to collect anticipatory data in the commerce and service sectors.

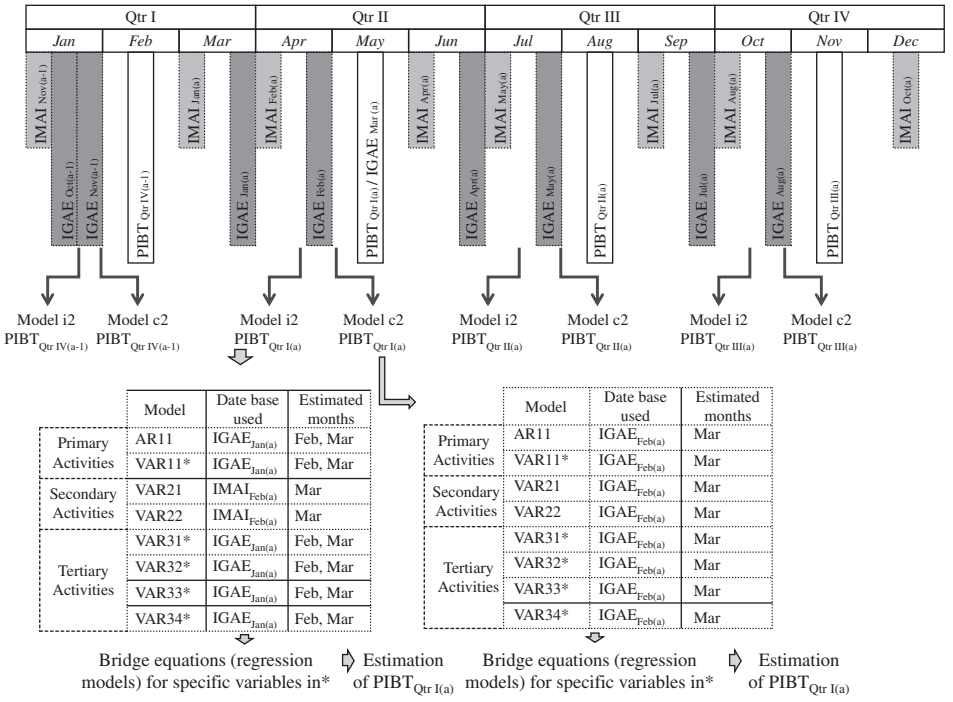
An advantage of the indirect approach employed here is that we could improve on the estimation of one of the Grand Activities without any need to modify the estimation of the other two. Nevertheless, it should be emphasized that the estimate of Total PIBT is reasonably good and better than each of the Grand Economic Activities estimates considered separately, both for the 15-day and 30-day delay estimation. The people in charge of operating the timely estimation system must be alert to the possibility of having access to more timely data and to some other useful indicators not yet employed by the models considered in this work in the future.

Appendix A. *Grouping of Subsectors, With NAICS Codes. Taken from INEGI (2007)*

Primary Activities
AGRIC (111) – Agriculture
GANAD (112) – Animal breeding and production
FOPEC* (113–115) – Forestry, logging, fishing and hunting
Secondary Activities
EXPYG (211) – Oil and gas extraction
MINER (212) – Mining
SEMIN (213) – Services related to mining
ELAGA (221–222) – Electric power generation, water and gas supply
CONST (236–238) – Construction
FDPYC (324) – Manufacturing of products derived from petroleum and coal
INQUI (325) – Chemical industry
FETRA (336) – Transportation equipment manufacturing
MANUF (311–316, 321–323, 326–327, 331–335, 337, 339) – Other manufacturing activities
Tertiary Activities
COMER (43–46) – Trade
TRANS* (481–488) – Transportation
MENS (491–492) – Messaging
ALMAC (493) – Warehousing services
TELEC* (511–512, 515–516, 518–519) – Mass media communication
OTELE* (517) – Other telecommunications
SEFIN* (521–524) – Financial and insurance services
SEINM* (531–533) – Real estate services and goods rental
SEPRO* (541) – Professional, scientific and technical services
CONED* (551, 561–562) – Head offices and business support services
SEDUC (611) – Educational services
SESAL* (621–624) – Health care and social assistances services
SEREC (711–713) – Recreation services
SEHOR (721–722) – Temporary accommodation services
SEOT* (811–814) – Other services
ACGOB* (931) – Government activities
SIFMI – Financial intermediation services indirectly measured

∗These variables lack information on some subsectors and require the use of bridge equations.

Appendix B. *Estimation Procedure Employed for Models i2 and c2 in a given Year "a"*



Appendix C. *Estimation Schedule for the Simulations (Historical and in Real Time)*

Simulation No. and type	Data available	Estimation date	Model type	PIBT estimate
1	IMAI	Apr/17/08	i2	2008:I
Historical	IGAE	Apr/29/08	c2	
2	IMAI	Jul/17/08	i2	2008:II
Historical	IGAE	Jul/29/08	c2	
3	IMAI	Oct/17/08	i2	2008:III
Historical	IGAE	Oct/29/08	c2	
4	IMAI	Jan/16/09	i2	2008:IV
Historical	IGAE	Jan/28/09	c2	
5	IMAI	Apr/17/09	i2	2009:I
Historical	IGAE	Apr/28/09	c2	
6	IMAI	Jul/17/09	i2	2009:II
Historical	IGAE	Jul/28/09	c2	
7	IMAI	Oct/16/09	i2	2009:III
Historical	IGAE	Oct/28/09	c2	
8	IMAI	Jan/12/10	i2	2009:IV
Historical	IGAE	Jan/27/10	c2	
9	IMAI	Apr/12/10	i2	2010:I
Historical	IGAE	Apr/27/10	c2	
10	IMAI	Jul/12/10	i2	2010:II
Real time	IGAE	Jul/27/10	c2	

6. References

Armah, N.A. and Swanson, N.R. (2008). Seeing Inside the Black Box: Using Diffusion Index Methodology to Construct Factor Proxies in Large Scale Macroeconomic Time Series Environments. Federal Reserve Bank of Philadelphia Working Paper No. 08–25.

Aruoba, S.B., Diebold, F.X., and Scotti, C. (2009). Real-Time Measurement of Business Conditions. Journal of Business and Economic Statistics, 27, 417–427.

Baffigi, A., Golinelli, R., and Parigi, G. (2004). Bridge Model to Forecast the Euro Area GDP. International Journal of Forecasting, 20, 447–460.

Bloem, A.M., Dippelsman, R.J., and Maehle, N.O. (2001). Manual de Cuentas Nacionales Trimestrales. Conceptos, fuentes de datos y compilación. Washington, D.C. International Monetary Fund.

Clements, M.P. and Galvao, A.B. (2008). Macroeconomic Forecasting with Mixed-Frequency Data: Forecasting Output Growth in the United States. Journal of Business and Economic Statistics, 26, 546–554.

Diebold, F.X. (2001). Elements of Forecasting, (2nd Edition). Cincinnati: South-Western.

Diron, M. (2006). Short-Term Forecasts of Euro Area Real GDP Growth. An Assessment of Real-Time Performance Based on Vintage Data. European Central Bank Working Paper No. 622.

Enders, W. (2003). Applied Econometric Time Series, (2nd Edition). New York: Wiley.

Fair, R.C. and Shiller, J. (1990). Comparing Information in Forecasts from Econometric Models. The American Economic Review, 80, 375–389.

Forni, M. and Reichlin, L. (1998). Let's Get Real: A Dynamic Factor Analytical Approach to Disaggregated Business Cycle. Review of Economic Studies, 65, 453–474.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. Journal of the American Statistical Association, 100, 830–840.

Geweke, J. (1977). The Dynamic Factor Analysis of Economic Time Series. Latent Variables in Socio-Economic Models, D.J. Aigner and A.S. Goldberger (eds). Amsterdam: North-Holland.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models. Unpublished manuscript, University of North Carolina. Available at: http://www.unc.edu/eghysels (accessed April, 2013).

Ghysels, E. (2012). Macroeconomics and the Reality of Mixed Frequency Data. Chapel Hill: Manuscript, University of North Carolina.

Granger, C.W.J. (1996). Can We Improve the Perceived Quality of Economic Forecasts? Journal of Applied Econometrics, 11, 455–473.

Guerrero, V.M. (1993). Time Series Analysis Supported by Power Transformations. Journal of Forecasting, 12, 37–48.

INEGI (2007). Sistema de Clasificación Industrial de América del Norte, México. SCIAN 2007, (Third edition). México: Instituto Nacional de Estadística, Geografía e Informática.

Katz, A.J. (2006). An Overview of BEA's Source Data and Estimating Methods for Quarterly GDP. Paper prepared for the 10th OECD-NBS Workshop on National Accounts. Paris, France, November 6-10, 2006.

Kitchen, J. and Monaco, R. (2003). Real-Time Forecasting in Practice: The U.S. Treasury Staff's Real-Time GDP Forecast System. Business Economics, October, 10–19.

Klein, L.R. and Sojo, E. (1989). Combinations of High and Low Frequency Data in Macroeconometric Models. Economics in Theory and Practice: An Eclectic Approach, L.R. Klein and J. Marquez (eds). Norwell, MA: Kluwer Academic Publishers.

Koenig, E.F., Dolmas, S., and Piger, J. (2003). The Use and Abuse of Real-Time Data in Economic Forecasting. The Review of Economics and Statistics, 85, 618–628.

Kuzin, V., Marcellino, M., and Schumacher, Ch. (2010). Pooling Versus Model Selection for Nowcasting with Many Predictors: An Application to German GDP. Presented at the 6th Colloquium on Modern Tools for Business Cycle Analysis. Luxembourg, 26–29 September. Available at: epp.eurostat.ec.europa.eu/portal/page/portal/euroindicators_conferences/6th_colloquium/ program_papers (accessed April 2013).

Lütkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Berlin: Springer-Verlag.

Maravall, A. (1993). Stochastic Linear Trends, Models and Estimators. Journal of Econometrics, 56, 5–37.

Mazzi, G.L. and Montana, G. (2009). A System of Rapid Estimates to Improve Real Time Monitoring of the Economic Situation: The Case of the Euro Area. Presented at the Seminar on Timeliness, methodology and comparability of rapid estimates of economic trends. Ottawa, 27-29 May. Available at: unstats.un.org/unsd/nationalaccount/workshops/2009/ottawa/ac188-2.asp (accessed April 2013).

Mazzi, G.L., Mitchell, J., Mouratidis, K., and Weale, M. (2009). The Euro-Area Recession and Nowcasting GDP Growth Using Statistical Models. Presented at the International Seminar on Early Warning and Business Cycle Indicators. Scheveningen, 14–16 December. Available at: unstats.un.org/unsd/nationalaccount/workshops/2009/netherlands/ac202-2.asp (accessed April 2013).

Mustapha, N. and Djolov, G. (2010). The Development and Production of GDP Flash Estimates in a Newly Industrialised Country: The Case of South Africa. Presented at the 6th Colloquium on Modern Tools for Business Cycle Analysis. Luxembourg, 26–29 September. Available at: epp.eurostat.ec.europa.eu/portal/page/portal/euroindicators_conferences/6th_colloquium/ program_papers (accessed April 2013).

Newey, W.K. and West, K. (1987). A Simple Positive Semi-Definite Heteroscedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55, 703–708.

Rünstler, G. and Sédillot, F. (2003). Short-Term Estimates of Euro Area Real GDP by Means of Monthly Data. European Central Bank Working Paper No. 276.

Sánchez, I. and Peña, D. (2001). Properties of Predictors in Overdifferenced Nearly Nonstationary Autoregression. Journal of Time Series Analysis, 22, 45–66.

Stock, J.H. and Watson, M.W. (2002). Macroeconomic Forecasting Using Diffusion Indices. Journal of Business and Economic Statistics, 20, 147–162.

UNECE Secretariat (2009). Rapid Estimates of GDP in CIS and Western Balkan Countries. Presented at the Seminar on Timeliness, methodology and comparability of rapid estimates of economic trends. Ottawa, 27-29 May. Available at: unstats.un.org/unsd/nationalaccount/workshops/2009/ottawa/ac188-2.asp (accessed April 2013).

Zadrozny, P.A. (1990). Estimating a Multivariate ARMA Model with Mixed-Frequency Data: An Application to Forecasting U.S. GNP at Monthly Intervals. Federal Reserve Bank of Atlanta Working Paper No. 90-6.

Zheng, I.Y. and Rossiter, J. (2006). Using Monthly Indicators to Predict Quarterly GDP. Bank of Canada Working Paper 2006-26.

Received November 2011
Revised September 2012
Accepted May 2013

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 425–465

Statistical Analysis of Noise-Multiplied Data Using Multiple Imputation

Martin Klein[1] *and Bimal Sinha*[2]

A statistical analysis of data that have been multiplied by randomly drawn noise variables in order to protect the confidentiality of individual values has recently drawn some attention. If the distribution generating the noise variables has low to moderate variance, then noise-multiplied data have been shown to yield accurate inferences in several typical parametric models under a formal likelihood-based analysis. However, the likelihood-based analysis is generally complicated due to the nonstandard and often complex nature of the distribution of the noise-perturbed sample even when the parent distribution is simple. This complexity places a burden on data users who must either develop the required statistical methods or implement the methods if already available or have access to specialized software perhaps yet to be developed. In this article we propose an alternate analysis of noise-multiplied data based on multiple imputation. Some advantages of this approach are that (1) the data user can analyze the released data as if it were never perturbed, and (2) the distribution of the noise variables does not need to be disclosed to the data user.

Key words: Combining rules; confidentiality; rejection sampling; statistical disclosure limitation; top coded data.

1. Introduction

When survey organizations and statistical agencies such as the U.S. Census Bureau release microdata to the public, a major concern is the control of disclosure risk, while ensuring fairly high quality and utility in the released data. Very often some popular statistical disclosure limitation (SDL) methods such as data swapping, multiple imputation, top/bottom coding (especially for income data), and perturbations with random noise are applied before releasing the data. Rubin (1993) proposed the use of the multiple imputation method to create synthetic microdata which would protect confidentiality by replacing actual microdata by random draws from a predictive distribution. Since then, rigorous statistical methods to use synthetic data for drawing valid inferences on relevant population parameters have been developed and used in many contexts (Little 1993;

[1] Martin Klein (Email: martin.klein@census.gov) is Research Mathematical Statistician in the Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, U.S.A.
[2] Bimal Sinha (Email: sinha@umbc.edu) is Research Mathematical Statistician in the Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, U.S.A. and Professor in the Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, U.S.A.
Acknowledgments: The authors thank Eric Slud for carefully reviewing the manuscript; Jerry Reiter for some valuable discussions; four anonymous referees and an associate editor for many helpful comments; and Joseph Schafer, Yves Thibaudeau, Tommy Wright and Laura Zayatz for encouragement. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Raghunathan et al. 2003; Reiter 2003, 2005; Reiter and Raghunathan 2007). An and Little (2007) also suggested multiple imputation methods as an alternative to top coding of extreme values and proposed two methods of data analysis with examples.

Noise perturbation of original microdata by addition or multiplication has also been advocated by some statisticians as a possible data confidentiality protection mechanism (Kim 1986; Kim and Winkler 1995, 2003; Little 1993), and recently there has been a renewed interest in this topic (Nayak et al. 2011; Sinha et al. 2012). In fact, Klein, Mathew, and Sinha (2013), hereafter referred to as Klein et al. (2013), developed likelihood-based data analysis methods under noise multiplication for drawing inference in several parametric models. They provided a comprehensive comparison of the above two methods, namely, multiple imputation and noise multiplication. Klein et al. (2013) commented that while standard and often *optimum* parametric inference based on the original data can be easily drawn for simple probability models, such an analysis is far from being close to optimum or even simple when noise multiplication is used. Hence their statistical analysis is essentially based on the asymptotic theory, requiring computational details of maximum likelihood estimation and calculations of the observed Fisher information matrices. Klein et al. (2013) also developed a similar analysis for top-coded data, which arise in many instances such as income and profit data, where values above a certain threshold C are coded and only the number m of values in the data set above C are reported along with all the original values below C. These authors considered statistical analysis based on unperturbed (i.e., original) data below C and noise-multiplied data above C instead of completely ignoring the data above C, and again provided a comparison with the statistical analysis reported in An and Little (2007), who carried out the analysis based on multiple imputation of the data above C in combination with the original values below C. In this article, we use the term *mixture* data, to refer to a data set in which values below a cut-off C are unperturbed, and values above C are perturbed via noise multiplication.

In the context of data analysis under noise perturbation, if the distribution generating the noise variables has low to moderate variance, then noise-multiplied data are expected to yield accurate inferences in some commonly used parametric models under a formal likelihood-based analysis (Klein et al. 2013). However, as noted by Klein et al. (2013), the likelihood-based analysis is generally complicated due to the nonstandard and often complex nature of the distribution of the noise-perturbed sample even when the parent distribution is simple (a striking example is analysis of noise-multiplied data under a *Pareto* distribution, typically used for income data, which we hope to address in a future communication). This complexity places a burden on data users who must either develop the required statistical methods or implement these methods if already available or have access to specialized software perhaps yet to be developed. Circumventing this difficulty is essentially the motivation behind this current research, where we propose an alternate simpler analysis of noise-multiplied data based on the familiar notion of multiple imputation. We believe that a proper blend of the two statistical methods as advocated here, namely, noise perturbation to protect confidentiality and multiple imputation for ease of subsequent statistical analysis of noise-multiplied data, will prove to be quite useful to both statistical agencies and data users. Some advantages of this approach are that (1) the data user can analyze the released data as if it were never perturbed (in conjunction with

the appropriate multiple imputation combining rules), and (2) the distribution of the noise variables does not need to be disclosed to the data user.

