NISS

The World's Simplest Survey Microsimulator (WSSM)

Alan F. Karr and Lawrence H. Cox

Technical Report 181
October 2012

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC
www.niss.org

The World's Simplest Survey Microsimulator (WSSM)

Alan F. Karr 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* dictated by current and anticipated budget pressures. In a larger sense, the issue may be *decision quality–cost* tradeoffs (Karr, 2012), because society may deem the resultant decisions—not the data—to be the end product of official statistics. In either case, whether official statistics agencies will be participants or bystanders as events unfold remains to be seen. The most pressing short-term need is to ensure that quality–cost tradeoffs be informed by scientific knowledge and reasoning. Efforts to do this, we believe, are hindered by a fundamental gap: currently, *survey science is not to any meaningful extent a laboratory science*. The World's Simplest Survey Microsimulator (WSSM) is a step toward filling this gap.

To explore three issues—need, utility and feasibility—surrounding simulation models for Federal surveys, in April 2011 the National Institute of Statistical Sciences (NISS) sponsored an interdisciplinary Workshop on Microsimulation Models for Surveys. Details and supporting papers for the workshop are available.[1] Cox (2012), and Cox (2013) delve deeper into these issues and offer a case for development of a simulation laboratory for Federal surveys. Karr et al. (2012) presents a prototype design for WSSM.

To make the issues more concrete, we pose the following question:

> Which of the following strategies most improves the quality of a household expenditure survey, such as the Consumer Expenditure Survey (CES) conducted by the U.S. Bureau of Labor Statistics (BLS):
>
> - A 10% increase in sample size?
> - A 10% decrease in measurement error?
> - Imposition of edit rules that replace "erroneous" data values by imputed values?
> - All of the above?
>
> and at what cost?

[1] At `http://www.niss.org/events/workshop-microsimulation-models-surveys`.

That we are currently not able even to frame these questions in a manner that allows them to be addressed confirms the breadth of the gap. In §4 we show how WSSM can answer this question.

Making survey science in part a laboratory science would have dramatic effect. But, of course, most "real-world" experiments are simply not feasible. One cannot answer our question by conducting the CES in four different ways, either over the next four years or by subsetting the population. "Expert opinion," while often insightful, especially with respect to survey operations, equally often amounts to little more than speculation.

Simulation is a feasible, powerful alternative, and there are precedents in official statistics. For instance, Karr (2011) used simulation to study the differences among several configurations of the K–12 longitudinal studies of students 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 was infeasible in this case. A microsimulation model for field operations and costs in the National Health Interview Survey (NHIS) is discussed in Chen (2008) and Chen (2012).

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

The remainder of this paper is organized in the following manner. In §2 we describe survey microsimulators in general, and in §3 we describe WSSM in particular. The results of the experiment just laid out appear in §4, while §5 contains discussion and conclusions.

2 What is a Survey Microsimulator?

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 dynamics of the survey process, including entities—subjects, people, interviewers, ... and their characteristics, and especially survey variables; interactions among the entities—interviews, nonresponse, callbacks, ...; costs, both fixed and variable; and operational decisions. A useful microsimulator must also be transparent enough that users can understand it, powerful enough to handle realistic scale, simple enough to conduct detailed experiments, and credible enough to be used.

Responding to these criteria and as its name implies, the WSSM is deliberately simple. We do not purport that the current version answers the question in §1 definitively, but it does show that there are differences among the strategies described there. Perhaps more important, WSSM supports sensitivity analyses demonstrating that even simple models can reflect methodological, policy and operational considerations, as well as inform the course of more elaborate modeling efforts in the future.

3 WSSM Version 1

Version 1 of WSSM has three essential characteristics. First, both the entire underlying population *and* the behavior on which the survey focuses are simulated, to serve as "ground truth" for calculating measures of data quality. Second, the complete survey process, including the survey responses themselves, is simulated. Finally, WSSM contains measures of data quality that quantify the fidelity of inferences drawn from the survey responses compared to the same inferences based on the population.

The focus of Version 1 of WSSM is *household surveys* involving interviews, via Web, telephone—computer-assisted telephone interview (CATI) and personal—computer-assisted personal interview (CAPI). Sample units are households, and the survey responses are amounts spent on various categories of goods and services, as well as demographic information about household members. The main objects simulated are:

Population: Categorical (integer-valued) frame variables, as well as categorical and numerical response variables—possibly satisfying constraint rules, together with a geographical location, a single stratum variable, a propensity to respond and item nonresponse probabilities.

Interviewers for both CATI and CAPI: Location, skill level, unit response probability factors, measurement error parameters, and costs.

The survey process: Selection of the sample; WEB, CATI and CAPI stages with interviewer assignment; unit nonresponse depending on subject and interviewer skill; up to three contact attempts, with increasing incentives, and omitted items at the last stage; item nonresponse; edit rules that either designate responses violating them for imputation or flag those responses; imputation of missing items and designated violations of edit rules, using means or resampling; weights reflecting the design and adjusted for unit nonresponse.

Costs: For CATI and CAPI contacts; for CATI and CAPI interviews, both per household and per person; for incentives; for out-of-location assignment of CAPI interviewers; and for data edits.

Data utility measures: Global measures that compare responses to the population: specifically, Hellinger distance for frame and categorical survey variables and Kullback–Liebler divergence for numerical survey variables. See §3.3 for details.

The population, sample, CATI interviewers and CAPI interviewers are simulated at the individual—agent—level.

