The World's Simplest Survey Microsimulator (WSSM)

Alan F. Karr, Satkartar K. Kinney and Lawrence H. Cox
National Institute of Statistical Sciences
Research Triangle Park, NC 27709

1 The Need

More sharply than in the past, the future of official statistics surveys is framed by *data quality–cost tradeoffs*. In fact, in a larger sense, the issue may be decision quality–cost tradeoffs (Karr, 2012). Current and prospective budget pressures dictate that these tradeoffs by informed by scientific knowledge. We believe that there is a fundamental gap in survey science—there is no *simulation laboratory for surveys*. The World's Simplest Survey Microsimulator (WSSM) is a step toward filling this gap.

Despite impressive advances in survey theory and methodology, many questions remain to which extant knowledge only suggests the answers. One examples is the effects on uncertainties of editing, imputation, record linkage, and statistical disclosure limitation (SDL) (Cox et al., 2011). Operational decisions reflect the accumulated experience of field personnel, but currently and may always lack a theoretical basis.

Making survey science in part a laboratory science would have dramatic effect. However, of course, many "real-world" experiments are simply not possible. Simulation is a feasible, powerful alternative. For instance, in Karr (2011), we used simulation to study the differences among several configurations of the K–12 longitudinal studies conducted by the National Center for Education Statistics (NCES). Among conclusions that arose is that continuation of even small numbers of students from one study to the next is of limited statistical value. Using the real world as a laboratory in this case was infeasible.

Simulation is used in settings ranging from social networks to healthcare. It can make a difference for surveys, because the future can only be more challenging than the past as problems such as use of administrative data and disappearance of land-line telephones become more intense.

2 What is a Survey Microsimulator?

As suggested in §1, a survey microsimulator is an *in silico* simulation laboratory for surveys—a modular, extensible computer model (set of programs) that is:

Agent-Based, with explicit representation of the survey process, including

- **Entities:** Subjects, people, interviewers, . . . and their characteristics, *including survey variables*

Interactions: Interviews, nonresponse, callbacks, . . .

Costs, both fixed and variable

Operational decisions.

Powerful enough to handle realistic scales

Simple enough to conduct serious experiments

Credible enough to be used.

Responding to these criteria, the WSSM is deliberately oversimplified. Its short-term purpose is to support sensitivity analyses that

- Demonstrate that even simple models can reflect methodological, policy and operational considerations;
- Inform the course of more elaborate modeling efforts in the future.

In particular, a key goal is to provoke suggestions for enrichments.

Version 1 of WSSM has three essential characteristics:

1. The entire underlying population *and* the behavior on which the survey is focused are both simulated, and serve as "ground truth" for calculating measures of data quality.
2. The survey responses themselves are simulated.
3. The measures of data quality quantify the fidelity of inferences drawn from the survey responses compared to the same inferences based on the entire population.

3 Generalities

The focus of WSSM is *household surveys* involving interviews: sample units are households, and the survey responses are amounts spent on various categories of goods and services.

We employ the following notation:

- $\mathcal{P}$ = population
- i = data subject/survey unit (e.g., a household)
- X_i = population attributes for unit i
- Y_i = behavioral attributes for unit i, about which the survey questions are meant to elicit information
- $\mathcal{S} \subseteq \mathcal{P}$ = sample for the survey
- W_i = sample weight of $i \in \mathcal{S}$

- IC_j = characteristics of interviewer j
- IA_i = interviewer assigned to unit i
- $\mathcal{R} \subseteq \mathcal{S}$ = set of (unit) respondents
- W_i^* = nonresponse adjusted weight of $i \in \mathcal{R}$
- D_i = responses to survey questions from unit $i \in \mathcal{R}$
- C_i = cost associated with (attempted, if $i \notin \mathcal{R}$) data collection from unit i.

4 Global Assumptions and Structure

Essentially everywhere there could be a *conditional independence* assumption, there is one. For instance, in §6, the $\{Y_i : i \in \mathcal{P}\}$ are conditionally independent given $\{X_i : i \in \mathcal{P}\}$.

The WSSM is *extensible*, and therefore *modular*. In the next sections, we describe its principal modules. Figure 1 contains a flowchart showing the relationships among them.

5 Population Module

The **Population Module** generates the target population $\mathcal{P}$ for the survey. For simplicity, the frame is assumed to be the target population.