The article is organized as follows. An overview of our proposed approach based on a general framework of fully noise-multiplied data is given in Section 2. Techniques of noise imputation from noise-multiplied data, which are essential for the proposed statistical analysis, are also presented in Section 2. This section also includes different methods of estimation of variance of the proposed parameter estimates. Section 3 contains our statistical analysis for *mixture* data. Details of computations for the normal and lognormal models are outlined in Section 4. An evaluation and comparison of the results with those under a formal likelihood-based analysis of noise-multiplied data (Klein et al. 2013) is presented in Section 5 through simulation. It turns out that the inferences obtained using the methodology of this article are comparable with, and just slightly less accurate than, those obtained in Klein et al. (2013). Section 6 presents a disclosure risk evaluation of the proposed method, discusses the benefits of the proposed method in comparison with synthetic data, and outlines how to extend this approach to multivariate data. Section 7 provides some concluding remarks, and the Appendices A, B and C contain proofs of some technical results.

2. Methodology for Fully Noise-Multiplied Data

2.1. General Framework

Suppose $y_1, \ldots, y_n \sim iid \sim f(y|\boldsymbol{\theta})$, independent of $r_1, \ldots, r_n \sim iid \sim h(r)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is an unknown $p \times 1$ parameter vector, and $h(r)$ is a known density (free of $\boldsymbol{\theta}$) such that $h(r) = 0$ if $r < 0$. It is assumed that $f(y|\boldsymbol{\theta})$ and $h(r)$ are the densities of continuous probability distributions. Define $z_i = y_i \times r_i$ for $i = 1, \ldots, n$. Let us write $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{r} = (r_1, \ldots, r_n)$, and $\boldsymbol{z} = (z_1, \ldots, z_n)$.

We note that the joint density of (z_i, r_i) is

$$g(z_i, r_i|\boldsymbol{\theta}) = f\left(\frac{z_i}{r_i}\,\middle|\,\boldsymbol{\theta}\right) h(r_i) r_i^{-1},$$

and the marginal density of z_i is

$$g(z_i|\boldsymbol{\theta}) = \int_0^\infty f\left(\frac{z_i}{\omega}\,\middle|\,\boldsymbol{\theta}\right) h(\omega)\omega^{-1} d\omega. \tag{1}$$

As clearly demonstrated in Klein et al. (2013), standard likelihood-based analysis of the noise-multiplied sample $\boldsymbol{z}$ in order to draw suitable inference about a scalar quantity $Q = Q(\boldsymbol{\theta})$ can be extremely complicated due to the form of $g(z_i|\boldsymbol{\theta})$, and the analysis also must be customized to the noise distribution $h(r)$. Instead, what we propose here is a procedure to *reconstruct* the original data $\boldsymbol{y}$ from reported sample $\boldsymbol{z}$ via *suitable* generation and division by noise terms, and enough replications of the recovered $\boldsymbol{y}$ data by applying multiple imputation method. Once this is accomplished, a data user can apply a simple and standard likelihood procedure to draw inference about $Q(\boldsymbol{\theta})$ based on each reconstructed $\boldsymbol{y}$ data as if it were never perturbed, and finally an application of some known combination rules would complete the task.

The advantages of the suggested approach, blending noise multiplication with multiple imputation, are the following:

1. to protect confidentiality through noise multiplication – satisfying data producer's desire,
2. to allow the data user to analyze the data as if it were never perturbed – satisfying data user's desire (the complexity of the analysis lies in the generation of the imputed values of the noise variables; and the burden of this task will fall on the data producer, not the user), and
3. to allow the data producer to hide information about the underlying noise distribution from data users.

The basic idea behind our procedure is to set it up as a missing data problem; we define the complete, observed, and missing data, respectively, as follows:

$$\boldsymbol{x}_c = \{(z_1, r_1), \ldots, (z_n, r_n)\}, \quad \boldsymbol{x}_{\text{obs}} = \{z_1, \ldots, z_n\}, \quad \boldsymbol{x}_{\text{mis}} = \{r_1, \ldots, r_n\}.$$

Obviously, if the complete data $\boldsymbol{x}_c$ were observed, one would simply recover the original data $y_i = z_i/r_i$, $i = 1, \ldots, n$, and proceed with the analysis in a straightforward manner under the parametric model $f(y|\boldsymbol{\theta})$. Treating the noise variables $r_1, \ldots, r_n$ as missing data, we impute these variables m times to obtain

$$\boldsymbol{x}_c^{*(j)} = \left\{ \left(z_1, r_1^{*(j)}\right), \ldots, \left(z_n, r_n^{*(j)}\right) \right\}, \quad j = 1, \ldots, m. \tag{2}$$

From $\boldsymbol{x}^{*(j)}$ we compute

$$\boldsymbol{y}^{*(j)} = \left\{ y_1^{*(j)}, \ldots, y_n^{*(j)} \right\} = \left\{ \frac{z_1}{r_1^{*(j)}}, \ldots, \frac{z_n}{r_n^{*(j)}} \right\}, \quad j = 1, \ldots, m. \tag{3}$$

The statistical agency would then release the m imputed data sets $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(m)}$, and each data set $\boldsymbol{y}^{*(j)}$ would be analyzed as if it were a random sample from $f(y|\boldsymbol{\theta})$. Thus, suppose that $\eta(\boldsymbol{y})$ is an estimator of $Q(\boldsymbol{\theta})$ based on the unperturbed data $\boldsymbol{y}$ and suppose that $v = v(\boldsymbol{y})$ is an estimator of the variance of $\eta(\boldsymbol{y})$, also computed based on $\boldsymbol{y}$. Often $\eta(\boldsymbol{y})$ will be the maximum likelihood estimator (MLE) of $Q(\boldsymbol{\theta})$, and $v(\boldsymbol{y})$ will be derived from the observed Fisher information matrix. One would then compute $\eta_j = \eta(\boldsymbol{y}^{*(j)})$ and $v_j = v(\boldsymbol{y}^{*(j)})$, the analogs of η and v, obtained from $\boldsymbol{y}^{*(j)}$, and apply a suitable combination rule to pool the information across the m simulations.

At this point two vital pieces of the proposed methodology need to be put together: (1) imputation of $\boldsymbol{r}$ from $\boldsymbol{z}$, which would be the responsibility of the statistical agency; and (2) combination rules for η_j and v_j from several imputations, which the data user would apply in order to analyze the released data. We discuss these two crucial points in Subsections 2.2 and 2.3, respectively.

2.2. *Imputation of the Noise Variables*

In this subsection we describe two procedures that a statistical agency can use to impute $\boldsymbol{r}$ from $\boldsymbol{z}$. Following Wang and Robins (1998), we refer to these two methods as the Type A and Type B imputation procedures.

Type A Imputation Procedure. Under the Type A procedure, the imputed values of $r_1, \ldots, r_n$ are obtained as draws from a posterior predictive distribution. We place a noninformative prior distribution $p(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$. In principle, sampling from the posterior predictive distribution of $r_1, \ldots, r_n$ can be done as follows:

1. Draw $\boldsymbol{\theta}^*$ from the posterior distribution of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$.
2. Draw $r_1^*, \ldots, r_n^*$ from the conditional distribution of $r_1, \ldots, r_n$ given $z_1, \ldots, z_n$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

The above steps are then repeated independently m times to get $\left(r_1^{*(j)}, \ldots, r_n^{*(j)}\right)$, $j = 1, \ldots, m$.

Notice that in step (1) above we use the posterior distribution of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$ as opposed to the posterior distribution of $\boldsymbol{\theta}$ given $y_1, \ldots, y_n$. Such a choice implies that we do not infuse any additional information into the imputes beyond what is provided by the noise-multiplied sample $\boldsymbol{z}$ and the knowledge of the noise-generating distribution $h(r)$. Step (2) above is equivalent to sampling each r_i from the conditional distribution of r_i given z_i and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. The *pdf* of this distribution is

$$h(r_i|z_i, \boldsymbol{\theta}) = \frac{f\left((z_i/r_i)|\boldsymbol{\theta}\right)h(r_i)r_i^{-1}}{\int_0^\infty f\left((z_i/\omega)|\boldsymbol{\theta}\right)h(\omega)\omega^{-1}d\omega}. \tag{4}$$

The sampling required in step (1) can be complicated due to the complex form of the joint density of $z_1, \ldots, z_n$. Certainly, in some cases, the sampling required in step (1) can be performed directly; for instance, if $\boldsymbol{\theta}$ is univariate then we can obtain a direct algorithm by inversion of the cumulative distribution function (numerically or otherwise). More generally, the data augmentation algorithm (Little and Rubin 2002; Tanner and Wong 1987) allows us to bypass the direct sampling from the posterior distribution of $\boldsymbol{\theta}$ given $z_1, \ldots, z_n$. Under the data augmentation method, we proceed as follows. Given a value $\boldsymbol{\theta}^{(t)}$ of $\boldsymbol{\theta}$ drawn at step t:

I. Draw $r_i^{(t+1)} \sim h(r|z_i, \boldsymbol{\theta}^{(t)})$ for $i = 1, \ldots, n$;
II. Draw $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|\boldsymbol{y}^{(t+1)})$ where $\boldsymbol{y}^{(t+1)} = \left(\left(z_1/r_1^{(t+1)}\right), \ldots, \left(z_n/r_n^{(t+1)}\right)\right)$, and $p(\boldsymbol{\theta}|y)$ is the posterior density of $\boldsymbol{\theta}$ given the original unperturbed data $\boldsymbol{y}$ (it is the functional form of $p(\boldsymbol{\theta}|\boldsymbol{y})$ which is relevant here).

The above process is run until t is large and one must, of course, select an initial value $\boldsymbol{\theta}^{(0)}$ to start the iterations. The final generations $\left(r_1^{(t)}, \ldots, r_n^{(t)}\right)$ and $\boldsymbol{\theta}^{(t)}$ form an approximate draw from the joint posterior distribution of $(r_1, \ldots, r_n)$ and $\boldsymbol{\theta}$ given $(z_1, \ldots, z_n)$. Thus, marginally, the final generation $\left(r_1^{(t)}, \ldots, r_n^{(t)}\right)$ is an approximate draw from the posterior predictive distribution of $(r_1, \ldots, r_n)$ given $(z_1, \ldots, z_n)$. This entire iterative process can be repeated independently m times to get the multiply imputed values of the noise variables. The data augmentation algorithm presented here is equivalent to Gibbs sampling. The goal here is to sample from $p(\boldsymbol{r}, \boldsymbol{\theta}|\boldsymbol{z})$, the joint posterior distribution of $(\boldsymbol{r}, \boldsymbol{\theta})$ given $\boldsymbol{z}$. Letting $p(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{\theta})$ denote the conditional density of $\boldsymbol{r}$ given $\boldsymbol{z}$ and $\boldsymbol{\theta}$, and letting $p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{r})$ denote the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{z}$ and $\boldsymbol{r}$, we note that the $(t+1)$th step of a Gibbs sampler would sample from the full conditionals such that $r^{(t+1)} \sim p(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{\theta}^{(t)})$ and $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|\boldsymbol{z}, \mathbf{r}^{(t+1)})$, and would continue until convergence. Alternate sampling from

these two full conditional distributions is equivalent to steps I and II of the data augmentation algorithm.

Sampling from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ in step (II) above will typically be straightforward, either directly or via an embedded Markov chain Monte Carlo step. Under the data augmentation algorithm, we still must sample from the conditional density $h(r|z, \boldsymbol{\theta})$ as defined in (4). The level of complexity here will depend on the form of $f(y|\boldsymbol{\theta})$ and $h(r)$. Usually, sampling from this conditional density will not be too difficult. The following result provides a general rejection algorithm (Devroye 1986; Robert and Casella 2005) to sample from $h(r|z, \boldsymbol{\theta})$ for any continuous $f(y|\boldsymbol{\theta})$, when the noise distribution is Uniform $(1 - \epsilon, 1 + \epsilon)$, that is, when

$$h(r) = \frac{1}{2\epsilon}, \quad 1 - \epsilon \le r \le 1 + \epsilon, \tag{5}$$

where $0 < \epsilon < 1$.

Proposition 1 *Suppose that $f(y|\boldsymbol{\theta})$ is a continuous probability density function, and let us write $f(y|\boldsymbol{\theta}) = c(\boldsymbol{\theta})q(y|\boldsymbol{\theta})$ where $c(\boldsymbol{\theta}) > 0$ is a normalizing constant. Let $M \equiv M(\boldsymbol{\theta}, \epsilon, z)$ be such that*

$$q\left(\frac{z}{r}\middle|\boldsymbol{\theta}\right) \le M \text{ for all } r \in [1 - \epsilon, \gamma]$$

where $\gamma \equiv \gamma(z, \epsilon) > 1 - \epsilon$. Then the following algorithm produces a random variable R having the density

$$h_U(r|z, \boldsymbol{\theta}) = \frac{q((z/r)|\boldsymbol{\theta})r^{-1}}{\int_{1-\epsilon}^{\gamma} q((z/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega}, \quad 1 - \epsilon \le r \le \gamma.$$

I. *Generate U, V as independent Uniform*$(0, 1)$ *and let $W = \gamma^V/(1 - \epsilon)^{V-1}$.*
II. *Accept $R = W$ if $U \le M^{-1}q((z/W)|\boldsymbol{\theta})$, otherwise reject W and return to step (I).*

The expected number of iterations of steps (I) and (II) required to obtain R is

$$\frac{M[\log(\gamma) - \log(1 - \epsilon)]}{\int_{1-\epsilon}^{\gamma} q((z/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega}.$$

The proof of Proposition 1 appears in Appendix A.

Remark 1. The conditional density of y_i given z_i and $\boldsymbol{\theta}$ is

$$f(y_i|z_i, \boldsymbol{\theta}) = \begin{cases} \dfrac{f(y_i|\boldsymbol{\theta})h(z_i/y_i)y_i^{-1}}{\int_0^{\infty} f((z_i/\omega)|\boldsymbol{\theta})h(\omega)\omega^{-1}d\omega}, & \text{if } 0 < z_i < \infty, \quad 0 < y_i < \infty, \\ \dfrac{f(y_i|\boldsymbol{\theta})h(z_i/y_i)(-y_i^{-1})}{\int_0^{\infty} f((z_i/\omega)|\boldsymbol{\theta})h(\omega)\omega^{-1}d\omega}, & \text{if } -\infty < z_i < 0, \quad -\infty < y_i < 0. \end{cases} \tag{6}$$

Drawing r_i^* from the conditional density $h(r_i|z_i, \boldsymbol{\theta}^*)$ defined in (4) and setting $y_i^* = z_i/r_i^*$ is equivalent to drawing y_i^* directly from the conditional density $f(y_i|z_i, \boldsymbol{\theta}^*)$ in the sense that given z_i and $\boldsymbol{\theta}^*$, the variable z_i/r_i^* has the density $f(y_i|z_i, \boldsymbol{\theta}^*)$.

Remark 2. As to the choice of $\boldsymbol{\theta}^{(0)}$, one can choose moment-based estimates (Nayak et al. 2011).

Remark 3. We have tacitly assumed in the above analysis that the posterior distribution of the parameter $\boldsymbol{\theta}$, given noise-multiplied data $\boldsymbol{z}$, is proper. In applications, this needs to be verified on a case by case basis because the posterior propriety under the original data $\boldsymbol{y}$, which may routinely hold under many parametric models, may *not* guarantee the same under $\boldsymbol{z}$ when an improper prior distribution for $\boldsymbol{\theta}$ is used. We refer to the technical report Klein and Sinha (2013) for an example. The same remark holds in the case of the posterior distribution of $\boldsymbol{\theta}$, given the mixture data. We have verified the posterior propriety in our specific applications for fully noise-multiplied data and mixture data in Appendices B and C, respectively.

Type B Imputation Procedure. In this procedure there is no Bayesian model specification. Instead, the unknown parameter $\boldsymbol{\theta}$ is set equal to $\hat{\boldsymbol{\theta}}_{\text{mle}}(\boldsymbol{z})$, the MLE based on the noise-multiplied data $\boldsymbol{z}$, which can often be computed via the EM algorithm (Klein et al. 2013). The imputed values of the noise variables are then randomly drawn such that

$$r_i^* \sim h\big(r|z_i, \hat{\boldsymbol{\theta}}_{\text{mle}}(\boldsymbol{z})\big), \quad \text{for } i = 1, \ldots, n. \tag{7}$$

The above sampling is repeated, independently, m times to obtain $\big(r_1^{*(j)}, \ldots, r_n^{*(j)}\big)$, $j = 1, \ldots, m$. If $h(r)$ is the uniform density (5), then Proposition 1 can be used to implement the sampling in (7).

2.3. Combination Rules for Analyzing the Released Data

We now present methods for analyzing the released data $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(m)}$. Naturally, under the proposed methodology, analysis of the released data would usually be the responsibility of the data user. The analysis involves first analyzing each $\boldsymbol{y}^{*(j)}$ as if it were a random sample from $f(y|\boldsymbol{\theta})$, and then suitably combining the results across $j = 1, \ldots, m$ to obtain final inference. We first present the combination rules of Rubin (1987), which should yield valid inferences when the agency uses the Type A method to impute the noise variables. Rubin's (1987) combination rules often work well, and are simple to apply; however, they may not be optimal, and hence we also consider alternative methods of Wang and Robins (1998).

Rubin's (1987) Rule for Type A Imputation. We assume here that the released data (3) are obtained using the Type A imputation procedure. The multiple imputation estimator of Q is

$$\bar{\eta}_m = \frac{1}{m}\sum_{j=1}^{m} \eta_j, \tag{8}$$

and the estimator of the variance of $\bar{\eta}_m$ is

$$T_m = \left(1 + \frac{1}{m}\right) b_m + \bar{v}_m, \tag{9}$$

where $b_m = (1/(m-1))\sum_{j=1}^{m}(\eta_j - \bar{\eta}_m)^2$ and $\bar{v}_m = (1/m)\sum_{j=1}^{m} v_j$. The point estimator $\bar{\eta}_m$ and its variance estimator T_m can now be used along with a normal cut-off

point to construct a confidence interval for Q. We can also use a t cut-off point based on setting the degrees of freedom equal to $(m-1)\left(1+a_m^{-1}\right)^2$ where $a_m = (1+m^{-1})b_m/\bar{v}_m$.