3.1 Structure

In this paper, we emphasize functionality of WSSM over the details of the software. Briefly, WSSM consists of four executable programs, written in the C language and compiled using GCC

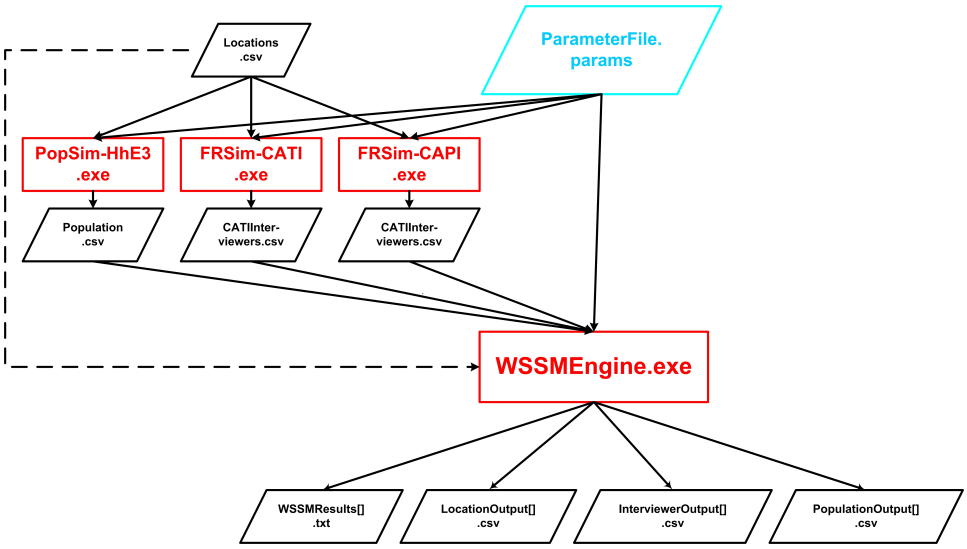

Figure 1: Flowchart for WSSM.

on Microsoft Windows:[2]

Population simulator `PopSim-HhE3.c`: $\sim$ 650 lines of source code; 91 KB executable.

Interviewer simulators `FRSim-CAPI.c` for CAPI interviewers and `FRSim-CATI.c` for CATI interviewers: each $\sim$ 300 lines of source code and KB executable.

Computational engine `WSSMEngine.c`: $\sim$ 3400 lines of source code; 152 KB executable.

There are also header files comprising $\sim$ 800 lines of code.

All four programs are executed from the Windows command line, and draw their inputs principally from a an ASCII text file—the *parameter file* discussed in §3.2. There is also a comma-separated value (CSV) file containing location information; in the current implementation, the locations are the 50 U.S. states and Washington, DC.

Figure 1 shows the relationships among the input files, the programs and the output files, which are discussed further in §3.6. All output files are either CSV or ASCII text.

3.2 The Parameter File

The most direct way to understand WSSM functionality in more detail is by means of the parameter file that is read by all four of the executable programs. This file is prepared using a text editor. As shown in Figure 2, most entries are of the form

[2]Fewer than 50 lines of the source code are specific to Windows, so porting to another operating system would be straightforward.

```
ParameterName = Value
```

For instance, the CSV file containing information about locations is `Locations.csv`, there are 51 locations, and each has four characteristics—a name, a cost factor that alters global interviewer costs, a price factor derived from the Consumer Price Index (CPI) that adjusts expenditures and the fraction of the national population living in that location.

The principal sections of the parameter file are as follows.

POPULATION The program used to simulate the population, the file containing population characteristics, the size of the population, constraints that must be satisfied by the survey variables. See §3.4 for details.

VARIABLES The names (and implicitly, the numbers of) three classes of variables: categorical frame variables, numerical survey variables and categorical survey variables. To illustrate, the four frame variables are the number of adults in the household, the number of children in the household, the age of the householder and the gender of the householder. The six numerical survey variables are total monthly income and monthly expenditure on education, housing, food, transportation and medical care. The five categorical survey variables are householder race, householder educational attainment, householder employment status the number of vehicles in the household and the number of household members who are students.

SURVEY The sample size, the sampling protocol (SRS—simple random sampling—is the only option implemented to date), and which of the Web, CATI and CAPI data collection stages are present.[3]

EDIT RULES These rules correct error in the response data, that is, violations of the constraint rules on the population, or of other specified relationships. In general, these violations are construed to be consequences of measurement error. No edit rules are present in the parameter file in Figure 2, but see §3.4 for elaboration and Figure 3 for examples.

ANALYSIS The imputation methods to be used for numerical and categorical survey variables. See §3.3.

CAPI INTERVIEWERS The program used to simulate CAPI interviewer characteristics, the CSV file in which their characteristics are stored, the number of interviewers, the fraction of them that are of high skill—to be explained momentarily, and the number of CAPI interviewer characteristics. These characteristics include the maximum number of interviewers, the minimum and maximum unit response probabilities, the minimum and maximum measurement error standard deviations and minimum and maximum interview costs per household and per person, the minimum and maximum costs per contact, the minimum and maximum costs for out-of-location interviews and the incentives offered with the first, second

[3] Always in this order, and, of course, only if previous stages did not produce a response.

and third interview attempts. Actual values of unit response probabilities, measurement error standard deviations various costs are chosen at random between the minimum and the average of the minimum and maximum values for low skill interviewers, and between the average and the maximum for high skill interviewers.

CATI INTERVIEWERS Similar information for CATI interviewers, which is omitted from Figure 2.

WEB A response probability, measurement error standard deviation, cost per contact, cost per unit, cost per person, number of contact attempts and incentive level, offered only at the first attempt.

Changing the parameters is straightforward: the user simply edits the parameter file. In §4 we show the changes associated with the experiment laid out in §1.

3.3 What WSSM Simulates

In this section, we elaborate on exactly what is simulated within the WSSM framework. An alternative description in more mathematical notation is in Karr et al. (2012). We stress that WSSM is a stochastic simulator: characteristics of the population, the interviewers and the survey process are random, chosen according to specified probability distributions and with specified parameters for those distributions. Currently, the distributions themselves, and in some cases, the parameters themselves are "hard-coded," appearing in the source code rather than being specified in the parameter file (§3.2).

Population. Individual households and members, together with their characteristics (frame and response variables, . . .) are simulated using probability distributions and parameter values that are hard-coded in `PopSim-HhE3.c`. Specifically,

- `Adult` is distributed on {1, 2, 3} with probabilities {0.35, 0.5, 0.15}.
- `HhAge` is uniformly distributed on {20, . . . , 75}.
- `Child` is uniformly distributed on {0, 1, 2, 3} with probabilities {.3, .3, .3, .1} when `Adult` ≥ 2 and `HhAge` ≥ 25, and is 0 otherwise.[4]
- `HhGend` is uniformly distributed on {0, 1}.
- The distribution of `Location` is that of the U.S. population, using the 2010 Census.
- `Stratum` is household size: `Adult` plus `Child`.
- `UnitResponseProbability` is uniformly distributed on [0.8.1.0], with a slight bias in favor of younger householders.