Generalities The Population Module creates (a subset of) the frame variables X_i for each member i of the target population. For experimental purposes, these can be taken from a public use dataset from one of the FedStats surveys, such as the American Community Survey (ACS).

Version 1 The "population" is an ACS microdata sample. There are four categorical attributes: number of adults in the household, number of children in the household, geography (at a level TBD) and the age of the head of household:

$$X_i = (\mathrm{NA}_i, \mathrm{NC}_i, \mathrm{G}_i, \mathrm{HHA}_i). \tag{1}$$

Extensions are to model undercoverage and add more attributes.

6 Behavior Module

A distinguishing characteristic of WSSM is that it *for all elements of the population it includes synthesized values of the behavioral variables* Y_i about which the survey is seeking to gather information. The **Behavior Module** generates these attributes for the population generated by the **Population Module**.

Generalities Conceptually, these are attributes Y_i will be synthesized from a previous, related survey, which is used to produce a conditional distribution $P(Y|X)$. Then, Y_i is generated by Monte Carlo simulation from $P(\cdot|X)$. If desired, multiple copies of the Y_i can be generated in order to assess replicate variability.

Version 1.0 Y consists of four continuous attributes—spending on housing, food, transportation and medical care:

$$Y_i = (\mathrm{SH}_i, \mathrm{SF}_i, \mathrm{ST}_i, \mathrm{SM}_i) = (Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}). \tag{2}$$

Extensions include categorical survey variables.

7 Sample Module

The **Sample Module** determines which units in the population are in the sample $\mathcal{S}$, and calculates their design weights $\{W_i : i \in \mathcal{S}\}$.

Generalities WSSM can accommodate almost any design that can be simulated. It is not necessary, but may be useful, to assume that every unit in the population has positive probability of being selected.

Version 1 Stratified sample based on the four X attributes, or simple random sampling.

Extensions include more complex designs.

8 Interviewer Characteristics Module

The **Interviewer Characteristics** module generates the population of interviewers for the survey.

Generalities The population of interviewer may be either finite or infinite. Each interviewer j is described by a vector IC_j of categorical or numerical characteristics.

Version 1 There is one (latent) binary characteristic of interviewers, for concreteness called *skill*. The population of interviewers is infinite. Those with Skill $= 1$ are more effective than those with Skill $= 0$.

Extensions include finite populations of interviewers, of whom a fixed number have skill equal to 1, as well as multi-dimensional characteristics.

9 Interviewer Assignment Module

The **Interviewer Assignment Module** models the process by which interviewers are assigned to sample units. Let IA_i denote the interviewer assigned to $i \in \mathcal{S}$.

Generalities When the population of interviewers is finite, construction of the assignment is a constrained optimization problem. When the population is infinite, the assignment is a (possibly randomized) function from possible values of X to possible values of the interviewer characteristics.

Version 1 Determined by a function $P(\text{Skill} = 1|X)$ that depends only on the age of the head of household.

Extensions are legion, including dynamic re-assignment of interviewers.

10 Unit Nonresponse Module

The **Unit Nonresponse Module** simulates which sample units $i \in \mathcal{S}$ are respondents.

Generalities Let $\text{UR}_i = 1$ if i is a respondent and $\text{UR}_i = 0$ otherwise. Then conceptually, generation of the UR_i requires a probability distribution

$$P(\text{UR}_i = 1|X_i, Y_i, \text{Skill}(\text{IA}_i)), \tag{3}$$

Let $\mathcal{R} = \{i \in \mathcal{S} : \text{UR}_i = 1\}$ denote the set of respondents.

This module also will adjust the sample weights $\{W_i : i \in \mathcal{S}\}$ for unit nonresponse, yielding adjusted weights $\{W_i^* : i \in \mathcal{R}\}$.

Version 1 The probability of unit nonresponse depends only on the skill of the interviewer, and is lower for skilled interviewers:

$$P(\text{UR}_i = 1|X_i, Y_i, \text{Skill}(\text{IA}_i)) = \begin{cases} p_0 & \text{if Skill}(\text{IA}_i) = 0 \\ p_1 & \text{if Skill}(\text{IA}_i) = 1 \end{cases}, \tag{4}$$

where $p_0 > p_1$, so that skilled interviewers are less likely to produce unit nonresponse.

The W_I^* will be calculated using weighting class adjustments based on the sample strata.

Extensions are multiple—for instance, allowing dependence on Y_i in (3) removes—or at least reduces—nonresponse bias.