Wang and Robins's (1998) Rule for Type A Imputation. Once again we assume that the released data (3) are obtained using the Type A imputation procedure. Let

$$\hat{\boldsymbol{\theta}}_j = \arg\max_{\boldsymbol{\theta}}\left\{\prod_{i=1}^{n} f\left(y_i^{*(j)}|\boldsymbol{\theta}\right)\right\}, j = 1, \ldots, m, \tag{10}$$

denote the MLE of $\boldsymbol{\theta}$ computed on the jth imputed data set $\boldsymbol{y}^{*(j)}$ under the model $f(y|\boldsymbol{\theta})$. The multiple imputation estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_A = (1/m)\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j$. By Wang and Robins (1998),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_p[\mathbf{0}, V_A], \text{ as } n \to \infty,$$

where $V_A = I_{\text{obs}}^{-1} + (1/m)I_c^{-1}J + (1/m)J'I_{\text{obs}}^{-1}J$, $J = I_{\text{mis}}I_c^{-1} = (I_c - I_{\text{obs}})I_c^{-1}$, and where I_c and I_{obs} are the $p \times p$ matrices defined by

$$I_c = E\left[-\left(\left(\frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial\,\theta_l \partial\,\theta_{l'}}\right)\right)\right] \text{ and } I_{obs} = E\left[-\left(\left(\frac{\partial^2 \log g(z|\boldsymbol{\theta})}{\partial\,\theta_l \partial\,\theta_{l'}}\right)\right)\right]. \tag{11}$$

Let $S_{ij}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)$ denote the $p \times 1$ score vector, with its lth element defined as

$$S_{ijl}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right) = \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial\,\theta_l}\Big|_{y=y_i^{*(j)}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j}, \; l = 1, \ldots, p, \; i = 1, \ldots, n, \; j = 1, \ldots, m;$$

and let $S_{ij}^*\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)$ denote the $p \times p$ matrix whose (l, l')th element is defined as

$$S_{ijll'}^*\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right) = \frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial\,\theta_l \partial\,\theta_{l'}}\Big|_{y=y_i^{*(j)}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j},$$

$$l, l' = 1, \ldots, p, \; i = 1, \ldots, n, \; j = 1, \ldots, m.$$

A consistent variance estimator $\hat{V}_A$ is obtained by estimating I_c *by*

$$\hat{I}_c = \frac{1}{m}\sum_{j=1}^{m}\hat{I}_{c,j}, \quad \hat{I}_{c,j} = -\frac{1}{n}\sum_{i=1}^{n} S_{ij}^*\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right), \tag{12}$$

and estimating I_{obs} by

$$\hat{I}_{\text{obs}} = \frac{1}{2nm(m-1)}\sum_{i=1}^{n}\sum_{j\neq j'=1}^{m}\left[S_{ij}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)S_{ij'}\left(y_i^{*(j')}, \hat{\boldsymbol{\theta}}_{j'}\right)' + S_{ij'}\left(y_i^{*(j')}, \hat{\boldsymbol{\theta}}_{j'}\right)S_{ij}\left(y_i^{*(j)}, \hat{\boldsymbol{\theta}}_j\right)'\right]. \tag{13}$$

For any given $Q(\boldsymbol{\theta})$, the variance of the multiple imputation estimator $Q(\hat{\boldsymbol{\theta}}_A)$ is obtained by applying the familiar δ-method, and Wald-type inferences can be directly applied to obtain confidence intervals.

Wang and Robins's (1998) Rule for Type B Imputation. We now assume that the released data (3) are obtained using the Type B imputation procedure. Let $\hat{\boldsymbol{\theta}}_j$ be defined

by (10). The multiple imputation estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_B = (1/m)\sum_{j=1}^{m} \hat{\boldsymbol{\theta}}_j$. By Wang and Robins (1998),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_p[0, V_B], \text{ as } n \to \infty,$$

where $V_B = I_{\text{obs}}^{-1} + (1/m)I_c^{-1}J = I_{\text{obs}}^{-1} + (1/m)I_c^{-1}(I_c - I_{\text{obs}})I_c^{-1}$ with I_c and I_{obs} defined in (11). A consistent variance estimator $\hat{V}_B$ is obtained by estimating I_c using (12) and estimating I_{obs} using (13). For any given $Q(\boldsymbol{\theta})$, the variance of the estimator $Q(\hat{\boldsymbol{\theta}}_B)$ is obtained by applying the familiar δ-method, and Wald-type inferences can be directly applied to obtain confidence intervals.

Remark 4. Wang and Robins (1998) provide a comparison between the Type A and Type B imputation procedures, and compare the corresponding variance estimators with Rubin's (1987) variance estimator T_m. Their observation is that the estimators $\hat{V}_A$ and $\hat{V}_B$ are consistent for V_A and V_B, respectively; and the Type B estimator $\hat{\boldsymbol{\theta}}_B$ will generally lead to more accurate inferences than $\hat{\boldsymbol{\theta}}_A$, because for finite m, $V_B < V_A$ (meaning $V_A - V_B$ is positive definite). Under the Type A procedure and for finite m, Rubin's (1987) variance estimator has a nondegenerate limiting distribution; however, the asymptotic mean is V_A, and thus T_m is also an appropriate estimator of variance (in defining Rubin's (1987) variance estimator, Wang and Robins (1998) multiply the quantity b_m by the sample size n to obtain a random variable that is bounded in probability). The variance estimator T_m would appear to underestimate the variance if applied in the Type B procedure because under the Type B procedure, if $m = \infty$, then T_m has a probability limit that is smaller than the asymptotic variance V_B (when $m = \infty$, $V_A = V_B = I_{\text{obs}}^{-1}$). However, under the Type A procedure, if $m = \infty$ then T_m is consistent for the asymptotic variance V_A. We refer to Rubin (1987) and Wang and Robins (1998) for further details.

3. Methodology for Mixture Data

Recall that the term mixture data in our context refers to a data set in which values below C are unperturbed and values above C are perturbed using noise multiplication. In this section we discuss the analysis of such data following the procedure outlined earlier, namely, by (*i*) suitably recovering the y-values above C via use of *reconstructed* noise terms and the noise-multiplied z-values along with or without their identities (below or above C), and (*ii*) providing multiple imputations of such y-values and methods to appropriately combine the original y-values and *reconstructed* y-values to draw inference on Q.

Let $C > 0$ denote the prescribed top code so that y-values above C are sensitive and hence cannot be reported/released. Given $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{r} = (r_1, \ldots, r_n)$, $\boldsymbol{z} = (z_1, \ldots, z_n)$ where $z_i = y_i \times r_i$, we define $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_n)$ with $\Delta_i = I(y_i \le C)$ and $x_i = y_i$ if $y_i \le C$, and $= z_i$ if $y_i > C$. Inference for $\boldsymbol{\theta}$ will be based on either (i) $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ or (ii) just $\{x_1, \ldots, x_n\}$. Under both the scenarios, which each guarantee that the sensitive y-values are protected, several data sets of the type $(y_1^*, \ldots, y_n^*)$ will be released along with a data analysis plan. We describe below the imputation and data analysis plans under both the scenarios.

Case (i). Here we generate r_i^* from the reported values of $(x_i, \Delta_i = 0)$ and compute $y_i^* = x_i / r_i^*$. Of course, if $\Delta_i = 1$ then we set $y_i^* = y_i$. Generation of r_i^* is done by sampling from the conditional distribution $h(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta})$ of r_i, given x_i, $\boldsymbol{\theta}$, and $\Delta_i = 0$, where

$$h(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta}) = \frac{f\left((x_i/r_i)|\boldsymbol{\theta}\right)h(r_i)r_i^{-1}}{\int_0^{(x_i/C)} f\left((x_i/\omega)|\boldsymbol{\theta}\right)h(\omega)\omega^{-1}d\omega}, \quad \text{for } 0 < r_i < \frac{x_i}{C} \tag{14}$$

(Klein et al. 2013) Note that the support of the above conditional distribution is such that $r_i^* \in (0, (x_i/C))$, and thus, if $\Delta_i = 0$, then $y_i^* = \left(x_i/r_i^*\right) > C$. That is, when $y_i > C$, the privacy-protected data point y_i^* has the desirable property that it will also be greater than C. When the noise distribution is the uniform density (5), then (14) can be written as

$$h_U(r_i|x_i, \Delta_i = 0, \boldsymbol{\theta}) = \frac{f((x_i/r_i)|\boldsymbol{\theta})r_i^{-1}}{\int_{1-\epsilon}^{\min\{(x_i/C), 1+\epsilon\}} f((x_i/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega}, \quad \text{for } 1-\epsilon \le r_i \le \min\left\{\frac{x_i}{C}, 1+\epsilon\right\}, \tag{15}$$

and Proposition 1 provides an algorithm for sampling from the above density (15).

Regarding choice of $\boldsymbol{\theta}$, we can proceed following the Type B method (Section 2) and use the MLE of $\boldsymbol{\theta}$ ($\hat{\boldsymbol{\theta}}_{\text{mle}}$) based on the data $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$. This will often be direct (via EM algorithm) in view of the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta)$ reported in Klein et al. (2013) and reproduced below:

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta) = \prod_{i=1}^{n} [f(x_i|\boldsymbol{\theta})]^{\Delta_i} \left[\int_0^{(x_i/C)} f\left(\frac{x_i}{r}|\theta\right)\frac{h(r)}{r}dr\right]^{1-\Delta_i}. \tag{16}$$

Alternatively, following Type A method discussed in Section 2, r^*-values can also be obtained as draws from a posterior predictive distribution. We place a noninformative prior distribution $p(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, and sampling from the posterior predictive distribution of $r_1, \ldots, r_n$ can be done as follows:

1. Draw $\boldsymbol{\theta}^*$ from the posterior distribution of $\boldsymbol{\theta}$ given $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$ using the likelihood $L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta)$ given above.
2. Draw r_i^* for those $i = 1, \ldots, n$ for which $\Delta_i = 0$, from the conditional distribution (14) of r_i, given x_i, $\Delta_i = 0$, and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

As mentioned in Section 2, the sampling required in step (1) above can be complicated due to the complex form of the joint density $L(\boldsymbol{\theta}|\boldsymbol{x}, \Delta)$. The data augmentation algorithm (Little and Rubin 2002; Tanner and Wong 1987) allows us to bypass the direct sampling from the posterior distribution of $\boldsymbol{\theta}$ given $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$.

Under the data augmentation method, given a value $\boldsymbol{\theta}^{(t)}$ of $\boldsymbol{\theta}$ drawn at step t:

I. Draw $r_i^{(t+1)} \sim h(r|x_i, \Delta_i = 0, \boldsymbol{\theta}^{(t)})$ for those $i = 1, \ldots, n$ for which $\Delta_i = 0$.
II. Draw $\boldsymbol{\theta}^{(t+1)} \sim p\left(\boldsymbol{\theta}|y_1^{(t+1)}, \ldots, y_n^{(t+1)}\right)$ where $y_i^{(t+1)} = x_i/r_i^{(t+1)}$ when $\Delta_i = 0$, and $y_i^{(t+1)} = x_i$ otherwise. Here $p(\boldsymbol{\theta}|\boldsymbol{y})$ stands for the posterior *pdf* of $\boldsymbol{\theta}$, given the original data $\boldsymbol{y}$ (only its functional form is used).

The above process is run until t is large and one must, of course, select an initial value $\boldsymbol{\theta}^{(0)}$ to start the iterations.

Case (ii). Here we generate $\left(r_i^{**}, \Delta_i^*\right)$ from the reported values of $(x_1, \ldots, x_n)$ and compute $y_i^{**} = \left(x_i / r_i^{**}\right)$ if $\Delta_i^* = 0$, and $y_i^{**} = x_i$, otherwise, $i = 1, \ldots, n$. This is done by using the conditional distribution $g(r, \delta | x, \boldsymbol{\theta})$ of r and Δ, given x and $\boldsymbol{\theta}$. Since $g(r, \delta | x, \boldsymbol{\theta}) = h(r | x, \delta, \boldsymbol{\theta}) \times \psi(\delta | x, \boldsymbol{\theta})$, and the conditional Bernoulli distribution of Δ, given x and $\boldsymbol{\theta}$, is readily given by

$$\psi(\delta = 1 | x, \boldsymbol{\theta}) = \Pr\{\Delta = 1 | x, \boldsymbol{\theta}\}$$

$$= \frac{f(x|\boldsymbol{\theta}) I(x < C)}{f(x|\boldsymbol{\theta}) I(x < C) + I(x > 0) \int_0^{x/C} f((x/r)|\boldsymbol{\theta}) h(r) r^{-1} dr} \tag{17}$$

(Klein et al. 2013), drawing of $\left(r_i^{**}, \Delta_i^*\right)$, given x_i and $\boldsymbol{\theta}$, is carried out by first randomly selecting Δ_i^* according to the above Bernoulli distribution, and then randomly choosing r_i^{**} if $\Delta_i^* = 0$ from the conditional distribution given by (14).

Again, in the above computations, following Type B approach, one can use the MLE of $\boldsymbol{\theta}$ (via EM algorithm) based on the $\boldsymbol{x}$-data alone whose likelihood is given by

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} \left[f(x_i|\boldsymbol{\theta}) I(x_i < C) + I(x_i > 0) \int_0^{x_i/C} f\left(\frac{x_i}{r} | \boldsymbol{\theta}\right) h(r) r^{-1} dr \right] \tag{18}$$

(Klein et al. 2013). Alternatively, one can proceed as in Type A method (sampling $r_1^{**}, \ldots, r_n^{**}$ from the posterior predictive distribution) by plugging in $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ that are random draws from the posterior distribution of $\boldsymbol{\theta}$, given $\boldsymbol{x}$, based on the above likelihood and choice of prior for $\boldsymbol{\theta}$. As noted in the previous case, here too a direct sampling of $\boldsymbol{\theta}$, given $\boldsymbol{x}$, can be complicated, and we can use the data augmentation algorithm suitably modified following the two steps indicated below.

1. Starting with an initial value of $\boldsymbol{\theta}$ and hence $\boldsymbol{\theta}^{(t)}$ at step t, draw $\left(r_i^{(t+1)}, \Delta_i^{(t+1)}\right)$ $h(r, \delta | x_i, \boldsymbol{\theta}^{(t)})$. This of course is accomplished by first drawing $\Delta_i^{(t+1)}$ and then $r_i^{(t+1)}$, in case $\Delta_i^{(t+1)} = 0$.
2. At step $(t + 1)$, draw $\boldsymbol{\theta}^{(t+1)}$ from the posterior distribution $p\left(\boldsymbol{\theta} | y_1^{(t+1)}, \ldots, y_n^{(t+1)}\right)$ of $\boldsymbol{\theta}$, where $y_i^{(t+1)} = x_i$ if $\Delta_i^{(t+1)} = 1$, and $y_i^{(t+1)} = x_i / r_i^{(t+1)}$ if $\Delta_i^{(t+1)} = 0$. Here, as before, the functional form of the *standard* posterior of $\boldsymbol{\theta}$, given $\boldsymbol{y}$, is used.

In both case (i) and case (ii), after recovering the multiply imputed complete data $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(m)}$ using the techniques described above, methods of parameter estimation, variance estimation, and confidence interval construction are the same as those discussed in Section 2 for fully noise-multiplied data. Naturally, in case (i) when information on the indicator variables Δ is used to generate y^*-values, data users will know exactly which y-values are original and which y-values have been noise-perturbed and de-perturbed. Of course, this need not happen in case (ii), thus providing more privacy protection with perhaps less accuracy. Thus the data producer (such as the Census Bureau) has a choice depending upon to what extent information about the released data should be provided to the data users.

4. Details for Normal and Lognormal Data

In this section we provide some details of the proposed methodology for normal and lognormal populations. Similar details for the exponential population appear in the technical report Klein and Sinha (2013).

4.1. Normal Data

We consider the case of a normal population with uniform noise, that is, we take $f(y|\boldsymbol{\theta}) = (1/(\sigma\sqrt{2\pi}))\exp[-(1/(2\sigma^2))(y-\mu)^2]$, $-\infty < y < \infty$, and we let $h(r)$ be the uniform density (5). We place a standard noninformative improper prior on (μ, σ^2):

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}, \quad -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty. \tag{19}$$

The posterior distribution of (μ, σ^2) given $\boldsymbol{y}$ is obtained as $p(\mu, \sigma^2|\boldsymbol{y}) = p(\mu|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y})$ where

$$(\sigma^2|\boldsymbol{y}) \sim \frac{(n-1)s^2}{\chi^2_{n-1}}, \quad (\mu|\sigma^2, \boldsymbol{y}) \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right), \tag{20}$$

with $\bar{y} = (1/n)\sum_{i=1}^n y_i$ and $s^2 = (1/(n-1))\sum_{i=1}^n (y_i - \bar{y})^2$ (Gelman et al. 2003). The conditional density $h(r|z, \boldsymbol{\theta})$ as defined in (4) now takes the form

$$h(r|z, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))((z/r)-\mu)^2]r^{-1}}{\int_{1-\epsilon}^{1+\epsilon} \exp[-(1/(2\sigma^2))((z/\omega)-\mu)^2]\omega^{-1}d\omega}, \ 1-\epsilon \le r \le 1+\epsilon. \tag{21}$$

We apply Proposition 1 to obtain an algorithm for sampling from this conditional density of r_i given z_i.