[4]This is merely to show what is possible, and not a statement about single parents.

```
*** Experiment7-HotDeck-BaseCase.params
*** WARNING: DO NOT CHANGE ANYTHING TO THE LEFT OF THE EQUAL SIGNS ***
*** 2012/10/14 ***
>>> MULTIPLE-USE
LocationCSVFile = Locations
NumberLocations = 51
NumberLocationCharacteristics = 4
>>> POPULATION
PopulationSimulator = PopSim-HhE3
PopulationCSVFile = Population7
PopulationSize = 100000
>>> VARIABLES
FrameVariableName = Adult
FrameVariableName = Child
FrameVariableName = HhAge
FrameVariableName = HhGend
CategoricalSurveyVariableName = Race
CategoricalSurveyVariableName = HhEdAt
CategoricalSurveyVariableName = HhEmSt
CategoricalSurveyVariableName = Vehicle
CategoricalSurveyVariableName = Student
NumericalSurveyVariableName = Income
NumericalSurveyVariableName = Education
NumericalSurveyVariableName = Housing
NumericalSurveyVariableName = Food
NumericalSurveyVariableName = Transp
NumericalSurveyVariableName = Medical
CONSTRAINTS ON POPULATION
BoundConstraint = Housing GE 0.0
BoundConstraint = Food GE 0.0
BoundConstraint = Transp GE 0.0
BoundConstraint = Medical GE 0.0
SumConstraint = Student LE Adult + Child
SumConstraint = Housing + Food + Transp + Medical LE Income
RatioConstraint = Food LE 1.0 * Housing
>>> SURVEY
SampleSize = 5000
SampleDesign = SRS
WEBStage = Yes
CATIStage = Yes
CAPIStage = Yes
>>> EDIT RULES
>>> EDIT COSTS
EditCostPerItem = 25.00
>>> ANALYSIS
NumericalImputationMethod = HotDeck
CategoricalImputationMethod = HotDeck
>>> CAPI INTERVIEWERS
CAPIInterviewerSimulator = FRSim-CAPI
CAPIInterviewerCSVFile = CAPIInterviewersB
CAPINumberInterviewers = 500
CAPIFractionHighSkillInterviewers = .25
CAPINumberInterviewerCharacteristics = 8
CAPIMaximumInterviews = 50
CAPIResponseProbMin = 0.1
CAPIResponseProbMax = 0.4
CAPINoiseStdDevMin = 100.0
CAPINoiseStdDevMax = 400.0
CAPICostUnitMin = 80.0
CAPICostUnitMax = 100.0
CAPICostPersonMin = 30.0
CAPICostPersonMax = 50.0
CAPICostContactMin = 20.0
CAPICostContactMax = 30.0
CAPICostOutOfLocationMin = 100.0
CAPICostOutOfLocationMax = 150.0
CAPINumberContactAttempts = 3
CAPIIncentiveAttempt1 = 15.0
CAPIIncentiveAttempt2 = 30.00
CAPIIncentiveAttempt3 = 50.00
>>> CATI INTERVIEWERS [ANALOGOUS TO CAPI]
>>> WEB
WEBResponseProb = 0.25
WEBNoiseStdDev = 500.0
WEBCostContact = 5.0
WEBCostUnit = 10.0
WEBCostPerson = 10.0
WEBNumberContactAttempts = 1
WEBIncentiveAttempt1 = 20.0
```

Figure 2: WSSM parameter file for the base case of the experiment described in §1 and §4.

- `Income` has a normal distribution $N(7000, 100000)$, and is then multiplied by the location-specific price factor.
- `Education`, `Housing`, `Food`, `Transp`, `Medical` are all normally distributed, but are correlated and multiplied by the location-specific price factor. Means depend on the numbers of adults and children.
- `Race` is distributed on {0, 1, 2, 3, 4} with probabilities {.5, .20, .20, .05, .05}.
- `HhEdAt` is distributed on {0, 1, 2, 3} with probabilities {.05, .6, .3, .1}.
- `HhEmSt` is uniformly distributed on {0, 1}.
- `Vehicle` is distributed over {0, 1, 2, 3, 4} with probabilities (.4, .4, .15, .05).
- `Student`: All children are students, and each adult is a student with probability .17.
- Item nonresponse probabilities are correlated, and are higher for those with higher incomes.

CATI and CAPI Interviewers. Individual interviewers are simulated; the numbers of interviewers are set in the parameter file. Interviewer characteristics are those listed in §3.2: a randomly chosen skill, unit response probability modifier, measurement error standard deviation, and costs per contact, unit and person, the latter modified by location-specific cost factors. Most of these parameters can be changed via the parameter file.

The Survey. Principal steps are to simulate:

1. Selection of the sample, which as noted previously is currently possible only via SRS.
2. As specified in the parameter file, up to three stages of data collection: WEB, CATI and CAPI, in that order. Numbers of contact attempts and incentives are set in the parameter file. For each stage, WSSM represents explicitly unit nonresponse, modeled as the product of sample case-dependent and interviewer-dependent factors; item nonresponse; and measurement error.
3. Data processing, including
    - Adjustment of weights for unit nonresponse, using the `Stratum` variable.
    - Application of the edit rules (§3.4), resulting in entries' being either designated for imputation or flagged (for later review that is not currently modeled in WSSM).
    - Imputation of missing items and, if prescribed in the parameter file, violators of the edit rules. Currently available options are: *HotDeck*, meaning resampling from sample cases with neither item response nor edit rule violations; *Mean (Mode)*, replacing by global means (numerical variables) or global modes (categorical variables); and *LocationMean (LocationMode)*, replacing by location-specific means (numerical variables) or location-specific modes (categorical variables).