11 Responses Module

The **Responses Module** simulates the data—respondents' answers D to the survey questions from their underlying behavior. The responses need not be "correct."

Generalities In general, these would given by conditional probabilities of the form

$$P(D_i|X_i, Y_i, \text{Skill}(\text{IA}_i), \text{UR}_i = 1). \tag{5}$$

Version 1 Recall that the behavior of interest Y consists of four numerical variables (§6). Assume that the four questions are of the form "How much did your household spend on [. . .] during the period [. . .]?" Responses are then

$$D_i = Y_i + \varepsilon_i, \tag{6}$$

where $\varepsilon_i \sim N(0, \sigma^2_{\text{Skill}(\text{IA}_i)})$, with $\sigma_0^2 > \sigma_1^2$. That is, the variance is higher for less skilled interviewers.

Extensions include item nonresponse, dependence of the variance on Y, dependence among responses, and more detailed modeling, e.g., of cognitive characteristics of questions.

Let $\mathcal{D} = \left\{(D_{i,1}, D_{i,2}, D_{i,3}, D_{i,4}) : i \in \mathcal{R}\right\}$ denote the set of responses.

12 Data Quality Module

The **Data Quality Module** compares estimates of population characteristics derived from the data $\mathcal{D}$ with the *known—because the entire population's behavior was simulated*—corresponding characteristics of the population.

Generalities As argued in Karr et al. (2006), Cox et al. (2011) and elsewhere, the ultimate measures of data quality are measures of the quality of decisions taken on the basis of $\mathcal{D}$. This perspective is currently too remote to be useful, and instead, measures based on the quality of inferences will be employed. Since $\left\{(Y_{i,1}, Y_{i,2}, Y_{i,3}, Y_{i,4}) : i \in \mathcal{P}\right\}$ is known, the comparison is then between inferences from the latter and corresponding inferences from $\mathcal{D}$.

Version 1 Two measures will be employed initially:

- *Population estimates.* Comparison of the (finite population) mean μ_Y and covariance matrix Σ_Y^2, which can be calculated exactly, to estimates $\widehat{\mu_Y}$ and $\widehat{\Sigma_Y^2}$ derived from $\mathcal{D}$ and the weights W_i^*.
- *Fitted tables.* Consider the four "mean" contingency tables in which cells are defined by possible values of the population attributes X and entries are averages of the four behavioral attributes over the corresponding subsets of $\mathcal{P}$. Each of these can be modeled using a variety of methods, including log-linear models (Bishop et al., 1975). The same models will be fit to these four tables and the corresponding tables derived from $\mathcal{D}$, and the results compared as in Dobra et al. (2002).

Extensions are many, based on ideas in various papers from NISS and others.

13 Cost Module

The **Cost Module** calculates the costs associated with the survey.

Generalities Survey costs comprise both fixed and variable components, borne by multiple organizations. Some costs, especially fixed agency costs and costs considered by contractors to be proprietary may be difficult to elucidate (Karr and Last, 2006).

Version 1 There will be solely a variable cost C_i for each unit i, which depends only on the skill level of the interviewer IA_i and whether i is a unit nonrespondent:

$$C_i = \begin{cases} c_{00} \text{ if } \mathrm{Skill}(\mathrm{IA}_i) = 0 \text{ and } \mathrm{UR}_i = 0 \\ c_{10} \text{ if } \mathrm{Skill}(\mathrm{IA}_i) = 1 \text{ and } \mathrm{UR}_i = 0 \\ c_{01} \text{ if } \mathrm{Skill}(\mathrm{IA}_i) = 0 \text{ and } \mathrm{UR}_i = 1 \\ c_{11} \text{ if } \mathrm{Skill}(\mathrm{IA}_i) = 0 \text{ and } \mathrm{UR}_i = 1 \end{cases} . \tag{7}$$

Extensions See Groves (2004). Ultimately, costs should also have stochastic elements.