Corollary 1 *The following algorithm produces a random variable R whose density is (21).*

I. *Generate U, V as independent Uniform*(0, 1) *and let* $W = (1+\epsilon)^V/(1-\epsilon)^{V-1}$.
II. *Accept* $R = W$ *if* $U \le \exp[(-1/(2\sigma^2))(z/W-\mu)^2]/M$, *otherwise reject W and return to step (I).*

If $z > 0$ *then the constant M is defined as*

$$M \equiv M(\mu, \sigma^2, \epsilon, z) = \begin{cases} \exp\left[-\frac{1}{2\sigma^2}(z/(1+\epsilon)-\mu)^2\right], & \text{if } \mu \le z/(1+\epsilon), \\ 1, & \text{if } z/(1+\epsilon) < \mu < z/(1-\epsilon), \\ \exp\left[-\frac{1}{2\sigma^2}(z/(1-\epsilon)-\mu)^2\right], & \text{if } \mu \ge z/(1-\epsilon). \end{cases}$$

and if $z < 0$ *then*

$$M \equiv M(\mu, \sigma^2, \epsilon, z) = \begin{cases} \exp\left[-\frac{1}{2\sigma^2}(z/(1-\epsilon)-\mu)^2\right], & \text{if } \mu \le z/(1-\epsilon), \\ 1, & \text{if } z/(1-\epsilon) < \mu < z/(1+\epsilon), \\ \exp\left[-\frac{1}{2\sigma^2}(z/(1+\epsilon)-\mu)^2\right], & \text{if } \mu \ge z/(1+\epsilon). \end{cases}$$

The expected number of iterations of steps (I) and (II) required to obtain R is

$$\frac{M[\log(1+\epsilon) - \log(1-\epsilon)]}{\int_{1-\epsilon}^{1+\epsilon} \exp[-(1/(2\sigma^2))((z/\omega)-\mu)^2]\omega^{-1}d\omega}.$$

In the case of mixture data, the conditional density (14) now becomes

$$h(r|x, \Delta = 0, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))(x/r-\mu)^2]r^{-1}}{\int_{1-\epsilon}^{\min\{(x/C),1+\epsilon\}} \exp[-(1/(2\sigma^2))(x/\omega-\mu)^2]\omega^{-1}d\omega}, \quad 1-\epsilon \le r \le \min\left\{\frac{x}{C}, 1+\epsilon\right\}, \tag{22}$$

and a simple modification of Corollary 1 yields an algorithm to sample from this *pdf*.

4.2. *Lognormal Data*

We next consider the case of the lognormal population: $f(y|\boldsymbol{\theta}) = (1/(y\sigma\sqrt{2\pi}))\exp[(-1/(2\sigma^2))(\log y - \mu)^2]$, $0 \le y < \infty$. We define a prior distribution on (μ, σ^2) as in (19). The posterior distribution of (μ, σ^2) is then given by (20) upon replacing each y_i by $\log(y_i)$.

Customized noise distribution for fully perturbed data. Let us take the noise density as:

$$h(r) = \frac{1}{r\xi\sqrt{2\pi}} \exp\left[-\frac{1}{2\xi^2}(\log r + \xi^2/2)^2\right], \quad 0 < r < \infty, \tag{23}$$

where $0 < \xi < \infty$, and $E(R) = 1$ and $\mathrm{Var}(R) = e^{\xi^2} - 1$. We note that $h(r)$ is a lognormal density such that $R \sim h(r) \Leftrightarrow \log(R) \sim N(-\xi^2/2, \xi^2)$. It then follows that $h(r|z, \boldsymbol{\theta})$ is also a lognormal density such that

$$R \sim h(r|z, \boldsymbol{\theta}) \Leftrightarrow \log(R) \sim N\left\{-\frac{\xi^2}{2} + \frac{\xi^2}{\sigma^2+\xi^2}\left[\log(z) + \frac{\xi^2}{2} - \mu\right], \frac{\sigma^2\xi^2}{\sigma^2+\xi^2}\right\}. \tag{24}$$

Uniform noise distribution. Suppose we take the noise distribution to be uniform as defined in (5). Then the conditional *pdf* (4) takes the form

$$h(r|z, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))(\log(z/r) - \mu)^2]}{\int_{1-\epsilon}^{1+\epsilon} \exp\ -(1/(2\sigma^2))(\log(z/\omega) - \mu)^2 d\omega}, \quad 1-\epsilon \le r \le 1+\epsilon \tag{25}$$

We apply Proposition 1 to obtain an algorithm for sampling from this conditional density of r_i given z_i.

Corollary 2 *The following algorithm produces a random variable R whose density is (25).*

I. *Generate U, V as independent Uniform*(0, 1) *and let* $W = (1+\epsilon)^V/(1-\epsilon)^{V-1}$.
II. *Accept* $R = W$ *if* $U \le Wz^{-1}\exp[-(1/(2\sigma^2))(\log(z/W) - \mu)^2]/M$, *otherwise reject W and return to step (I).*

The constant M is defined as

$$M \equiv M(\mu, \sigma^2, \epsilon, z)$$

$$= \begin{cases} (1+\epsilon) z^{-1} \exp\left[-\frac{1}{2\sigma^2} (\log(z/(1+\epsilon)) - \mu)^2 \right], & \text{if } e^{\mu - \sigma^2} \le z/(1+\epsilon), \\ \exp\left[-\mu + \frac{\sigma^2}{2} \right], & \text{if } z/(1+\epsilon) < e^{\mu-\sigma^2} < z/(1-\epsilon), \\ (1-\epsilon) z^{-1} \exp\left[-\frac{1}{2\sigma^2} (\log(z/(1-\epsilon)) - \mu)^2 \right], & \text{if } e^{\mu-\sigma^2} \ge z/(1-\epsilon). \end{cases}$$

The expected number of iterations of steps (I) and (II) required to obtain R is

$$\frac{M[\log(1+\epsilon) - \log(1-\epsilon)]}{\int_{1-\epsilon}^{1+\epsilon} z^{-1} \exp[-(1/(2\sigma^2))(\log(z/\omega) - \mu)^2] d\omega}.$$

In the case of mixture data, the conditional density (14) now becomes

$$h(r|x, \Delta = 0, \boldsymbol{\theta}) = \frac{\exp[-(1/(2\sigma^2))(\log(x/r) - \mu)^2]}{\int_{1-\epsilon}^{\min\{(x/C), 1+\epsilon\}} \exp[-(1/(2\sigma^2))(\log(x/\omega) - \mu)^2] d\omega}, \quad (26)$$

$$1 - \epsilon \le r \le \min\left\{ \frac{x}{C}, 1 + \epsilon \right\},$$

and a simple modification of Corollary 2 yields an algorithm to sample from this *pdf*.

5. Simulation Study to Assess Accuracy of Inference

We use simulation to study the finite sample properties of point estimators, variance estimators, and confidence intervals obtained from noise-multiplied data. We consider the cases of normal and lognormal populations in conjunction with uniform and customized noise distributions as far as possible, as outlined in Section 4. The results for the exponential population are similar to the normal and lognormal, and appear in the technical report Klein and Sinha (2013). One may expect that the simpler method of data analysis proposed in this paper may lead to less accurate inferences than a formal likelihood-based analysis of fully noise-multiplied and mixture data. However, if the inferences derived using the proposed methodology are not substantially less accurate, then the proposed method may be preferable, in some cases, because of its simplicity. Thus the primary goals of this section are essentially to (1) compare the proposed methods with the likelihood-based method reported in Klein et al. (2013), and (2) to assess and compare the finite sample performance of Rubin's (1987) estimation methods with those of Wang and Robins (1998) under our settings of fully noise-multiplied and mixture data.

Each of the tables discussed below is based on a simulation with 5,000 iterations and $m = 5$ imputations of the noise variables generated at each iteration. We choose $m = 5$ because this is a fairly small number of imputations which may be conveniently used in practice. In each of the 5,000 iterations, five independent runs of the data augmentation algorithm, each having 50 iterations, are used to obtain the Type A imputations. Some

exploratory analysis indicated that 50 iterations of the data augmentation algorithm provided an adequate approximation in the chosen simulation settings. All results are obtained using the statistical computing software R (R Development Core Team 2011).

5.1. Fully Noise-Multiplied Data

Table 1 provides results for the case of a normal population when the parameter of interest is either the mean μ or the variance σ^2; and Table 2 provides results for the case of a lognormal population when the parameter of interest is either the mean $e^{\mu+\sigma^2/2}$ or the .95 quantile $e^{\mu+1.645\sigma}$. For each distribution we consider samples sizes $n = 100$ and $n = 500$, but we only display results for the former sample size. Each table displays results for several different methods which are summarized below.

UD: Analysis based on the unperturbed data $\boldsymbol{y}$.

NM10UIB: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type B imputation method and the associated combining rules of Wang and Robins (1998).

NM10UIA1: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type A imputation method and Rubin's (1987) combining rules with the normal cut-off point for confidence interval construction.

NM10UIA2: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type A imputation method and Rubin's (1987) combining rules with the t cut-off point for confidence interval construction.

NM10UIA3: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the Type A imputation method and the associated combining rules of Wang and Robins (1998).

NM10UL: Analysis based on noise-multiplied data with $h(r)$ defined by (5), $\epsilon = .10$, and using the formal likelihood based method of analysis of Klein et al. (2013).

NM10CIB, NM10CIA1, NM10CIA2, NM10CIA3, NM10CL: These methods are defined analogously to the methods above, but $h(r)$ is now the customized noise distribution (23) (for lognormal data); the parameters δ and ξ appearing in $h(r)$ are chosen so that if $R \sim h(r)$, then $\text{Var}(R) = (\epsilon^2)/3$, the variance of the Uniform $(1 - \epsilon, 1 + \epsilon)$ distribution with $\epsilon = 0.10$.

The remaining methods appearing in these tables are similar to the corresponding methods mentioned above after making the appropriate change to the parameter ϵ in the referenced Uniform$(1 - \epsilon, 1 + \epsilon)$ distribution. For each method and each parameter of interest, we display the root mean squared error of the estimator (RMSE), bias of the estimator, standard deviation of the estimator (SD), average over simulation runs of the estimated standard deviation of the estimator ($\widehat{\text{SD}}$), empirical coverage probability of the associated confidence interval (Cvg.), and average length (over simulation iterations) of the corresponding confidence interval relative to the average length of the confidence interval computed from the unperturbed data (Rel. Len.). In each case the nominal coverage probability of the confidence interval is 0.95. For computing an estimate of the standard deviation of an estimator, we simply compute the square root of the appropriate variance estimator. For computing the estimator $\eta(\boldsymbol{y})$ and variance estimator $v(\boldsymbol{y})$ of

Table 1. Inference under fully perturbed N ($\mu = 0$, $\sigma^2 = 1$) data with $n = 100$

	Parameter of interest is the mean μ						Parameter of interest is the variance σ^2					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	99.99	3.24	99.94	99.40	94.66	1.0000	143.89	−6.70	143.73	140.47	92.92	1.0000
NM10UIB	100.11	3.18	100.06	100.99	95.16	1.0160	145.92	−6.02	145.79	148.67	93.06	1.0584
NM10UIA1	100.10	3.12	100.05	99.62	94.80	1.0021	145.87	−6.34	145.74	142.14	92.66	1.0119
NM10UIA2	100.10	3.12	100.05	99.62	94.80	1.0021	145.87	−6.34	145.74	142.14	92.66	1.0119
NM10UIA3	100.10	3.12	100.05	101.24	95.12	1.0185	145.87	−6.34	145.74	149.76	93.36	1.0661
NM10UL	100.10	3.13	100.05	99.58	94.74	1.0018	145.59	−6.32	145.46	141.87	92.62	1.0100
NM20UIB	100.92	3.27	100.87	101.48	95.20	1.0209	150.15	−4.91	150.07	152.80	93.56	1.0878
NM20UIA1	100.83	3.10	100.79	100.26	94.92	1.0086	150.45	−4.17	150.39	146.89	93.02	1.0457
NM20UIA2	100.83	3.10	100.79	100.26	94.92	1.0087	150.45	−4.17	150.39	146.89	93.02	1.0458
NM20UIA3	100.83	3.10	100.79	101.78	95.38	1.0238	150.45	−4.17	150.39	154.09	93.70	1.0969
NM20UL	100.74	3.09	100.69	100.11	94.94	1.0071	149.43	−4.84	149.35	145.74	93.10	1.0375
NM50UIB	103.96	3.39	103.90	104.18	94.80	1.0480	170.21	−4.83	170.15	173.55	93.26	1.2355
NM50UIA1	104.11	3.46	104.06	103.53	94.40	1.0415	171.79	1.78	171.78	169.74	93.12	1.2083
NM50UIA2	104.11	3.46	104.06	103.53	94.52	1.0438	171.79	1.78	171.78	169.74	93.16	1.2109
NM50UIA3	104.11	3.46	104.06	104.79	94.56	1.0541	171.79	1.78	171.78	176.64	93.78	1.2575
NM50UL	103.31	3.29	103.26	102.64	94.52	1.0326	167.38	−4.24	167.32	164.29	93.16	1.1695

Table 2. Inference under fully perturbed LN ($\mu = 0$, $\sigma^2 = 1$) data with $n = 100$

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{SD}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{SD}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	202.26	1.56	202.26	201.82	93.88	1.0000	799.81	−11.83	799.72	793.27	93.16	1.0000
NM10UIB	202.80	1.69	202.79	208.31	94.10	1.0321	802.34	−11.16	802.26	826.38	93.16	1.0417
NM10UIA1	203.18	1.83	203.17	202.46	93.64	1.0032	803.98	−10.57	803.91	796.22	92.88	1.0037
NM10UIA2	203.18	1.83	203.17	202.46	93.64	1.0032	803.98	−10.57	803.91	796.22	92.88	1.0037
NM10UIA3	203.18	1.83	203.17	208.34	94.16	1.0323	803.98	−10.57	803.91	826.50	93.30	1.0419
NM10UL	202.91	1.70	202.90	202.31	93.62	1.0025	802.80	−11.16	802.72	795.55	92.78	1.0029
NM10CIB	202.72	1.48	202.71	208.30	93.92	1.0321	801.97	−12.25	801.88	826.34	93.34	1.0417
NM10CIA1	202.81	1.52	202.80	202.38	93.80	1.0028	802.40	−11.87	802.31	795.89	93.04	1.0033
NM10CIA2	202.81	1.52	202.80	202.38	93.80	1.0028	802.40	−11.87	802.31	795.89	93.04	1.0033
NM10CIA3	202.81	1.52	202.80	208.29	94.02	1.0320	802.40	−11.87	802.31	826.26	93.38	1.0416
NM10CL	202.68	1.41	202.68	202.25	93.84	1.0021	801.56	−12.39	801.47	795.26	93.20	1.0025
NM20UIB	204.60	2.55	204.59	210.24	94.16	1.0417	811.20	−7.89	811.16	835.35	93.26	1.0530
NM20UIA1	204.76	2.21	204.75	204.24	93.84	1.0120	811.69	−9.16	811.63	804.47	93.02	1.0141
NM20UIA2	204.76	2.21	204.75	204.24	93.84	1.0122	811.69	−9.16	811.63	804.47	93.02	1.0144
NM20UIA3	204.76	2.21	204.75	210.16	94.02	1.0413	811.69	−9.16	811.63	834.97	93.34	1.0526
NM20UL	204.33	2.29	204.32	203.83	93.94	1.0099	810.06	−8.76	810.02	802.52	93.34	1.0117
NM20CIB	204.59	2.05	204.58	209.93	94.18	1.0402	810.41	−11.38	810.33	834.00	93.22	1.0513
NM20CIA1	204.41	1.72	204.40	204.04	94.02	1.0110	809.98	−12.28	809.89	803.51	92.98	1.0129
NM20CIA2	204.41	1.72	204.40	204.04	94.04	1.0112	809.98	−12.28	809.89	803.51	93.00	1.0132
NM20CIA3	204.41	1.72	204.40	209.88	94.08	1.0399	809.98	−12.28	809.89	833.77	93.28	1.0511
NM20CL	204.06	1.62	204.05	203.56	93.98	1.0086	808.43	−12.73	808.33	801.31	92.92	1.0101
NM50UIB	217.16	1.62	217.16	221.96	94.06	1.0998	866.70	−16.33	866.55	890.55	93.30	1.1226
NM50UIA1	217.31	2.95	217.29	216.77	93.44	1.0741	867.67	−9.31	867.62	862.13	92.64	1.0868

Table 2. Continued

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
NM50UIA2	217.31	2.95	217.29	216.77	93.56	1.0810	867.67	−9.31	867.62	862.13	92.78	1.0960
NM50UIA3	217.31	2.95	217.29	222.23	93.62	1.1012	867.67	−9.31	867.62	891.63	92.78	1.1240
NM50UL	214.82	0.82	214.81	213.53	93.52	1.0580	855.59	−17.25	855.41	847.91	92.86	1.0689
NM50CIB	214.35	3.42	214.32	220.94	93.96	1.0948	854.98	−7.29	854.95	885.62	93.58	1.1164
NM50CIA1	215.22	4.67	215.17	215.77	93.84	1.0691	857.50	−1.24	857.50	857.56	93.16	1.0810
NM50CIA2	215.22	4.67	215.17	215.77	93.94	1.0749	857.50	−1.24	857.50	857.56	93.32	1.0888
NM50CIA3	215.22	4.67	215.17	221.25	94.02	1.0963	857.50	−1.24	857.50	886.83	93.50	1.1179
NM50CL	212.48	2.53	212.46	212.80	93.96	1.0544	845.95	−9.46	845.90	844.25	93.00	1.0643

Subsection 2.2, we use the maximum likelihood estimator and inverse of observed Fisher information, respectively. All results shown for unperturbed data use Wald-type inferences based on the maximum likelihood estimator and observed Fisher information. The following is a summary of the simulation results of Tables 1–2.