National Estimates. For categorical survey variables, WSSM calculates and reports Horvitz–Thompson estimators (Horvitz, 1952) of the marginal distribution of each categorical survey variable, based on the sample, unit respondents, and final data; and of the mean and covariance matrix of the entire set of numerical survey variables, based on the sample, unit respondents, and final data, as well as the corresponding objects for the population.

Data Quality Measures. A central strength of WSSM is that since the actual values of the survey response variables are simulated for all units in the population, comparisons are possible among all of the following: actual values for the population; actual values for the sample; actual values for unit respondents; final values, incorporating measurement error, edit and imputation, for unit respondents.

The numerical measures used to quantify these comparisons are well-known "metrics" for discrete and continuous probability distributions. For categorical variables, WSSM uses *Hellinger distance*: the Hellinger distance between distributions P and Q on a finite set C (Think of C as cells in a contingency table.) is

$$\mathrm{HD}(P, Q) = \sum_{c \in C} \left(\sqrt{P_c} - \sqrt{Q_c} \right)^2 . \tag{1}$$

WSSM applies (1) to all (frame or categorical survey) variables simultaneously, which requires an appropriate data structures for the associated contingency tables; lists of (cell coordinates, cell count) pairs are used (Karr et al., 2007), in order to exploit sparsity.

For continuous variables, WSSM employs *Kullback–Liebler divergence*, which for density functions $f > 0$ and $g > 0$ on $\mathcal{R}^d$ is given by

$$\mathrm{KL}(f, g) = \int_{\mathcal{R}^d} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx. \tag{2}$$

In practice, the numerical integration necessary to calculate $\mathrm{KL}(f, g)$ using (2) is very difficult to implement even for $d = 3$, let alone for the six numerical survey variables in our experiment. WSSM instead employs an approximation based on the assumption that both densities are multivariate normal. If $f = N(\mu_0, \Sigma_0)$ and $g = N(\mu_1, \Sigma_1)$, then

$$\mathrm{KL}(f, g) = \frac{1}{2} \left[\mathrm{tr} \left(\Sigma_1^{-1} \Sigma_0 \right) + (\mu_1 - \mu_0)^T - \ln \left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)} \right) - d \right], \tag{3}$$

where $\mathrm{tr}(M)$ and $\det(M)$ are the trace and determinant of the matrix M. In WSSM, the matrix inversion in (3) is performed by means of Gaussian elimination.

WSSM calculates and reports: *for frame variables*, Hellinger distances between the population and the sample and between the population and the unit respondents; for *categorical survey variables*, Hellinger distances between the population and the sample, between the population and the unit respondents, and between the population and the final data; and for *numerical survey variables*, Kullback–Liebler divergence between the population and the sample, between the population and the unit respondents, and between the population and the final data.

3.4 Constraint Rules and Edit Rules

WSSM allows four different kinds of *constraint rules* that must be satisfied by the survey variables for each population unit and are applied by the population simulator, as well four finds of *edit rules* that can applied by the computational engine, which offers two options when rules are violated. Responses violating edit rules can be either imputed (in the case of HotDeck imputation, from records satisfying all edit rules) or simply flagged for later analysis. Such analysis is not now modeled in WSSM.

Although they are logically distinct, the constraint rules and the edit rules are linked by the long–practiced concept of data edits. Because population elements are generated stochastically, they may not satisfy physical constraints or plausibility relationships. The constraint rules prevent violation of such constraints or relationships. For instance, the population in `Population7.csv`, the input file for our experiment, is required to satisfy the constraint rules shown in Figure 3. This figure shows the syntax for three classes of rules:

BoundConstraints of the form $V \leq c$ or $V \geq c$, where V is a response variable and c is a constant. So, the first BoundConstraint in Figure 3 requires that `Housing` ≥ 0.

SumConstraints of the form $V_1 + ... + V_k \leq V_0$ or $V_1 + ... + V_k \geq V_0$, where the V_j can be frame variables or response variables.[5] The first SumConstraint in Figure 3 requires that the number of students in a household (a response variable) not exceed the number of adults (a frame variable) plus the number of children (another frame variable).

RatioConstraints of the form $V_1/V_2 \leq c$ or $V_1/V_2 \geq c$, where V_1 and V_2 are frame or response variables and c is a constant. Only the "$\leq$" form is implemented.

The fourth class of constraint rules is:

ConsistencyConstraints of the form "$V_1 = a$ is inconsistent with $V_2 = b$," where V_1 and V_2 are frame or categorical response variables, and a and b are constants. A typical example is "`Age` = 2 is inconsistent with `MaritalStatus` = Married."

Edit rules correct violations of the constraint rules caused by measurement error. WSSM currently has only one generic form of measurement error, which is presumed to cover such phenomena as misinterpreted questions, "lying" by respondents and interviewer error. The edit rules corresponding to the constraint rules in Figure 3 appear in Figure 4. It is not logically necessary that the two sets of rules be identical, but they are easiest to understand when they are identical. The edit rules, when invoked with the "Impute" option, force responses to satisfy the same constraints as the population does.

The syntax for edit rules is almost identical to that for constraint rules, with the addition of the Flag/Impute option. The "Impute" option forces imputation of all survey variables appearing in the rule, and similarly for the "Flag" option. There is no attempt to determine which variable(s) is (are) at fault.[6]

[5] At least one of them must be a response variable.

[6] In some settings, an action of the form "In this case, replace 'Married' by 'Single'." might be prescribed. Cur-

```
CONSTRAINTS ON POPULATION
BoundConstraint = Housing GE 0.0
BoundConstraint = Food GE 0.0
BoundConstraint = Transp GE 0.0
BoundConstraint = Medical GE 0.0
SumConstraint = Student LE Adult + Child
SumConstraint = Housing + Food + Transp + Medical LE Income
RatioConstraint = Food LE 1.0 * Housing
```

Figure 3: Excerpt from the WSSM parameter file for the experiment described in §1 and §4, showing the constraints on the survey variables.

```
>>> EDIT RULES
BoundEdit = Housing GE 0.0 Impute
BoundEdit = Food GE 0.0 Impute
BoundEdit = Transp GE 0.0 Impute
BoundEdit = Medical GE 0.0 Impute
SumEdit = Student LE Adult + Child Impute
SumEdit = Housing + Food + Transp + Medical LE Income Impute
RatioEdit = Food LE 1.0 * Housing Impute
>>> EDIT COSTS
EditCostPerItem = 25.00
```

Figure 4: Excerpt from the WSSM parameter file for the "edit rules" case of the experiment described in §1 and §4. The edit rules are identical to the constraint rules in Figure 3.