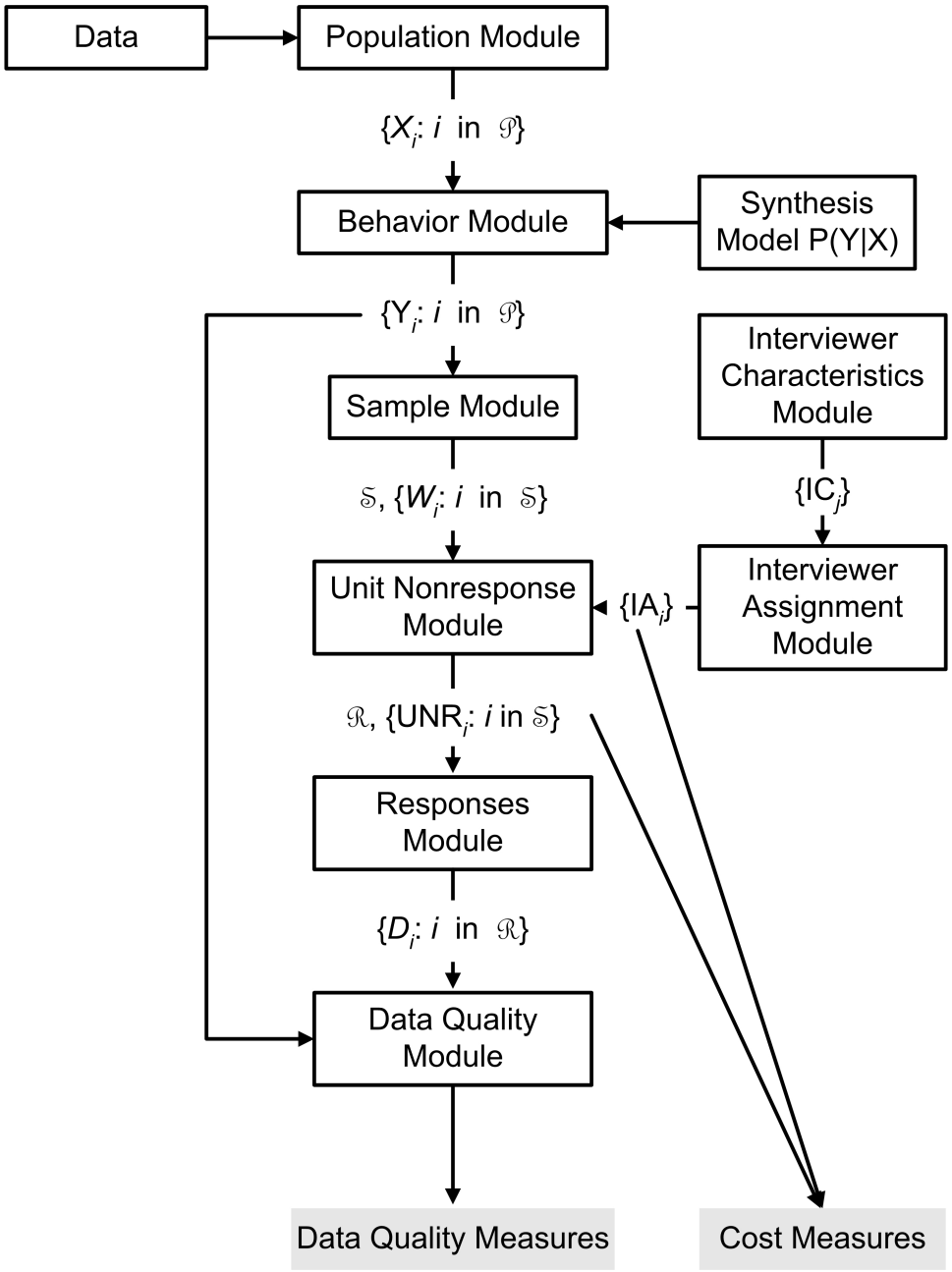

Figure 1: Flowchart for WSSM.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Cox, L. H., Karr, A. F., and Kinney, S. K. (2011). Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act (with discussion). *Int. Statist. Rev.*, 79(2):160–199.

Dobra, A., Fienberg, S. E., Karr, A. F., and Sanil, A. P. (2002). Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544.

Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley, New York.

Karr, A. F. (2011). National Institute of Statistical Sciences Configuration and Data Integration Technical Panel: Final report. Technical report, National Center for Education Statistics, Washington. NCES publication 2011607, available on-line at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011607.

Karr, A. F. (2012). Discussion on statistical use of administrative data: Old and new challenges. *Statist. Neederlandica*, 66(1):80–84.

Karr, A. F. and Last, M. (2006). Survey costs: Workshop report and white paper. Technical Report 161, National Institute of Statistical Sciences. Available on-line at http://niss.org/sites/default/files/tr161.pdf.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173.