1. In terms of RMSE, bias, and SD of point estimators, as well as average confidence interval length, the proposed methods of analysis are generally only slightly less accurate than the corresponding likelihood-based analysis.
2. In terms of coverage probability of confidence intervals, the multiple imputation-based and formal likelihood-based methods of analysis yield similar results.
3. We consider Uniform$(1 - \epsilon, 1 + \epsilon)$ noise distributions with $\epsilon = 0.1$, 0.2, and 0.5, or equivalent (in terms of variance) customized noise distributions. Generally, for noise distributions with $\epsilon = 0.1$ and 0.2, the proposed analysis based on the noise-multiplied data results only in a slight loss of accuracy in comparison with that based on unperturbed data. When the noise distribution has a larger variance (i.e., when $\epsilon = 0.5$) we notice that the bias of the resulting estimators generally remains small, while the SD clearly increases. When the parameter of interest is the mean, the noise-multiplied data with $\epsilon = 0.5$ still appear to provide inferences with only a slight loss of accuracy compared with the unperturbed data. In contrast, when the parameter of interest is the normal variance as in the right-hand panel of Table 1, the loss of accuracy in terms of SD and hence RMSE appears to be more substantial when ϵ increases to 0.5. We refer to Klein et al. (2013) for a detailed study of the properties of noise-multiplied data.
4. We observe very little difference in the bias, SD, and RMSE of estimators derived under the Type A imputation procedure versus those derived under the Type B imputation procedure.
5. In each table, the column $\widehat{\text{SD}}$ provides the finite sample mean of each of the multiple imputation standard deviation estimators (square root of variance estimators) presented in Section 2. Thus we can compare the finite sample bias of Rubin's (1987) standard deviation estimator of Subsection 2.2 with that of Wang and Robins's (1998) standard deviation estimators of Subsection 2.3 under our setting of noise multiplication. We find that the mean of both of Wang and Robins's (1998) standard deviation estimators is generally larger than the mean of Rubin's (1987) standard deviation estimator. From these numerical results it appears that we cannot make any general statement about which estimators possess the smallest bias, because none of these estimators uniformly dominates the other in terms of minimization of bias. With a larger sample size of $n = 500$ (results not displayed here), we find that all standard deviation estimators have similar expectation; this statement is especially true for the normal case. With the sample size of $n = 100$ we notice in Table 1 that the mean of Rubin's (1987) SD estimator is slightly less than the true SD while both of Wang and Robins's (1998) estimators have a mean slightly larger than the true SD. We should point out that this slight negative bias of Rubin's (1987) SD estimator is most likely due to the fact that the SD estimator based on the original data is itself slightly downward-biased. In the lognormal case, for the sample size $n = 100$ of

Table 2, we notice that Rubin's (1987) estimator is nearly unbiased for the true SD while Wang and Robins's (1998) estimators tend to overestimate the true SD more substantially.

6. When the customized noise distribution is available (e.g., exponential and lognormal cases), the results obtained under the customized noise distribution are quite similar to those obtained under the equivalent (in terms of variance) uniform noise distribution.
7. For confidence interval construction based on Rubin's (1987) variance estimator, the interval based on the normal cut-off point performs very similarly to the interval based on the t cut-off point.
8. The data augmentation algorithm, used by the Type A methods to sample from the posterior predictive distribution of $\boldsymbol{r}$, given the noise-multiplied data, appears to provide an adequate approximation.

5.2. *Mixture Data*

We now study the properties of estimators derived from mixture data as presented in Section 3. Table 3 provides results for the case of a normal population, and Table 4 provides results for the case of a lognormal population. The parameters of interest in each case are the same as in the previous subsection, and the top-coding threshold value C is set equal to the 0.90 quantile of the population. The methods in the rows of Tables 3–4 are as described in the previous subsection, except that each ends with either .i or .ii to indicate either case (i) or case (ii) of Section 3, respectively. The conclusions here are generally in line with those of the previous subsection. Below are some additional findings.

1. Rubin's (1987) SD estimator in this case tends to exhibit very little bias.
2. Generally we find here that the noise multiplication methods yield quite accurate inferences, even more so than in the case of full noise multiplication; this finding is expected since with mixture data only a subset of the original observations are noise-perturbed.
3. As expected, the inferences derived under the case (i) data scenario (observe $(\boldsymbol{x}, \Delta)$) are generally more accurate than those derived under the case (ii) data scenario (observe only $\boldsymbol{x}$), but for the noise distributions considered, the differences in accuracy generally are not too substantial.

6. Further Evaluations and Extensions

6.1. *Disclosure Risk Evaluation*

In this section we report the results of a numerical study designed to give an indication of the amount of disclosure protection provided by the proposed methodology. To be specific, we determine how tightly the m draws $y_i^{*(1)}, \ldots, y_i^{*(m)}$ are centered around the true value y_i, and how well the average and median of these m draws approximate the true value y_i. We consider both the fully noise-multiplied data and mixture data scenarios.

Table 3. Inference for mixture N ($\mu = 0$, $\sigma^2 = 1$) data with C = .90 quantile and n = 100

	Parameter of interest is the mean μ						Parameter of interest is the variance σ^2					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	98.70	−1.21	98.69	99.30	94.50	1.0000	139.88	−9.00	139.59	140.15	93.68	1.0000
NMl0UIB.i	98.81	−1.18	98.81	101.00	94.88	1.0171	140.72	−8.81	140.45	149.18	94.10	1.0645
NMl0UIAl.i	98.79	−1.17	98.78	99.37	94.46	1.0007	140.62	−8.75	140.35	140.88	93.48	1.0053
NM10UIA2.i	98.79	−1.17	98.78	99.37	94.46	1.0007	140.62	−8.75	140.35	140.88	93.48	1.0053
NM10UIA3.i	98.79	−1.17	98.78	101.01	94.82	1.0172	140.62	−8.75	140.35	149.17	94.14	1.0644
NMl0UL.i	98.81	−1.19	98.80	99.36	94.48	1.0005	140.54	−8.87	140.26	140.74	93.56	1.0042
NMl0UIB.ii	98.83	−1.15	98.83	101.01	94.78	1.0172	140.71	−8.67	140.45	149.20	94.16	1.0646
NMl0UIAl.ii	98.81	−1.20	98.81	99.37	94.50	1.0006	140.76	−8.89	140.48	140.87	93.52	1.0052
NM10UIA2.ii	98.81	−1.20	98.81	99.37	94.50	1.0006	140.76	−8.89	140.48	140.87	93.52	1.0052
NM10UIA3.ii	98.81	−1.20	98.81	101.00	94.84	1.0171	140.76	−8.89	140.48	149.21	94.06	1.0646
NMl0UL.ii	98.81	−1.20	98.80	99.36	94.38	1.0006	140.54	−8.88	140.26	140.75	93.54	1.0043
NM20UIB.i	99.23	−1.12	99.22	101.10	94.70	1.0181	142.24	−8.52	141.99	150.74	93.92	1.0756
NM20UIAl.i	99.13	−0.97	99.13	99.55	94.48	1.0025	142.10	−7.89	141.88	142.68	93.64	1.0180
NM20UIA2.i	99.13	−0.97	99.13	99.55	94.48	1.0025	142.10	−7.89	141.88	142.68	93.64	1.0186
NM20UIA3.i	99.13	−0.97	99.13	101.13	94.90	1.0184	142.10	−7.89	141.88	150.71	94.12	1.0753
NM20UL.i	99.09	−1.06	99.09	99.51	94.42	1.0021	141.77	−8.20	141.54	142.24	93.56	1.0149
NM20UIB.ii	99.17	−1.11	99.17	101.13	94.78	1.0184	142.12	−8.37	141.88	150.76	93.90	1.0757
NM20UIAl.ii	99.13	−0.96	99.12	99.58	94.36	1.0028	142.61	−7.76	142.39	142.80	93.40	1.0189
NM20UIA2.ii	99.13	−0.96	99.12	99.58	94.36	1.0028	142.61	−7.76	142.39	142.80	93.44	1.0195
NM20UIA3.ii	99.13	−0.96	99.12	101.16	94.62	1.0187	142.61	−7.76	142.39	150.79	94.02	1.0760
NM20UL.ii	99.10	−1.07	99.09	99.52	94.40	1.0022	141.92	−8.25	141.68	142.31	93.44	1.0154
NM50UIB.i	99.67	−0.59	99.66	101.41	94.56	1.0212	148.43	−6.19	148.30	155.53	94.04	1.1098
NM50UIAl.i	99.77	−0.05	99.77	100.18	94.32	1.0088	149.25	−3.94	149.20	148.33	93.72	1.0584

Table 3. Continued

	Parameter of interest is the mean μ						Parameter of interest is the variance σ^2					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
NM50UIA2.i	99.77	−0.05	99.77	100.18	94.32	1.0089	149.25	−3.94	149.20	148.33	93.78	1.0630
NM50UIA3.i	99.77	−0.05	99.77	101.53	94.64	1.0224	149.25	−3.94	149.20	155.79	94.08	1.1116
NM50UL.i	99.55	−0.57	99.54	99.96	94.32	1.0066	147.32	−6.08	147.19	146.70	93.66	1.0467
NM50UIB.ii	99.99	−0.64	99.99	101.82	94.86	1.0254	150.46	−6.41	150.33	157.79	93.84	1.1259
NM50UIAl.ii	100.07	−0.01	100.07	100.60	94.44	1.0130	150.68	−3.90	150.63	150.30	93.64	1.0724
NM50UIA2.ii	100.07	−0.01	100.07	100.60	94.46	1.0133	150.68	−3.90	150.63	150.30	93.70	1.0791
NM50UIA3.ii	100.07	−0.01	100.07	101.98	94.76	1.0270	150.68	−3.90	150.63	158.04	94.08	1.1277
NM50UL.ii	99.74	−0.72	99.74	100.29	94.48	1.0100	148.93	−6.48	148.79	148.34	93.66	1.0584

Table 4. Inference for mixture LN ($\mu = 0$, $\sigma^2 = 1$) data with $C = .90$ quantile and $n = 100$

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
UD	99.45	2.10	99.43	99.21	94.78	1.0000	781.78	1.65	781.78	794.90	93.48	1.0000
NMl0UIB.i	99.46	2.11	99.44	100.87	95.08	1.0167	783.15	1.95	783.15	824.70	93.62	1.0375
NMl0UIAl.i	99.46	2.11	99.43	99.23	94.78	1.0002	783.32	2.04	783.32	796.04	93.40	1.0014
NM10UIA2.i	99.46	2.11	99.43	99.23	94.78	1.0002	783.32	2.04	783.32	796.04	93.40	1.0014
NM10UIA3.i	99.46	2.11	99.43	100.87	95.10	1.0167	783.32	2.04	783.32	824.66	93.72	1.0374
NMl0UL.i	99.47	2.11	99.45	99.22	94.78	1.0002	783.09	1.97	783.09	795.82	93.38	1.0011
NMl0UIB.ii	99.48	2.11	99.46	100.87	95.06	1.0167	783.92	2.22	783.92	824.75	93.72	1.0376
NMl0UIAl.ii	99.47	2.09	99.45	99.22	94.72	1.0002	783.18	1.64	783.18	796.00	93.36	1.0014
NM10UIA2.ii	99.47	2.09	99.45	99.22	94.72	1.0002	783.18	1.64	783.18	796.00	93.36	1.0014
NM10UIA3.ii	99.47	2.09	99.45	100.86	95.10	1.0167	783.18	1.64	783.18	824.70	93.70	1.0375
NMl0UL.ii	99.47	2.11	99.45	99.23	94.72	1.0002	783.15	1.98	783.14	795.85	93.36	1.0012
NM20UIB.i	99.50	2.10	99.47	100.89	95.12	1.0169	787.17	2.26	787.17	827.71	93.60	1.0413
NM20UIAl.i	99.47	2.10	99.45	99.27	94.82	1.0006	786.76	2.44	786.76	798.97	93.30	1.0051
NM20UIA2.i	99.47	2.10	99.45	99.27	94.82	1.0006	786.76	2.44	786.76	798.97	93.30	1.0052
NM20UIA3.i	99.47	2.10	99.45	100.89	95.04	1.0170	786.76	2.44	786.76	827.52	93.82	1.0410
NM20UL.i	99.49	2.08	99.47	99.26	94.80	1.0005	785.69	1.62	785.69	798.04	93.34	1.0039
NM20UIB.ii	99.50	2.08	99.47	100.90	95.04	1.0170	786.09	1.92	786.09	827.94	93.66	1.0416
NM20UIAl.ii	99.51	2.09	99.49	99.28	94.84	1.0008	787.30	2.51	787.30	799.37	93.44	1.0056
NM20UIA2.ii	99.51	2.09	99.49	99.28	94.84	1.0008	787.30	2.51	787.30	799.37	93.44	1.0057
NM20UIA3.ii	99.51	2.09	99.49	100.91	95.06	1.0171	787.30	2.51	787.30	827.80	93.72	1.0414
NM20UL.ii	99.50	2.07	99.48	99.26	94.76	1.0006	785.97	1.54	785.97	798.27	93.36	1.0042
NM50UIB.i	99.83	2.33	99.80	101.09	95.24	1.0189	804.56	9.96	804.50	842.34	93.76	1.0597
NM50UIAl.i	99.84	2.46	99.81	99.58	94.90	1.0037	803.02	12.96	802.92	816.58	93.50	1.0273

Table 4. Continued

	Parameter of interest is the mean $e^{\mu+\sigma^2/2}$						Parameter of interest is the .95 quantile $e^{\mu+1.645\sigma}$					
	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.	RMSE $\times 10^3$	Bias $\times 10^3$	SD $\times 10^3$	$\widehat{\text{SD}}$ $\times 10^3$	Cvg. %	Rel. Len.
NM50UIA2.i	99.84	2.46	99.81	99.58	94.90	1.0038	803.02	12.96	802.92	816.58	93.50	1.0282
NM50UIA3.i	99.84	2.46	99.81	101.12	95.12	1.0193	803.02	12.96	802.92	842.43	93.96	1.0598
NM50UL.i	99.73	2.32	99.71	99.50	94.86	1.0029	798.51	8.40	798.47	811.47	93.56	1.0208
NM50UIB.ii	100.05	2.42	100.02	101.32	95.18	1.0213	809.84	12.40	809.75	850.03	93.74	1.0694
NM50UIAl.ii	100.07	2.55	100.04	99.78	94.78	1.0058	809.88	14.73	809.75	822.84	93.68	1.0351
NM50UIA2.ii	100.07	2.55	100.04	99.78	94.78	1.0058	809.88	14.73	809.75	822.84	93.70	1.0366
NM50UIA3.ii	100.07	2.55	100.04	101.34	95.12	1.0215	809.88	14.73	809.75	850.50	93.78	1.0699
NM50UL.ii	99.96	2.40	99.93	99.68	94.66	1.0047	803.94	10.09	803.87	817.17	93.54	1.0280

Table 5. Illustration of y, z, and y-values for fully perturbed LN ($\mu = 0$, $\sigma^2 = 1$) data with n = 100 and uniform noise*

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
0.26	0.1	0.26	5	0.25	0.25	0.28	0.27	0.29	0.29	0.26	0.27	0.27	0.27	0.28	0.28
	0.2	0.26		0.23	0.24	0.29	0.27	0.29	0.30	0.22	0.24	0.25	0.27	0.31	0.31
	0.5	0.29		0.20	0.26	0.38	0.34	0.42	0.46	0.28	0.34	0.37	0.39	0.43	0.55
	0.1	0.26	5000	0.24	0.25	0.27	0.27	0.28	0.29	0.24	0.25	0.27	0.27	0.28	0.29
	0.2	0.26		0.22	0.24	0.27	0.27	0.29	0.32	0.22	0.24	0.27	0.27	0.29	0.32
	0.5	0.29		0.19	0.27	0.35	0.36	0.45	0.58	0.19	0.27	0.35	0.36	0.45	0.58
0.96	0.1	1.05	5	1.01	1.09	1.11	1.11	1.17	1.17	0.96	1.02	1.05	1.07	1.14	1.16
	0.2	1.07		0.94	0.95	1.02	1.02	1.07	1.10	0.94	1.11	1.16	1.16	1.29	1.32
	0.5	0.76		0.64	0.67	0.95	0.96	1.10	1.46	0.63	0.64	0.86	0.93	1.22	1.29
	0.1	1.05	5000	0.96	1.00	1.05	1.05	1.10	1.17	0.96	1.00	1.05	1.05	1.10	1.17
	0.2	1.07		0.89	0.98	1.07	1.09	1.19	1.34	0.89	0.98	1.07	1.09	1.19	1.34
	0.5	0.76		0.51	0.65	0.81	0.87	1.05	1.53	0.51	0.63	0.81	0.87	1.07	1.52
2.98	0.1	3.26	5	3.08	3.09	3.14	3.17	3.27	3.29	2.98	3.01	3.10	3.20	3.31	3.63
	0.2	2.56		2.18	2.21	2.62	2.50	2.71	2.78	2.21	2.43	2.55	2.56	2.56	3.03
	0.5	2.84		1.92	1.98	2.13	3.00	4.39	4.58	2.58	3.12	3.83	3.68	4.01	4.87
	0.1	3.26	5000	2.97	3.10	3.25	3.26	3.42	3.63	2.97	3.10	3.25	3.26	3.42	3.62
	0.2	2.56		2.13	2.31	2.53	2.57	2.81	3.20	2.13	2.31	2.53	2.57	2.80	3.20
	0.5	2.84		1.89	2.16	2.58	2.87	3.35	5.68	1.89	2.17	2.57	2.86	3.30	5.68
8.95	0.1	8.13	5	8.00	8.73	8.88	8.70	8.93	8.97	7.43	7.61	8.11	8.03	8.40	8.58
	0.2	9.06		8.24	8.40	9.14	9.17	9.66	10.38	8.24	8.48	9.28	9.09	9.54	9.91
	0.5	7.22		5.04	5.27	5.38	6.55	5.88	11.18	5.01	5.13	5.78	5.66	5.99	6.40
	0.1	8.13	5000	7.39	7.70	8.06	8.11	8.48	9.03	7.39	7.70	8.07	8.11	8.48	9.04
	0.2	9.06		7.55	8.06	8.72	8.95	9.71	11.32	7.55	8.05	8.74	8.96	9.73	11.32
	0.5	7.22		4.82	5.32	6.08	6.75	7.49	14.41	4.81	5.33	6.10	6.75	7.61	14.42
18.21	0.1	19.03	5	17.59	17.62	19.01	18.89	19.14	21.06	17.42	19.13	19.58	19.28	19.66	20.62
	0.2	21.79		19.99	21.68	24.91	23.79	26.19	26.19	18.69	20.59	24.11	23.11	25.33	26.84