3.5 Running WSSM

All WSSM executables are invoked from the command line, with syntax of the form

`WSSMEngine ParameterFileName`

Figure 5 shows the associated screen output for the base case in our experiment. The machine employed has reasonable capabilities: Microsoft Windows 7 operating system, 6-core processor and 32 GB of memory; it is not stressed. With a proper batch mode capability, which is under development, sensitivity analyses comprising 10,000 cases can be run in one day.

3.6 WSSM Output

As shown in Figure 1, each of `PopSim-HhE3.c`, `FRSim-CAPI.c` and `FRSim-CATI.c` produces a single CSV output file that provides input to `WSSMEngine`. These files can also be analyzed statistically. For instance, Figure 6 contains histograms of the six numerical survey variables in `Population7.csv`, which is the population file used in our experiment.

`WSSMEngine` produces five output files, which are also shown in Figure 1. The file naming convention is `[NAME]_YEAR_MONTH_DAY_HOUR_MINUTE_SECOND.csv`.[7] Four of these are CSV files meant primarily for statistical analysis:

`InterviewerOutput_2012_10_15_11_41_20.csv` contains one record per interviewer, with information such as mode, location, unit response probability, cost parameters, assigned

rently, WSSM has no such capability.

[7] This convention ensures that files are never overwritten accidentally.

```
e:\NISS\SurveyMicrosimulator\WSSM-v1\Software>WSSMEngine Experiment7\Experiment7-HotDeck-BaseCase

Parameter names read from WSSMParameterNames.txt: 70 names
Parameter values read from Experiment7\Experiment7-HotDeck-BaseCase.params and parsed: 84 parameters

Arrays initialized
Location-specific data read from Locations.csv; 51 locations
Population data read from Population7.csv; population size = 100000
Population means, covariances, categorical variable tables calculated
Sample of size 5000 drawn using SRS, and sampling weights generated
Sample frame and categorical survey variable tables generated
Sample means and covariance for numerical survey variables calculated
Web data collection initialized
CATI interviewer data read from CATIInterviewersB.csv: 250 interviewers
CAPI interviewer data read from CAPIInterviewersB.csv: 500 interviewers
Web responses generated: 738 responses
CATI interviews generated: 1503 responses
CAPI interviews generated: 1080 responses
Total unit respondents: 3321
Weights adjusted for unit nonresponse
Respondent frame and categorical survey variable tables generated
Respondent means and covariances for numerical survey variable tables generated
Item nonresponses generated
Edit rules parsed: 0 rules
Edits executed: 0 rules, 0 entries flagged, 0 imputations generated
Imputation completed using HotDeck for numerical and HotDeck for categorical: 0 edit imputations, 20
96 missing value imputations
Categorical response variable tables generated
Response variable statistics calculated
Hellinger distances calculated
K-L divergences for numerical survey variables calculated
Cost calculations completed; total cost = $1,040,487

Output file written to WSSMResults_2012_10_15_13_23_21.txt

CSV population data written to PopulationOutput_2012_10_15_13_23_21.csv
CSV interviewer data written to InterviewerOutput_2012_10_15_13_23_21.csv
CSV location data written to LocationOutput_2012_10_15_13_23_21.csv
Categorical variable tables written to TableOutput_2012_10_15_13_23_21.csv

WSSMEngine Execution time: 7.960 seconds
Memory usage: 16% of 32693 MB

e:\NISS\SurveyMicrosimulator\WSSM-v1\Software>_
```

Figure 5: Screen output when `WSSMEngine` is run on the "all options" parameter file `Experiment7-HotDeckBaseCase.params`.

and completed interviews, and incurred costs. An excerpt is shown in Figure 7. An illustrative analysis is the histogram of interviewer-level costs in Figure 8.

`LocationOutput_2012_10_15_11_41_201.csv` contains one record per location, with location 0 corresponding to the entire US. The information includes population and sample counts, data quality measures and costs; see the column headings in Figure 9.

`PopulationOutput_2012_10_15_11_41_20.csv` contains complete information for every unit of the population, with units in the sample preceding those not sampled. There are 55 variables for each unit, including frame and survey variables, the assigned interviewer, unit and item nonresponse status and costs. Figure 10 shows the histogram of (total) cost over the sample.

`TableOutput_2012_10_15_11_41_20.csv` is a specialized file containing the full contingency tables for the frame variables and categorical survey variables.

The fifth WSSM output file is a text file, meant for reading by human analysts. Its name is of the form `WSSMResults_2012_10_15_11_41_20.txt`, and most of it appears in Figures 11 and 12. Since this file is virtually self-explanatory, we note only that its main components are:

- Run information, especially the software version and the names of the input files, together with selected parameters.

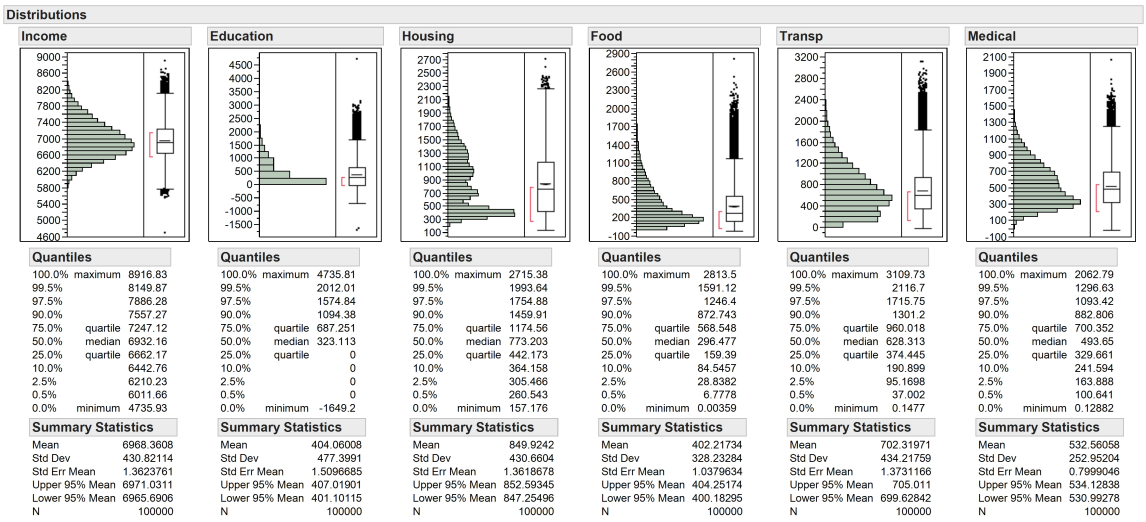

Figure 6: Histograms of the six numerical survey variables, from `Population7.csv`.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Mode	Loc	unitRespI	NoiseVar		Cost/Con	Cost/Unit	Cost/Pers	Cost/OutOfLoc		MaxInt	AssignInt	ComplInt		ContCost	UnitCost	PersonCo	OutOfLoc	TotalCost
2	WEB	N/A	0.25	250000		5	10	18570	0		0	5000	759		28795	7590	18570	0	54955
3	CATI	CA	0.338142	22844.5		18.97671	0	28.50276	0		100	22	5		157.837	0	399.0387	0	556.8757
4	CATI	GA	0.291931	9094.811		14.53688	0	12.26234	0		100	15	4		107.2213	0	159.4104	0	266.6317
5	CATI	IL	0.212046	14234.22		10.01999	0	16.17145	0		100	7	2		50.07996	0	80.85727	0	130.9372
6	CATI	PA	0.257686	16680.14		12.49229	0	11.4301	0		100	15	6		142.4306	0	194.3116	0	336.7423
7	CATI	TX	0.224464	21692.09		12.99326	0	12.66121	0		100	17	4		71.97302	0	177.257	0	249.23
8	CATI	CA	0.350743	35500		17.9046	0	25.47014	0		100	21	8		254.8552	0	687.6937	0	942.5489
9	CATI	GA	0.241209	7412.097		11.00955	0	14.48866	0		100	14	2		43.02866	0	28.97732	0	72.00598
10	CATI	IL	0.221232	20867.52		10.72588	0	18.07062	0		100	21	6		94.3553	0	289.1299	0	383.4852
11	CATI	PA	0.278588	12427.72		11.80654	0	19.28007	0		100	11	4		102.6458	0	154.2405	0	256.8863
12	CATI	TX	0.215754	16522.25		10.54216	0	12.98288	0		100	19	4		83.25297	0	142.8117	0	226.0646
13	CATI	CA	0.28858	16003.49		10.27772	0	17.00064	0		100	21	6		101.944	0	255.0096	0	356.9536
14	CATI	GA	0.20799	7794.485		13.0163	0	12.64138	0		100	22	10		206.1956	0	391.8827	0	598.0783
15	CATI	IL	0.293457	10172.21		11.13025	0	17.76605	0		100	16	7		179.6933	0	248.7246	0	428.4179
16	CATI	PA	0.252153	20802.47		13.87783	0	15.08652	0		100	25	7		173.7783	0	286.6439	0	460.4222
17	CATI	TX	0.279165	9590.499		14.63881	0	10.92258	0		100	23	8		215.6658	0	185.6838	0	401.3495
18	CATI	CA	0.293674	20262.73		14.68246	0	16.37562	0		100	20	9		294.6017	0	360.2637	0	654.8654
19	CATI	GA	0.206894	17643.71		10.96332	0	15.08591	0		100	21	6		150.5965	0	256.4605	0	407.057
20	CATI	IL	0.221473	9291.513		12.05695	0	10.67019	0		100	14	4		68.22779	0	117.372	0	185.5998
21	CATI	PA	0.252184	12730.9		11.00192	0	13.82275	0		100	24	8		172.0231	0	304.1005	0	476.1235
22	CATI	TX	0.396243	38467.83		19.6495	0	25.1851	0		100	14	9		300.4434	0	705.1827	0	1005.626
23	CATI	CA	0.230918	10050.52		14.24253	0	16.66524	0		100	14	4		119.6977	0	133.3219	0	253.0197
24	CATI	GA	0.239027	11896.36		14.5024	0	12.1897	0		100	15	3		87.51198	0	97.51762	0	185.0296
25	CATI	IL	0.321638	39356.92		16.58788	0	20.14496	0		100	17	4		119.5273	0	201.4496	0	320.9769
26	CATI	PA	0.214582	11624.65		12.07007	0	15.24644	0		100	21	5		97.42042	0	152.4644	0	249.8848
27	CATI	TX	0.34351	25263.66		18.61919	0	28.53511	0		100	21	8		300.6687	0	542.1671	0	842.8358
28	CATI	CA	0.393814	31265.41		16.59612	0	21.87658	0		100	23	8		222.5573	0	546.9146	0	769.4719
29	CATI	GA	0.299332	12145.52		12.26737	0	18.98373	0		100	17	8		174.9411	0	227.8048	0	402.7459
30	CATI	IL	0.229673	12048.48		11.07105	0	19.83764	0		100	22	4		86.42628	0	297.5646	0	383.9909
31	CATI	PA	0.34294	32689.53		16.36219	0	20.73244	0		100	17	7		198.6219	0	228.0569	0	426.6788
32	CATI	TX	0.230952	17480.58		13.48949	0	19.20133	0		100	14	2		36.97897	0	38.40266	0	75.38163
33	CATI	CA	0.280435	10595.51		14.02661	0	13.10526	0		100	13	6		114.1597	0	144.1578	0	258.3175
34	CATI	GA	0.268255	13261.73		13.62407	0	17.42241	0		100	11	5		106.7444	0	226.4913	0	333.2357
35	CATI	IL	0.264541	10548.3		12.38533	0	11.94861	0		100	15	5		99.31196	0	215.0749	0	314.3869
36	CATI	PA	0.2966	16768.79		12.03589	0	19.58098	0		100	13	5		109.2512	0	293.7147	0	402.9659
37	CATI	TX	0.291528	5627.623		12.55715	0	19.22361	0		100	18	6		117.9	0	422.9194	0	540.8194
38	CATI	CA	0.228999	13999.76		10.54109	0	10.98758	0		100	17	2		41.62328	0	65.92547	0	107.5488
39	CATI	GA	0.360024	25300.38		15.39888	0	25.42497	0		100	27	12		352.5787	0	686.4742	0	1039.053
40	CATI	IL	0.345369	30173.33		15.84964	0	24.34309	0		100	16	4		99.24818	0	292.1171	0	391.3652

Figure 7: Excerpt from the interviewer output file.