Table 5. Continued

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
	0.5	20.42		13.79	14.84	15.93	15.46	16.15	16.58	13.96	14.42	17.64	16.64	18.31	18.88
	0.1	19.03	5000	17.30	17.95	18.75	18.90	19.77	21.14	17.30	17.96	18.77	18.90	19.74	21.14
	0.2	21.79		18.16	19.21	20.62	21.21	22.86	27.23	18.16	19.24	20.69	21.27	22.94	27.24
	0.5	20.42		13.61	14.71	16.37	17.92	19.42	40.62	13.61	14.70	16.25	17.88	19.35	40.81

Table 6. Illustration of y, z, and y-values for fully perturbed LN ($\mu = 0$, $\sigma^2 = 1$) data with n = 100 and customized noise*

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
0.26	0.1	0.28	5	0.25	0.28	0.28	0.28	0.29	0.30	0.28	0.28	0.29	0.29	0.29	0.30
	0.2	0.22		0.19	0.20	0.22	0.22	0.24	0.25	0.18	0.19	0.22	0.22	0.22	0.27
	0.5	0.31		0.24	0.34	0.36	0.35	0.37	0.44	0.36	0.37	0.37	0.38	0.39	0.42
	0.1	0.28	5000	0.23	0.27	0.28	0.28	0.29	0.35	0.23	0.27	0.28	0.28	0.29	0.34
	0.2	0.22		0.15	0.21	0.22	0.23	0.24	0.34	0.14	0.21	0.22	0.23	0.24	0.33
	0.5	0.31		0.14	0.29	0.36	0.37	0.43	0.93	0.16	0.30	0.36	0.37	0.43	0.90
0.96	0.1	1.05	5	0.97	1.03	1.05	1.04	1.05	1.09	0.94	1.00	1.07	1.04	1.11	1.11
	0.2	1.08		0.91	0.95	0.98	1.08	1.27	1.31	0.95	1.01	1.15	1.12	1.21	1.29
	0.5	0.75		0.61	0.79	0.79	0.85	0.93	1.10	0.62	0.68	0.69	0.71	0.76	0.82
	0.1	1.05	5000	0.85	1.01	1.05	1.05	1.09	1.28	0.85	1.01	1.05	1.05	1.09	1.30
	0.2	1.08		0.74	1.01	1.09	1.10	1.18	1.83	0.71	1.01	1.09	1.10	1.18	1.58
	0.5	0.75		0.33	0.68	0.81	0.84	0.98	1.97	0.25	0.68	0.81	0.85	0.98	2.48
2.98	0.1	2.98	5	2.91	2.95	2.99	3.07	3.26	3.26	2.85	2.89	2.94	2.99	3.13	3.14
	0.2	3.08		2.46	2.76	2.93	2.82	2.95	3.02	2.85	2.93	3.30	3.26	3.45	3.79
	0.5	3.10		2.90	2.92	3.59	3.52	4.06	4.13	1.97	2.80	2.81	3.48	3.29	6.52
	0.1	2.98	5000	2.39	2.87	2.99	2.99	3.10	3.58	2.49	2.87	2.98	2.99	3.10	3.69
	0.2	3.08		1.98	2.83	3.06	3.08	3.30	4.48	2.04	2.85	3.08	3.09	3.32	4.56
	0.5	3.10		0.96	2.52	3.02	3.14	3.65	6.91	1.00	2.51	2.99	3.12	3.60	8.03
8.95	0.1	7.75	5	7.41	7.54	7.70	7.72	7.80	8.16	7.21	7.23	7.56	7.54	7.70	7.99
	0.2	7.74		5.86	7.36	7.54	7.46	7.80	8.77	6.38	6.89	7.23	7.75	8.96	9.27
	0.5	8.28		5.20	6.56	8.56	8.27	10.45	10.57	6.05	6.49	6.59	7.12	7.99	8.46
	0.1	7.75	5000	6.22	7.42	7.71	7.72	8.02	9.39	6.20	7.42	7.72	7.73	8.02	9.38
	0.2	7.74		4.66	7.05	7.63	7.68	8.24	12.45	4.80	7.05	7.60	7.66	8.22	11.55
	0.5	8.28		2.71	6.24	7.48	7.75	8.96	17.58	2.47	6.18	7.42	7.71	8.97	21.01
18.21	0.1	18.01	5	17.18	17.26	18.06	18.00	18.59	18.88	16.32	17.16	18.16	17.76	18.53	18.62
	0.2	18.00		14.41	15.40	17.19	16.83	18.17	18.96	18.92	19.12	19.98	20.63	21.22	23.90

Table 6. Continued

				y^*-value											
				Type B method						Type A method					
y-value	ϵ	z-value	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
	0.5	31.32		25.48	29.49	30.15	29.78	31.63	32.16	18.39	24.32	26.99	25.35	28.17	28.88
	0.1	18.01	5000	14.43	17.18	17.86	17.89	18.56	22.38	14.57	17.18	17.87	17.90	18.56	22.77
	0.2	18.00		11.63	16.22	17.53	17.65	18.92	27.71	10.68	16.21	17.46	17.60	18.88	26.66
	0.5	31.32		10.55	21.20	25.48	26.53	30.73	71.57	9.68	21.39	25.58	26.59	30.77	68.18

Case of Fully Noise-Multiplied Data. Tables 5 and 6 report the results of the numerical study for evaluating the disclosure risk in the case of full noise multiplication. In Table 5, $f(y|\boldsymbol{\theta})$ is the lognormal density as in Subsection 4.2 with $\mu = 0$, $\sigma^2 = 1$, and the table shows, for a few selected y_i values, the corresponding z_i values, and a summary of the distribution of the associated values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. The z-values are shown for the cases of the uniform noise density (5) with $\epsilon = 0.1$, 0.2, and 0.5; and the minimum, 1st quartile, median, mean, 3rd quartile, and maximum of the associated values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ are displayed for two cases: $m = 5$ and $m = 5{,}000$. While such a large value as $m = 5{,}000$ may not be used in practice, we consider this large m in order to obtain an accurate picture of the distribution of released values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. Of course for the case $m = 5$, the minimum, 1st quartile, median, 3rd quartile, and maximum are simply the ordered values of $y_i^{*(1)}, \ldots, y_i^{*(5)}$, respectively. Furthermore, results for both the Type A and Type B imputation methods for y^*-values are shown in the table. Table 6 reports similar results for lognormal except that instead of uniform, we use the customized noise distribution for lognormal data as defined in Subsection 4.2, with variances matching those of the Uniform$(1 - \epsilon, 1 + \epsilon)$ density with $\epsilon = 0.1$, 0.2, and 0.5. The following is a summary of the results of Tables 5 and 6.

1. As the variation in the noise distribution $h(r)$ increases (i.e., as ϵ increases), the dispersion in $y_i^{*(1)}, \ldots, y_i^{*(m)}$ also increases. Therefore, as one would expect, the amount of privacy protection provided by this method increases with the variance of the noise-generating distribution.
2. Generally, even for large m, one does not recover the original y_i by averaging or computing the median of the imputed copies $y_i^{*(1)}, \ldots, y_i^{*(m)}$. Usually we find that the noise-multiplied observation z_i is contained between the 1st and 3rd quartiles of $y_i^{*(1)}, \ldots, y_i^{*(m)}$, but interestingly, the y_i value may not be contained between these quartiles. In fact, when ϵ is small, the distribution of the $y_i^{*(1)}, \ldots, y_i^{*(m)}$ values tends to be concentrated around z_i and not y_i. However, when the noise multiplication results in a large perturbation as in the bottom row of Table 6 where $y_i = 18.21$ and $z_i = 31.32$, then we find that the distribution of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ is shifted downward toward y_i, yet still the original value of $y_i = 18.21$ is not contained between the 1st and 3rd quartiles of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. This finding gives some indication that the method does provide some correction of an extreme z_i value, while at the same time does not disclose the original y_i value.
3. Comparing the results of the Type A and Type B imputation procedures, we find them to be quite similar.
4. The results for the uniform and customized noise distributions are similar, although the uniform noise does tend to give a slightly larger interquartile range of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ than the customized noise, thus providing perhaps slightly more privacy protection.

Case of Mixture Data. Table 7 reports the results of the numerical study for evaluating the disclosure risk in the case of mixture data. The population density $f(y|\boldsymbol{\theta})$ is again the lognormal density as in Subsection 4.2 with $\mu = 0$, $\sigma^2 = 1$, the top-coding threshold is $C = 3.60$ which is the 0.90 quantile of the population density (rounded to two decimal places), and the table shows, for three particular y_i values, the corresponding x_i value, and

Table 7. Illustration of y, x, and y-values for mixture LN ($\mu = 0$, $\sigma^2 = 1$) data with C = .90 quantile, n = 100 and uniform noise*

					y^*-value											
					Type B method						Type A method					
y-value	ϵ	x-value	case	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
5.71	0.1	6.01	(i)	5	5.73	5.74	6.03	6.10	6.49	6.52	5.53	5.71	5.83	5.84	5.89	6.23
	0.2	6.38			5.32	5.47	5.55	5.63	5.79	6.02	5.40	5.50	6.09	6.17	6.34	7.52
	0.5	5.09			3.77	4.13	4.15	4.58	4.96	5.90	3.78	3.78	6.02	5.95	6.03	10.11
	0.1	6.01		5000	5.46	5.70	5.97	6.00	6.30	6.68	5.46	5.69	5.97	6.00	6.29	6.68
	0.2	6.38			5.32	5.70	6.23	6.36	6.95	7.98	5.32	5.70	6.19	6.33	6.87	7.98
	0.5	5.09			3.60	4.02	4.63	5.10	5.78	10.12	3.60	4.02	4.68	5.13	5.79	10.16
	0.1	6.01	(ii)	5	5.61	6.02	6.12	6.11	6.38	6.39	5.57	5.71	5.74	5.92	6.04	6.56
	0.2	6.38			5.57	5.81	6.07	6.14	6.42	6.84	5.39	5.50	5.64	6.05	6.32	7.38
	0.5	5.09			3.62	3.99	6.51	5.54	6.60	6.96	4.73	5.15	5.21	5.84	6.85	7.28
	0.1	6.01		5000	5.46	5.70	5.98	6.01	6.30	6.68	5.46	5.70	5.97	6.00	6.29	6.68
	0.2	6.38			5.32	5.70	6.20	6.34	6.88	7.98	5.32	5.70	6.21	6.34	6.91	7.98
	0.5	5.09			3.60	4.04	4.69	5.14	5.82	10.16	3.60	4.01	4.65	5.15	5.87	10.18
3.75	0.1	3.40	(i)	5	3.65	3.68	3.71	3.69	3.71	3.73	3.61	3.61	3.65	3.66	3.70	3.74
	0.2	3.38			3.61	3.80	3.80	3.88	4.02	4.15	3.63	3.63	3.78	3.85	4.00	4.20
	0.5	1.94			3.60	3.61	3.64	3.65	3.65	3.77	3.63	3.66	3.73	3.71	3.73	3.80
	0.1	3.40		5000	3.60	3.64	3.69	3.69	3.73	3.77	3.60	3.64	3.68	3.69	3.73	3.77
	0.2	3.38			3.60	3.73	3.88	3.89	4.04	4.22	3.60	3.73	3.87	3.89	4.04	4.22
	0.5	1.94			3.60	3.66	3.73	3.73	3.80	3.87	3.60	3.67	3.73	3.73	3.80	3.87
	0.1	3.40	(ii)	5	3.40	3.40	3.40	3.46	3.40	3.71	3.40	3.40	3.40	3.47	3.40	3.77
	0.2	3.38			3.38	3.38	3.38	3.49	3.38	3.92	3.38	3.38	3.38	3.38	3.38	3.38
	0.5	1.94			1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94
	0.1	3.40		5000	3.40	3.40	3.40	3.45	3.40	3.77	3.40	3.40	3.40	3.45	3.40	3.77
	0.2	3.38			3.38	3.38	3.38	3.50	3.38	4.22	3.38	3.38	3.38	3.50	3.38	4.22
	0.5	1.94			1.94	1.94	1.94	1.98	1.94	3.87	1.94	1.94	1.94	1.98	1.94	3.87
3.56	0.1	3.56	(i)	5	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.2	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56

Table 7. Continued

					y^*-value											
					Type B method						Type A method					
y-value	ϵ	x-value	case	m	min	q_1	med	mean	q_3	max	min	q_1	med	mean	q_3	max
	0.5	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.1	3.56		5000	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.2	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.5	3.56			3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56	3.56
	0.1	3.56	(ii)	5	3.56	3.56	3.56	3.58	3.56	3.65	3.56	3.56	3.56	3.65	3.72	3.84
	0.2	3.56			3.56	3.56	3.56	3.64	3.56	3.95	3.56	3.56	3.56	3.56	3.56	3.56
	0.5	3.56			3.56	3.56	4.13	4.27	4.94	5.15	3.56	3.56	3.56	4.62	5.40	7.04
	0.1	3.56		5000	3.56	3.56	3.56	3.62	3.65	3.96	3.56	3.56	3.56	3.62	3.65	3.96
	0.2	3.56			3.56	3.56	3.56	3.68	3.69	4.45	3.56	3.56	3.56	3.69	3.70	4.45
	0.5	3.56			3.56	3.56	3.56	3.85	3.62	7.12	3.56	3.56	3.56	3.85	3.60	7.12

distribution of the associated values of $y_i^{*(1)}, \ldots, y_i^{*(m)}$. In this table, the x-values are shown for the cases of the uniform noise density (5) with $\epsilon = 0.1$, 0.2, and 0.5; and the minimum, 1st quartile, median, mean, 3rd quartile, and maximum of $y_i^{*(1)}, \ldots, y_i^{*(m)}$ are displayed for the cases $m = 5$ and $m = 5{,}000$. Results are shown for both cases (i) and (ii) of Section 3 and for both the Type A and Type B imputation methods. Most of the findings here are similar to those of the case of full noise multiplication. Below is a summary of findings from Table 7 which highlights the similarities and differences in privacy protection between cases (i) and (ii) of Section 3.

1. The first part of the table shows results when the y-value is $y_i = 5.71$, which is, of course, greater than the top-coding threshold $C = 3.60$. It happens here that each of the displayed noise-multiplied values is also larger than C. Therefore, based on each of the x-values shown, we know with certainty that $\Delta_i = 0$ (that is, the conditional probability (17) equals 0), and hence the case (ii) method will always impute this particular Δ_i value correctly. Here, the properties of the replications $y_i^{*(1)}, \ldots, y_i^{*(m)}$ for both cases (i) and (ii) are similar to each other and similar to those noted for the full noise multiplication case (replications not centered at y_i, dispersion increasing with ϵ, etc.). Note that the imputations under case (i) may be of slightly higher quality, since the estimate of $\boldsymbol{\theta}$ (either posterior draw or MLE) needed to generate the imputations may be of higher quality when based on case (i) data.
2. The second part of the table shows results when $y_i = 3.75$, which is again greater than $C = 3.60$, but each of the displayed x-values happen to fall in the interval $((1 - \epsilon)C, C)$. When the x-value falls in this interval, the indicator Δ_i cannot be determined from x_i with certainty (that is, the conditional probability (17) does not equal 0 or 1). Therefore, the case (ii) method will sometimes (with a probability governed by (17)), impute Δ_i by the value one, and hence release the noise-multiplied data point as the y^*-value. Here it is interesting to look at the $\epsilon = 0.50$ case where $x_i = 1.94$ because in this case we see a large difference between the results in cases (i) and (ii). In case (i) we use the information that $\Delta_i = 0$ when generating imputations, and hence the released y^*-values are more similar to the original y-value. In case (ii) we do not have this knowledge about the true value of Δ_i. Since the noise-multiplied observation is fairly small, Δ_i is often imputed as 1 in case (ii). Therefore, under case (ii), the noise-multiplied data point is often directly released in the replications $y_i^{*(1)}, \ldots, y_i^{*(m)}$ and a user who sees these data would not immediately know if the value repeated several times in the released $y_i^{*(1)}, \ldots, y_i^{*(m)}$ was the original y_i or its noise perturbed version.
3. The third part of the table shows results with $y_i = 3.56$. In this case, the y-value is less than the top-coding threshold $C = 3.60$, while each of the x-values happen to fall in the interval $((1 - \epsilon)C, C)$. Therefore, the value of Δ_i cannot be determined with certainty from x_i (the conditional probability (17) does not equal 0 or 1). Thus, the case (ii) method sometimes imputes Δ_i by 0, and in these cases the released y^* will not be equal to the original y-value, since it will be divided by a random draw from (14). In this situation, unlike the situation described in item (2) directly above, the value repeated several times in the replications $y_i^{*(1)}, \ldots, y_i^{*(m)}$ for case (ii) is the original observation, not its noise-perturbed version. In this case, the case

(i) method, which uses knowledge of $\Delta_i = 1$, always sets the released $y*$-value to the true value of y.