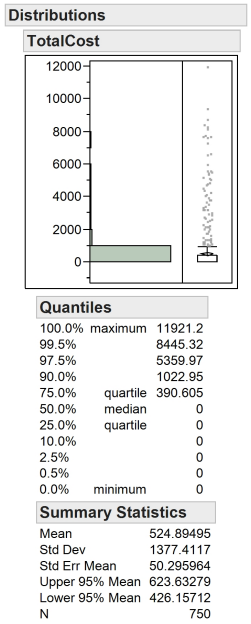

Figure 8: Histogram of total costs incurred by CAPI and CATI interviewers.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Location	PopCount	SampCou	WEBResp	CATIResp	CAPIResp	TotalResp	WtPopCount		EditImput	MVImputations		HDFrPop5	HDFrPop	KLNumSu	KLNumSu	KLNumSu	HDCatSur	HDCatSur	HDCatRespPopResp	Contact	Unit	Person	Incentive	OutOfLoc	Edit	Total
2	Overall	100000	5000	759	1438	1143	3340	100000		1651	1637		0.030542	0.069603	0.001667	0.002957	0.014565	0.004972	0.006644	2.581055	301560	104148.9	186992.9	458655	1510.54	0	0	1052467
3	AL	1543	81	19	26	8	53	1576.123		20	31		0.605029	0.657552	0.101343	0.204185	0.112683	0.314894	0.500039	4.777428	4331.81	788.25	2013.58	6710	0	0	0	13843.64
4	AK	230	8	2	3	1	6	180.5003		9	3		0.795987	0.821964	0.557052	0.544952	0.431916	2.457389	20.18369	-1.#IND00	373.32	94.7	264.27	575	0	0	0	1307.29
5	AR	976	40	6	15	12	33	992.3646		18	11		0.670472	0.698824	0.123407	0.163019	0.197286	0.2182	0.59253	4.935609	2325.64	965.87	1561.64	3205	0	0	0	8058.16
6	AZ	2112	97	17	27	16	60	1796.2		32	30		0.57538	0.671117	0.106281	0.213616	0.135277	0.128077	0.2539	3.228057	5609.99	1485.22	2851.99	8955	0	0	0	18902.2
7	CA	12039	601	81	164	136	381	11406.33		169	174		0.249409	0.377012	0.036025	0.043139	0.050489	0.037936	0.053182	2.162404	37052.8	13642.75	21542.8	57360	0	0	0	129598.4
8	CO	1670	78	8	20	16	44	1320.175		25	15		0.610963	0.68784	0.115844	0.159689	0.125563	0.140702	0.407672	2.800191	5077.59	1375.87	2419.96	8020	0	0	0	16893.42
9	CT	1119	52	9	15	10	34	1024.973		11	23		0.669142	0.744399	0.164994	0.170452	0.196111	0.295326	0.563218	2.686186	3065.93	1099.42	1853.18	4670	0	0	0	10688.53
10	DC	192	15	3	3	5	11	328.0558		13	1		0.686311	0.743221	0.17497	0.186035	0.236967	1.566863	2.935238	7.904821	920.04	613.43	796.57	1395	864.69	0	0	4589.72
11	DE	318	18	3	5	0	8	242.95		4	2		0.738476	0.814953	0.214309	0.22211	0.250079	1.207893	5.14009	10.34528	1187.39	30	332.13	1820	0	0	0	3369.52
12	FL	6100	363	50	104	115	269	8077.481		100	138		0.348395	0.457956	0.037715	0.065321	0.048474	0.052984	0.082132	3.378072	22753.24	9770.49	16931.12	32455	0	0	0	81909.84
13	GA	3021	162	16	37	54	107	3216.602		49	48		0.496558	0.599423	0.073203	0.109147	0.09583	0.094744	0.12237	2.900735	11302.46	4964.6	7217.58	16665	0	0	0	40149.65
14	HI	434	12	2	5	1	8	235.3004		5	5		0.802417	0.861356	0.357085	0.344798	0.311113	1.770028	2.970401	7.827796	607.4	91.38	363.65	975	0	0	0	2037.42
15	IA	980	44	3	16	14	33	991.9741		10	19		0.646262	0.72079	0.308828	0.201217	0.262545	0.338352	0.376108	1.713845	2607.32	1012.6	1990.94	3935	0	0	0	9545.86
16	ID	499	23	2	10	2	14	426.0405		4	6		0.743714	0.810407	0.268224	0.275782	0.205182	0.867735	1.296091	2.307892	1264.67	166.45	664.79	1975	0	0	0	4070.91
17	IL	4105	186	36	55	31	122	3640.791		71	53		0.454277	0.567308	0.07677	0.108555	0.079705	0.088544	0.088697	4.386236	10470.74	3061.16	5656.31	16425	0	0	0	35613.21
18	IN	2033	103	10	31	33	74	2205.882		55	32		0.554847	0.59985	0.102293	0.173556	0.128953	0.296091	0.300102	3.408807	7586.92	2823.49	4461.82	9980	0	0	0	24852.24
19	KS	873	38	3	9	7	19	565.0421		8	11		0.70621	0.788536	0.125077	0.301374	0.332653	0.459526	0.700976	2.82678	2609.97	522.96	926.46	4200	0	0	0	8259.4
20	KY	1418	76	9	14	17	40	1209.838		19	22		0.618965	0.721989	0.148591	0.140728	0.206718	0.26265	0.604464	3.322123	5068.93	1291.69	2085.93	8145	0	0	0	16591.55
21	LA	1420	71	11	18	13	42	1264.379		25	27		0.633047	0.705092	0.104168	0.151346	0.135244	0.251725	0.459193	3.756143	4544.87	1061.71	2380.36	6870	0	0	0	14856.94