6.2. Comparison with Synthetic Data

The methodology developed in this article is designed to enable statistical agencies to release privacy-protected data that can be readily analyzed by data users. The methods of (partially) synthetic data developed in Reiter (2003) are designed for the same purpose, and hence a comparison of our methodology with that of Reiter (2003) is in order. A general criticism of noise multiplication is that a proper statistical analysis of noise-multiplied data is complicated for data users. The results of this article show how to remedy this criticism by making the analysis as simple (for the data user) as the analysis of synthetic data (we showed that Rubin's (1987) combining rules can be used here, and these rules are only slightly different from those of Reiter (2003)). Since the methodology of this article gives very similar results to the full likelihood-based analysis of noise-multiplied data developed in Klein et al. (2013), we believe that the pertinent comparison is that of synthetic data versus noise multiplication, assuming a valid data analysis is performed in both cases. Such a comparison, in terms of data quality, is precisely the topic of Klein et al. (2013). We note that synthetic data certainly has benefits, as it has been thoroughly studied in recent years, and successfully applied to complex multivariate data sets. At the same time, the methodology of this article can be extended to multivariate data as outlined in the subsection below. An advantage of noise multiplication over synthetic data is that noise multiplication allows the statistical agency to precisely control the quality of the released data, and also the level of privacy protection, through the choice of $h(r)$. For instance, when $h(r)$ is the uniform density (5), the extensive numerical results of Klein et al. (2013) show, for some univariate parametric models, precisely how to select ϵ so that the quality of inferences are equivalent to, less than, or greater than, the quality of inferences derived under synthetic data. Indeed, the ability to choose $h(r)$ provides the statistical agency with a very fine level of control over the data quality and privacy protection, and such an explicit tuning mechanism is not present in standard synthetic data methodology. Further privacy guarantees under noise multiplication can be made, for instance, by taking $h(r)$ to be a density such as

$$h(r) = \frac{1}{2(\epsilon - \xi)}, \text{ if } r \in (1 - \epsilon, 1 - \xi) \cup (1 + \xi, 1 + \epsilon), \tag{27}$$

where $0 < \xi < \epsilon < 1$. Notice that the noise density (27) implies that the noise multiplier r is always a distance ξ away from 1, and hence we are guaranteed that the relative distance between the original observation y and noise-multiplied observation z is $|(z - y)/y| > \xi$.

6.3. Extensions for Multivariate Data

So far in this article we assumed that the original data, $y_1, \ldots, y_n$, consist of a set of n independent random variables whose support is a subset $\mathbb{R}$. In this section, we outline an extension of our methodology to the case of multivariate and fully noise-multiplied data. In the multivariate case, we assume that the original data consist of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, a set of n independent $k \times 1$ dimensional random vectors. Thus we suppose that

$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim iid \sim f(\boldsymbol{y}|\boldsymbol{\theta})$, independent of $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n \sim iid \sim h(\boldsymbol{r})$ where $f(\,y|\boldsymbol{\theta})$ and $h(r)$ are densities of continuous probability distributions whose support is a subset of $\mathbb{R}^k$. As before, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is an unknown $p \times 1$ parameter vector, and now $h(\boldsymbol{r})$ is a known density such that $h(\boldsymbol{r}) = 0$ if any component of the vector r is less than zero. Writing $\boldsymbol{y}_i = (\,y_{i1}, \ldots, y_{ik})$ and $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{ik})$, the fully noise-multiplied data are now defined by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik}) = (\,y_{i1}r_{i1}, , \ldots, y_{ik}r_{ik})$, $i = 1, , \ldots, n$.

The joint density of $(\boldsymbol{z}_i, \boldsymbol{r}_i)$ is

$$g(\boldsymbol{z}_i, \boldsymbol{r}_i|\boldsymbol{\theta}) = f\left(\frac{z_{i1}}{r_{i1}}, \ldots, \frac{z_{ik}}{r_{ik}}\middle|\boldsymbol{\theta}\right) h(r_{i1}, \ldots, r_{ik}) \left[\prod_{l=1}^{k} r_{il}^{-1}\right],$$

the marginal density of $\boldsymbol{z}_i$ is

$$g(\boldsymbol{z}_i|\theta) = \int_0^\infty \ldots \int_0^\infty f\left(\frac{z_{i1}}{\omega_{i1}}, \ldots, \frac{z_{ik}}{\omega_{ik}}\middle|\boldsymbol{\theta}\right) h(\omega_{i1}, \ldots, \omega_{ik}) \left[\prod_{l=1}^{k} \omega_{il}^{-1}\right] d\omega_{i1} \ldots d\omega_{ik},$$

and hence the conditional density of $\boldsymbol{r}_i$ given $\boldsymbol{z}_i$ is

$$h(\boldsymbol{r}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) = \frac{f((z_{i1}/r_{i1}), \ldots, (z_{ik}/r_{ik})|\boldsymbol{\theta}) h(r_{i1}, \ldots, r_{ik}) \left[\prod_{l=1}^{k} r_{il}^{-1}\right]}{\int_0^\infty \ldots \int_0^\infty f((z_{i1}/\omega_{i1}), \ldots, (z_{ik}/\omega_{ik})|\boldsymbol{\theta}) h(\omega_{i1}, \ldots, \omega_{ik}) \left[\prod_{l=1}^{k} \omega_{il}^{-1}\right] d\omega_{i1} \ldots d\omega_{ik}}. \tag{28}$$

The complete, observed, and missing data are defined, respectively, as

$$\boldsymbol{x}_c = \{(\boldsymbol{z}_1, \boldsymbol{r}_1), \ldots, (\boldsymbol{z}_n, \boldsymbol{r}_n)\}, \quad \boldsymbol{x}_{\text{obs}} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}, \quad \boldsymbol{x}_{\text{mis}} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}.$$

The noise vectors $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ are imputed m times to obtain

$$\begin{aligned} \boldsymbol{x}_c^{*(j)} &= \left\{\left(\boldsymbol{z}_1, \boldsymbol{r}_1^{*(j)}\right), \ldots, \left(\boldsymbol{z}_n, \boldsymbol{r}_n^{*(j)}\right)\right\} \\ &= \left\{(z_{11}, \ldots, z_{1k}), \left(r_{11}^{*(j)}, \ldots, r_{1k}^{*(j)}\right), \ldots, (z_{n1}, \ldots, z_{nk}), \left(r_{n1}^{*(j)}, \ldots, r_{nk}^{*(j)}\right)\right\}, j = 1, \ldots, m, \end{aligned}$$

and the privacy-protected data are obtained as

$$\begin{aligned} \boldsymbol{y}^{*(j)} &= \left\{\boldsymbol{y}_1^{*(j)}, \ldots, \boldsymbol{y}_n^{*(j)}\right\} = \left\{\left(y_{11}^{*(j)}, \ldots, y_{1k}^{*(j)}\right), \ldots, \left(y_{n1}^{*(j)}, \ldots, y_{nk}^{*(j)}\right)\right\} \\ &= \left\{\left(\frac{z_{11}}{r_{11}^{*(j)}}, \ldots, \frac{z_{1k}}{r_{1k}^{*(j)}}\right), \ldots, \left(\frac{z_{n1}}{r_{n1}^{*(j)}}, \ldots, \frac{z_{nk}}{r_{nk}^{*(j)}}\right)\right\}, j = 1, \ldots, k. \end{aligned} \tag{29}$$

The methods of Subsection 2.2 can be used to impute the noise vectors, and the methods of Subsection 2.3 can be used to analyze the privacy-protected data given in (29). Conceptually, the methods of Subsections 2.2 and 2.3 can be readily applied to multivariate data. For instance, a data user wishing to draw inference about the correlation between y_{i1} and y_{i2} would set $Q(\boldsymbol{\theta}) = \text{Corr}(\,y_{i1}, y_{i2}|\boldsymbol{\theta})$, and apply methods of Subsection 2.3. For the statistical agency generating the imputations, there is perhaps one extension needed in applying the methods of Subsection 2.2, because when generating the imputations (either Type A or Type B), instead of sampling from the univariate

conditional density (4), we must now sample from the k-dimensional multivariate conditional density (28). In the univariate case we used Proposition 1 to extract samples from (4) when one takes the noise-generating density to be (5). In the multivariate case, a generalization of Proposition 1 can be used to sample random vectors from (28), when the noise-generating distribution is the following k-dimensional uniform density (which is a straightforward generalization of (5)):

$$h(r_1,\ldots,r_k)=\frac{1}{2^k\prod_{l=1}^{k}\epsilon_l},\ \text{for}\ (r_1,\ldots,r_k)\in[1-\epsilon_1,1+\epsilon_1]\times\cdots\times[1-\epsilon_k,1+\epsilon_k], \quad (30)$$

where $0<\epsilon_1,\ldots,\epsilon_k<1$. The generalization of Proposition 1 is stated below as Proposition 2; the proof is similar to that of Proposition 1 and hence is omitted.

Proposition 2 *Suppose that $f(\mathbf{y}|\boldsymbol{\theta})$ is a continuous probability density function of a k-dimensional distribution, and let us write $f(\mathbf{y}|\boldsymbol{\theta})=c(\boldsymbol{\theta})q(\mathbf{y}|\boldsymbol{\theta})$ where $c(\boldsymbol{\theta})>0$ is a normalizing constant. Let $M\equiv M(\boldsymbol{\theta},\epsilon_1,\ldots,\epsilon_k,\mathbf{z})$ be such that*

$$q\left(\frac{z_1}{r_1},\ldots,\frac{z_k}{r_k}\middle|\boldsymbol{\theta}\right)\le M\ \text{for all}\ (r_1,\ldots,r_k)$$

$$\in[1-\epsilon_1,1+\epsilon_1]\times\cdots\times[1-\epsilon_k,1+\epsilon_k].$$

Then the following algorithm produces a random vector $(R_1,\ldots,R_k)$ having the density

$$h_U(r_1,\ldots,r_k|z_1,\ldots,z_k,\boldsymbol{\theta})$$

$$=\frac{q((z_1/r_1),\ldots,(z_k/r_k)|\boldsymbol{\theta})\left[\prod_{l=1}^{k}r_l^{-1}\right]}{\int_{1-\epsilon_k}^{1+\epsilon_k}\cdots\int_{1+\epsilon_1}^{1+\epsilon_1}q((z_1/\omega_1),\ldots,(z_k/\omega_k)|\boldsymbol{\theta})\left[\prod_{l=1}^{k}\omega_l^{-1}\right]d\omega_1\ldots d\omega_k},$$

$$\text{for}\ (r_1,\ldots,r_k)\in[1-\epsilon_1,1+\epsilon_1]\times\cdots\times[1-\epsilon_k,1+\epsilon_k].$$

I. *Generate $U, V_1,\ldots,V_k$ as independent Uniform $(0,\ 1)$ and let $W_l=(1+\epsilon_l)^{V_l}/(1-\epsilon_l)^{V_l-1}$ for $l=1,\ldots,k$.*
II. *Accept $(R_1,\ldots,R_k)=(W_1,\ldots,W_k)$ if $U\le M^{-1}q((z_1/W_1),\ldots,(z_k/W_k)|\boldsymbol{\theta})$, otherwise reject the vector $(W_1,\ldots,W_k)$ and return to step (I).*

The expected number of iterations of steps (I) and (II) required to obtain $(R_1,\ldots,R_k)$ is

$$\frac{M\prod_{l=1}^{k}\log\left[\dfrac{1+\epsilon_l}{1-\epsilon_l}\right]}{\int_{1-\epsilon_k}^{1+\epsilon_k}\cdots\int_{1-\epsilon_1}^{1+\epsilon_1}q((z_1/\omega_1),\ldots,(z_k/\omega_k)|\boldsymbol{\theta})\left[\prod_{l=1}^{k}\omega_l^{-1}\right]d\omega_1\ldots d\omega_k}.$$

Remark 5. In this section we briefly outlined the multivariate extension for the case of fully noise-multiplied data; that is, where $\mathbf{y}_1,\ldots,\mathbf{y}_n\sim iid\sim\mathbf{Y}$ and each component of $\mathbf{Y}$ requires protection from disclosure. We note that the methodology outlined in this section allows one to use different levels of privacy protection for each component of $\mathbf{Y}$ through the choice of $\epsilon_1,\ldots,\epsilon_k$ in (30). Other scenarios are certainly possible; for instance, it

may be that $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ where the variable Y_1 must always be protected, Y_2 requires protection only if it exceeds a fixed threshold $C > 0$, and Y_3 does not require any protection. We intend to address such issues in a future communication.

7. Concluding Remarks

There are perhaps two rigorous ways of producing privacy-protected data: multiple imputation and noise perturbation. Klein et al. (2013) show that the likelihood-based method of analysis of noise-multiplied data can yield accurate inferences under several standard parametric models and compare favorably with the standard multiple imputation-based analysis methods of Reiter (2003) and An and Little (2007). Since the likelihood of the noise-multiplied data is often complex, one wonders if an alternative simpler and fairly accurate data analysis method can be developed based on such kind of privacy-protected data. With precisely this objective in mind, we have shown in this article that a proper application of multiple imputation leads to such an analysis. In implementing the proposed method under a standard parametric model $f(y|\boldsymbol{\theta})$, the most complex part is generally simulation from the conditional densities (4) or (14), and this part would be the responsibility of the data producer, not the data user. We have provided Proposition 1 which gives an exact algorithm to sample from (4) and (14) for general continuous $f(y|\boldsymbol{\theta})$, when $h(r)$ is the uniform distribution (5). Moreover, we have seen that in the lognormal case under full noise multiplication, if one uses the customized noise distribution, then the conditional density (4) takes a standard form from which sampling is straightforward. Simulation results based on sample sizes of 100 and 500 indicate that the multiple imputation-based analysis, as developed in this article, generally results in only a slight loss of accuracy in comparison to the formal likelihood-based analysis. Our simulation results also indicate that both the Rubin (1987) and Wang and Robins (1998) combining rules exhibit adequate performance in the selected sample settings. We have also reported some additional numerical results for evaluating the amount of privacy protection offered by the method. These results showed that one does not recover the original observation simply by averaging the multiply imputed copies of the original value.

In conclusion, we observe that, from a data user's perspective, our method does require a knowledge of the underlying parametric model of the original y-data so that efficient model-based estimates can be used to analyze the reconstructed y^*-values. In this article we assumed that the model used by the agency to multiply impute the original data, namely $f(y|\boldsymbol{\theta})$, is the same model adopted by the data user to analyze the released data. However, in practice this may not be the case (see Meng 1994 and Robins and Wang 2000 for a discussion of possible consequences of model misspecification). In any event, modeling by data users, if necessary, will be based on the released y^*-values, and *not* on the noise-multiplied z-values. It is expected that the sampling behaviors of y-values and y^*-values would be similar. This is in the same spirit as in the case of synthetic data usage where a data user will either be informed about the original model or try to build up a reasonable model based on the released synthetic data. We should also point out that in practice, most data sets have a complex multivariate structure. We briefly outlined how our methodology can be extended to multivariate data. In a future communication we intend to investigate the robustness of the multiple imputation-based analysis to

discrepancies between the imputation and analysis models, and to further develop the multivariate extensions of the proposed method.

Appendix A

Proof of Proposition 1. This is a rejection sampling algorithm where the target density $h_U(r|z, \boldsymbol{\theta})$ is proportional to $s_{\text{target}}(r) = q((z/r)|\boldsymbol{\theta})r^{-1}$, $1 - \epsilon \le r \le \gamma$, and the instrumental density is $s_{\text{instr}}(r) = r^{-1}/(\log(\gamma) - \log(1-\epsilon))$, $1 - \epsilon \le r \le \gamma$. To fill in the details, first note that since $f(y|\boldsymbol{\theta})$ is continuous in y, it follows that $q((z/r)|\boldsymbol{\theta})$ is continuous as a function of r, on the interval $[1 - \epsilon, \gamma]$, and thus the bounding constant M exists. Then we see that

$$\frac{s_{\text{target}}(r)}{s_{\text{instr}}(r)} = [\log(\gamma) - \log(1-\epsilon)]q\left(\frac{z}{r}|\boldsymbol{\theta}\right) \le [\log(\gamma) - \log(1-\epsilon)]M, \tag{31}$$

for all $r \in [1 - \epsilon, \gamma]$. Note that the cumulative distribution function corresponding to $s_{\text{instr}}(r)$ is $\mathbb{S}_{\text{instr}}(r) = (\log(r) - \log(1-\epsilon))/(\log(\gamma) - \log(1-\epsilon))$, $1 - \epsilon \le r \le \gamma$, and the inverse of this distribution function is $\mathbb{S}^{-1}_{\text{instr}}(u) = \gamma^u/(1-\epsilon)^{u-1}$, $0 \le u \le 1$. Thus, by the inversion method (Devroye 1986), step (I) is equivalent to independently drawing $U \sim \text{Uniform}(0,1)$ and W from the density $s_{\text{instr}}(r)$. Since $M^{-1}s_{\text{target}}(W)/([\log(\gamma) - \log(1-\epsilon)]s_{\text{instr}}(W)) = q(z/w|\boldsymbol{\theta})/M$, step (II) is equivalent to accepting W *if* $U \le M^{-1}s_{\text{target}}(W)/([\log(\gamma) - \log(1-\epsilon)]s_{\text{instr}}(W))$, which is the usual rejection step based on the bound in (31). Finally, we use the well-known fact that the expected number of iterations of the rejection algorithm is equal to the bounding constant in (31) times the normalizing constant for $s_{\text{target}}(r)$, i.e., $[\log(\gamma) - \log(1-\epsilon)]M/[\int_{1-\epsilon}^{\gamma} q((z/\omega)|\boldsymbol{\theta})\omega^{-1}d\omega]$.

Appendix B

Here we provide proofs of the posterior propriety of $\boldsymbol{\theta}$, given the fully noise-multiplied data $\boldsymbol{z}$, for normal and lognormal distributions.

Normal distribution. Here $g(z|\boldsymbol{\theta}) \propto (1/\sigma)\int \exp[-((z/r) - \mu)^2/(2\sigma^2)](h(r)/r)dr$. Writing down the joint *pdf* of $z_1, \ldots, z_n$, it is obvious that upon integrating out μ with respect to (wrt) the Lebesgue measure and σ wrt the flat or noninformative prior, we end up with the expression $U(\boldsymbol{z})$ given by

$$U(\boldsymbol{z}) = \int \cdots \int \left[\sum_{i=1}^{n} \frac{z_i^2}{r_i^2} - \frac{\left(\sum_{i=1}^{n}(z_i/r_i)\right)^2}{n}\right]^{-n-\delta} \frac{h(r_1)\ldots h(r_n)}{r_1 \ldots r_n} dr_1 \ldots dr_n$$

where $\delta \ge 0$. To prove that $U(\boldsymbol{z})$ is finite for any given $\boldsymbol{z}$, note that

$$\left[\sum_{i=1}^{n} \frac{z_i^2}{r_i^2} - \frac{\sum_{i=1}^{n}(z_i/r_i)^2}{n}\right] = \frac{1}{2}\sum_{i,\,j=1}^{n}\left(\frac{z_i}{r_i} - \frac{z_j}{r_j}\right)^2 \ge \frac{1}{2}\left[\frac{z_1}{r_1} - \frac{z_2}{r_2}\right]^2$$

for any pair $(z_1, z_2; r_1, r_2)$, Assume without any loss of generality that $z_1 > z_2$, and note that

$[(z_1/r_1) - (z_2/r_2)]^2 = [(z_1/z_2) - (r_1/r_2)]^2 \times z_2^2 r_1^{-2}$. Then under the *condition*

$$\int_r \frac{h(r)}{r} dr = K_1 < \infty, \quad \int_{r_1 \le r_2} r_1^{2(n+\delta)-1} r_2^{-1} h(r_1) h(r_2) dr_1 dr_2 = K_2 < \infty, \tag{32}$$

$U(z)$ is bounded above by

$$U(z) \le 2^{n+\delta} K_1^{n-2} \left[\frac{z_1}{z_2} - 1\right]^{-2(n+\delta)} \left[\int_{r_1 \le r_2} r_1^{2(n+\delta)-1} r_2^{-1} h(r_1) h(r_2) dr_1 dr_2\right] < \infty.$$

In particular, when $R \sim$ Uniform$(1 - \epsilon, 1 + \epsilon)$, the above condition is trivially satisfied!

Lognormal distribution. Here $g(z|\boldsymbol{\theta}) \propto (1/z\sigma) \int \exp[-(\log(z/r) - \mu)^2/(2\sigma^2)] h(r) dr$. Writing down the joint density of $z_1, \ldots, z_n$, and putting $u = \log(z/r)$, it is obvious that upon integrating out μ wrt the Lebesgue measure and σ wrt the flat or noninformative prior, we end up with the expression $U(z)$ given by

$$U(z) = \int_{r_1} \cdots \int_{r_n} \left[\sum_{i=1}^n (u_i - \bar{u})^2\right]^{-2(n+\delta)} h(r_1) \cdots h(r_n) dr_1 \cdots dr_n$$

where $\delta \ge 0$. To prove that $U(z)$ is finite for any given z, note as in the normal case that when $z_1 > z_2$ (without any loss of generality),

$$\left[\sum_{i=1}^n (u_i - \bar{u})^2\right] = \frac{1}{2} \sum_{i,j=1}^n (u_i - u_j)^2 \ge \frac{1}{2}(u_1 - u_2)^2 = \frac{1}{2}\left[\log\left(\frac{z_1}{z_2}\right) - \log\left(\frac{r_1}{r_2}\right)\right]^2$$

$$\ge \frac{1}{2}\left[\log\left(\frac{z_1}{z_2}\right)\right]^2$$

for $r_1 < r_2$. Hence $U(z)$ is always finite, since $\int_{r_1 < r_2} h(r_1) h(r_2) dr_1 dr_2 < \infty$.

Appendix C

Here we provide proofs of the posterior propriety of $\boldsymbol{\theta}$, given the mixture data, for normal and lognormal distributions. We consider two cases depending on the nature of mixture data that will be released.

Case (i): Nature of data $\{(x_1, \Delta_1), \cdots, (x_n, \Delta_n)\}$.

Normal distribution. Given the data $\{(x_1, \Delta_1), \ldots, (x_n, \Delta_n)\}$, let $I_1 = \{i : \Delta_i = 1\}$ and $I_0 = \{i : \Delta_i = 0\}$. Then the normal likelihood $L(\boldsymbol{\theta}|\text{data})$, apart from a constant, can be expressed as

$$L(\boldsymbol{\theta}|\text{data}) \propto \sigma^{-n} \left[\exp\left(-\sum_{i \in I_1} \frac{(x_i - \mu)^2}{2\sigma^2}\right)\right]$$

$$\times \left[\prod_{i \in I_0} \int_0^{(x_i/c)} \exp\left(-\frac{((x_i/r_i) - \mu)^2}{2\sigma^2}\right) \frac{h(r_i)}{r_i} I(x_i > 0) dr_i\right].$$

It is then obvious that upon integrating out μ wrt the Lebesgue measure and σ wrt the flat or noninformative prior, we end up with the expression U (data) given by

$$U(\text{data}) = \prod_{i\in I_0}\int_0^{(x_i/c)} I(x_i > 0)\left[\sum_{i\in I_1} x_i^2 + \sum_{i\in I_0}\frac{x_i^2}{r_i^2} - \frac{\left(\sum_{i\in I_1} x_i + \sum_{i\in I_0}(x_i/r_i)\right)^2}{n}\right]^{-n-\delta} \frac{h(r_i)}{r_i}dr_i.$$

Writing $v_i = x_i/r_i$ for $i \in I_0$, the expression $\Psi(\text{data}) = \sum_{i\in I_1} x_i^2 + \sum_{i\in I_0} x_i^2/r_i^2 - \left(\sum_{i\in I_1} x_i + \sum_{i\in I_0}(x_i/r_i)\right)^2/n$ is readily simplified as $[S_1^2 + S_0^2 + rs(\bar{x}_1 - \bar{x}_0)^2](r+s)^{-1}$ where r and s are the cardinalities of I_1 and I_0, respectively, and $\left(\bar{x}_1, S_1^2\right)$ and $\left(\bar{x}_0, S_0^2\right)$ are the sample means and variances of the data in the two subgroups I_1 and I_0, respectively.

When I_1 is nonempty, an obvious lower bound of Ψ(data) is $S_1^2/(r+s)$, and if I_1 is empty, $\Psi(\text{data}) = S_0^2/n$. In the first case, U(data) is finite whenever $\int_0^{(x_i/c)}(h(r)/r)dr < \infty$ for $i \in I_0$. In the second case, we proceed as in the fully noise-perturbed case for normal and conclude that U (data) is finite under the conditions stated in (32) except that the bounds of r_i in the integrals are replaced by x_i/C. In particular, for uniform noise distribution, the conditions trivially hold.

Lognormal distribution. Proceeding as in the normal case with $u = \log(x/r)$, and breaking up the sum in the exponent into two parts corresponding to I_1 and I_0, we get the finiteness of corresponding U(data) under noninformative priors of μ and σ when the noise distribution is uniform.

Case (ii): Nature of data $(x_1, \ldots, x_n)$.

Normal distribution. Upon carefully examining the joint *pdf* of the data $\boldsymbol{x}$, given by (18), let us split the entire data into three mutually exclusive sets:

$$I_1 = \{i : x_i < 0\}, \quad I_2 = \{i : 0 < x_i < C\}, \quad I_3 = \{i : x_i > C\}.$$

It is now clear from standard computations under the normal distribution that whenever I_1 is non-empty, the posterior of (μ, σ) under a flat or noninformative prior of (μ, σ) will be proper. This is because the rest of the joint *pdf* arising out of I_2 and I_3 can be bounded under a uniform noise distribution or even under a general $h(.)$ under very mild conditions, and the retained part under I_1 will lead to propriety of the posterior. Likewise, if I_1 is empty but I_3 is non-empty, we can easily bound the terms in I_2, and proceed as in the fully noise-perturbed case for data in I_3 and show that the posterior is proper. Lastly, assume that the entire data fall in I_2, resulting in the joint *pdf* $L(\boldsymbol{\theta}|\text{data} \in I_2)$ as a product of terms of the type

$$f(x_i|\boldsymbol{\theta}) + \int_0^{(x_i/c)} f\left(\frac{x_i}{r_i}|\boldsymbol{\theta}\right)\frac{h(r_i)}{r_i}dr_i < \int_0^{(x_i/c)}\left[f(x_i|\boldsymbol{\theta})\frac{C}{x_{(1)}} + f\left(\frac{x_i}{r_i}|\boldsymbol{\theta}\right)\frac{h(r_i)}{r_i}\right]dr_i$$

where $x_{(1)} = \min(x_i)$. Let us now carefully check the product of the above integrands under the normal distribution, which will be first integrated wrt (μ, σ) under a flat or noninformative prior, and later wrt the noise variables which we take to be *iid* uniform. Obviously this product will be a sum of mixed terms of the following two types which are

relevant to check the propriety of the resultant posterior:

$$\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i\in J_1}(x_i-\mu)^2+\sum_{i\in J_2}\left(\frac{x_i}{r_i}-\mu\right)^2\right)\right]$$

where J_1 and J_2 form a partition of $\{1,\ldots,n\}$. It is now immediate that the terms of the first type (standard normal theory without any noise perturbation) will lead to a proper posterior of (μ,σ). Likewise, from our previous computations under the fully noise-perturbed case, it follows that the terms of the second type will also lead to propriety of the posterior of μ and σ under a uniform noise distribution.

Lognormal distribution. Proceeding as in the normal case above by replacing x/r by $u = \log(x/r)$, we get the posterior propriety of μ and σ under flat or noninformative priors when the noise is uniform. We omit the details.

8. References

An, D. and Little, R.J.A. (2007). Multiple Imputation: An Alternative to Top Coding for Statistical Disclosure Control. Journal of Royal Statistical Society, Series A, 170, 923–940.

Devroye, L. (1986). Non-Uniform Random Variate Generation: Springer.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). Bayesian Data Analysis, (second edition). Chapman & Hall/CRC.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 303–308.

Kim, J.J. and Winkler, W.E. (1995). Masking Microdata Files. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 114–119.

Kim, J.J. and Winkler, W.E. (2003). Multiplicative Noise for Masking Continuous Data. Statistical Research Division Research Report Series (Statistics #2003-01). U.S. Census Bureau. Available at: www.census.gov/srd/papers/pdf/rrs2003-01.pdf (accessed May 14, 2012).

Klein, M., Mathew, T., and Sinha, B. (2013). A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication. Center for Statistical Research & Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-02). U.S. Census Bureau. Available at: www.census.gov/srd/papers/ pdf/rrs2013-02.pdf (accessed Jan. 23, 2013).

Klein, M. and Sinha, B. (2013). Statistical Analysis of Noise Multiplied Data Using Multiple Imputation. Center for Statistical Research and Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-01). U.S. Census Bureau. Available at: www.census.gov/srd/papers/pdf/rrs2013-01.pdf (accessed Jan. 23, 2013).

Little, R.J.A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9, 407–426.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis With Missing Data, (second edition). Wiley.

Meng, X.L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. Statistical Science, 9, 538–558.

Nayak, T., Sinha, B.K., and Zayatz, L. (2011). Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection. Journal of Official Statistics, 27, 527–544.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at: www.R-project.org/.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.

Reiter, J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29, 181–188.

Reiter, J.P. (2005). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. Journal of Royal Statistical Society, Series A, 168, 185–205.

Reiter, J.P. and Raghunathan, T.E. (2007). The Multiple Adaptations of Multiple Imputation. Journal of American Statistical Association, 102, 1462–1471.

Robert, C.P. and Casella, G. (2005). Monte Carlo Statistical Methods, (second edition). Springer.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys: Wiley.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461–468.

Robins, J.M. and Wang, N. (2000). Inference for Imputation Estimators. Biometrika, 87, 113–124.

Sinha, B.K., Nayak, T., and Zayatz, L. (2012). Privacy Protection and Quantile Estimation From Noise Multiplied Data. Sankhya, Series B, 73, 297–315.

Tanner, M.A. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). Journal of the American Statistical Association, 82, 528–550.

Wang, N. and Robins, J.M. (1998). Large-Sample Theory for Parametric Multiple Imputation Procedures. Biometrika, 85, 935–948.

Received September 2012
Revised February 2013
Accepted May 2013

Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 467–469

Book Review

Books for review are to be sent to the Book Review Editor Jaki S. McCarthy, USDA/NASS, Research and Development Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.
Email: jaki_mccarthy@nass.usda.gov

Sabine Häder, Michael Häder, and Mike Kühne (Eds). *Telephone Surveys in Europe: Research and Practice*. Berlin: Springer-Verlag, 2012. ISBN 978-3-642-25410-9, 326 pp, €139.05.

The goal of *Telephone Surveys in Europe* is to provide a European perspective on the subject matter. The authors acknowledge the size and impact of American literature's contribution to the topic area but emphasise the distinctiveness of Europe and the need to consider the impact of cultural differences from the USA. This provides the motivation and key focus for the book.

The book is divided into five parts, covering: the development of telephone surveys in a selected number of European countries; associated sampling solutions; issues around weighting and nonresponse; data quality and finishes with recommendations. A useful summary of the book's contents and objectives is provided at the start of the book.

Part one of the book provides a collection of views and research evidence focused on the development of the design and implementation of surveys across Europe. The geographical and individual infrastructure of the respective countries covered within the book is used as an explanation of how surveys have developed. Researchers and national scientific institutions provide perspectives from Russia, the Netherlands, Switzerland, Finland, Italy, Portugal and the UK.

In Chapter 1, the Russian contribution to the book provides an interesting overview of how and why face-to-face surveys have continued to dominate in Russia, despite the rapid expansion of landline and mobile coverage. The author cites the impact of geography, availability of technology, political landscape and culture on dominant modes of data collection. Consideration is given to the challenges of producing adequate samples for telephone surveys due to landline coverage and lack of a national telephone register. While this makes for an interesting read, it was neither apparent at whom this level of detail is aimed nor where the information could be usefully applied.

Chapter 2 from the Netherlands was easy to read and informative. The author, Beukenhorst, provides a clear explanation of the popularity of face-to-face interviewing and the emergence of telephone interviewing. This provided a nice contrast to the scene set for the previous chapter in Russia, where telephone penetration had not reached the levels

of the Netherlands, and the later chapter from Finland, where attitudes to mobile phone registration differed from the other countries presented within the book.

An interesting debate is presented by Beukenhorst around the possibility of an increase in satisficing from those respondents who answer a mobile phone when on the move as well as issues around associated bias. This was quite thought provoking.

The contribution from Switzerland (Chapter 3) sets out some unique country-specific elements of telephone surveys, but in general a similar picture to that in other countries is presented. It was reassuring to find such commonalities across countries. However, it would have been good to see the key points from each chapter combined into a succinct position across Europe and contrasted with the USA. There are some good points made within part one of the book, but these are buried under detailed information that at times feels quite repetitive.

Contributions of particular note were those from Beukenhorst around satisficing (Chapter 2); Poggio and Callegaro's assertion of mobile and internet access rather than ownership as a better indicator of survey response (Chapter 6); Vicente and Reis's discussion of respondent distraction and multitasking when using a mobile phone and differences in completion rates (Chapter 7).

Part two of the book is divided into three chapters that look at the difficulty of contacting people by phone, sampling frames from a market research perspective, and mobile- and landline-onlys in dual-frame-approaches. The aim of Chapter 8 is to determine the potential bias caused by variations in accessibility and inclusion in telephone directories. Social integration, political opinion and sociodemographic characteristics are considered. The authors present a well thought-out and executed piece of research that utilises two large Swiss surveys (the ESS and EVS) and the EVS nonrespondent survey. A measure of the thoroughness of this work is the acknowledgement of the impact of the quality of questions on analysis. The authors use the reliability of questions across all modes to help inform which variables to include in their analysis; this serves as an important reminder to the reader. The results from the analysis are clearly illustrated through a series of tables and figures throughout the chapter.

Chapter 8 makes for an informative, well-written read that stimulates both thoughts and questions. A reasonable critique of the strengths and weaknesses of the work is provided. The authors provide the acronyms rather than providing the full survey title for the surveys. It would have been useful for the surveys' full titles to have been provided to enable readers to find out more about these surveys to further critique this work. While the book itself sets out to inform the methodology on telephone surveys, this chapter provides a nice platform for the debate on mixed mode data collection.

Similarly, Chapter 9 considers characteristics of respondents, but this time in relation to respondent mobile network connection and type of contract. The authors discuss the sampling frames and parameters of five European countries based on market research. This is where the book would have benefited from stronger links to earlier chapters. The discussion around the challenges of using telephone directories and random digit dialling is quite limited compared to some of the earlier discussions in part one of the book.

The introduction of weighting to the book begins with a discussion from Germany on the benefits of weighting for unequal inclusion and nonresponse using a dual frame sample. The chapter is clearly written and draws on research presented in an earlier

chapter. Reasons behind the methodology, the process of review and refinement and final conclusions make for an informative read.

The aim of the final part of the book is to make recommendations based on information in earlier chapters. However, the links between the last three chapters and earlier contributions is at times quite tenuous. Chapter 17 introduces a new concept of reciprocity based on the author's experimental work. This chapter links to earlier parts of the book, in that respondent reluctance to participate in surveys is raised. However, the discussion around this important topic is limited to the initial contact with respondents. This is followed by a discussion around the statistical and cost-related problems of an "optimal" dual frame approach to sampling and data quality. This chapter (Chapter 18) nicely sums up discussions from previous chapters, although it provides a further option rather than any firm conclusion from earlier discussions. The concluding chapter again introduces a new dimension rather than drawing together the proceeding chapters with a detailed account of an approach to fieldwork management.

The authors note that the book is written for "scientists and practitioners who deal with theory and application of telephone surveys in academic and market research". It would be helpful if the audience for this book were clearly identified and the structure appropriately tailored. There would also be great benefit from providing cross references between chapters to help the flow for the reader.

There are some useful contributions in this book but it appears to be more of a compendium of research findings. On average each chapter is approximately eight pages in length; this does not give the sort of depth required by survey professionals. Given the way the book is organised, it provides a useful compendium of research findings and discussions that may be useful to "career young" professionals looking for a general overview of telephone survey methodology.

Jayne Olney
Data Collection Methodology
Office for National Statistics
Government Buildings
Cardiff Road
Newport NP10 8XG
Telephone: 01633 456291
Email: Jayne.Olney@ons.gov.uk