22	MA	2052	113	24	23	25	72	2164.508		29	44		0.56482	0.63607	0.141927	0.226832	0.123827	0.173399	0.259955	2.38021	6646.66	2149.49	3517.33	10190	0	0	0	22503.49
23	MD	1911	106	20	29	27	76	2263.679		45	33		0.595986	0.631595	0.107882	0.13082	0.102871	0.230845	0.456458	3.823516	6105.1	2804.86	4505.71	9405	0	0	0	22820.68
24	ME	459	23	3	5	7	15	456.2613		4	9		0.726031	0.780245	0.268427	0.354975	0.322241	0.885462	1.012152	2.34524	1452.39	700.62	1141.11	2255	0	0	0	5549.12
25	MI	3147	163	22	49	50	121	3608.647		74	58		0.49085	0.554638	0.066724	0.103078	0.183245	0.108843	0.132276	3.967217	10331.92	4111.49	7040.07	14400	0	0	0	35883.47
26	MN	1646	100	16	24	23	63	1888.863		24	32		0.580138	0.654097	0.127155	0.159968	0.116746	0.059602	0.099016	2.449062	5900.87	2112.8	3754.5	9740	0	0	0	21508.17
27	MO	1908	94	14	28	27	69	2066.985		23	31		0.59257	0.661755	0.053938	0.097287	0.164336	0.125141	0.267498	2.554679	5088.79	2471.7	4348.32	8290	0	0	0	20198.8
28	MS	957	50	8	14	9	31	923.7341		9	14		0.671099	0.719054	0.165609	0.216178	0.218947	0.161524	0.532356	3.616906	2973.29	723.72	1508.32	4540	0	0	0	9745.33
29	MT	293	17	3	7	3	13	388.2111		11	7		0.716403	0.760661	0.206764	0.230183	0.289986	1.354111	2.259916	4.784829	860.78	212.43	445.96	1325	267.65	0	0	3111.81
30	NE	637	31	1	10	5	16	479.4817		6	10		0.714947	0.805437	0.201322	0.261498	0.383421	0.502169	1.211937	3.398771	2214.3	395.91	951.17	3430	0	0	0	6991.38
31	NC	3123	154	29	43	41	113	3393.056		71	64		0.504193	0.583635	0.088618	0.115439	0.195194	0.054938	0.069414	3.307652	8943.76	3097.85	5279.47	13330	0	0	0	30651.08
32	ND	225	9	0	3	1	4	120.8053		2	4		0.771908	0.858753	0.318245	0.361632	0.514771	1.685272	23.36649	25.21417	680.31	72.86	229.56	1000	108.64	0	0	2091.37
33	NH	417	16	4	2	3	9	266.3832		2	3		0.758826	0.813287	0.184186	0.412162	0.365027	0.988601	2.753149	2.912657	986.23	278.92	336.95	1510	0	0	0	3112.1
34	NJ	2745	140	26	40	31	97	2887.959		62	38		0.536485	0.619606	0.070327	0.127862	0.134343	0.108185	0.282283	2.885015	8311.35	3456.89	6655.33	12240	0	0	0	30663.57
35	NM	644	38	5	14	8	27	807.2023		13	10		0.668322	0.729693	0.214974	0.241966	0.334736	0.403692	0.631574	1.365005	2406.73	656.87	1693.48	3355	0	0	0	8112.09
36	NV	851	36	4	14	4	22	654.7858		9	11		0.693309	0.753962	0.176489	0.285419	0.369054	0.630361	0.911242	1.327779	2031.85	332.83	853.54	3280	0	0	0	6498.22
37	NY	6206	315	48	102	53	203	6070.526		85	104		0.370071	0.492598	0.052645	0.104149	0.07799	0.058715	0.097126	2.546993	18374.06	5584.54	10895	28895	0	0	0	63748.6
38	OH	3646	196	32	49	31	112	3362.046		69	59		0.489926	0.593735	0.07022	0.138611	0.057984	0.093531	0.152329	3.303334	12739.27	2946.65	5927.45	19415	0	0	0	41028.37
39	OK	1201	58	8	20	3	31	926.2991		8	8		0.66501	0.765213	0.239641	0.150916	0.271781	0.320644	0.806727	2.941434	3596.4	305.19	1599.29	5380	0	0	0	10880.88
40	OR	1239	58	5	19	7	31	921.0257		18	13		0.647487	0.745133	0.13561	0.232177	0.19458	0.340462	1.059938	1.684869	3936.01	539.05	1287.83	5980	0	0	0	11742.89

Figure 9: Excerpt from the location output file.

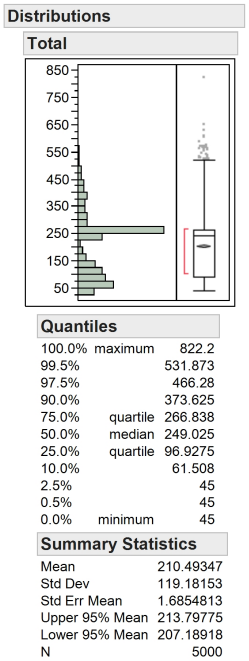

Figure 10: Histogram of total costs incurred by sampled units.

- Counts of the population, sample and respondents by mode.
- Frame variables, including Hellinger distances between the population and the sample and the population and the unit respondents.
- Item nonresponse, edit and imputation counts.
- For numerical survey variables, the population, sample and unit respondent means and covariances, as well as the Horvitz–Thompson estimates, and also Kullback–Liebler divergences (population to sample, to unit respondents and to Horvitz–Thompson estimates).
- For categorical survey variables, one-dimensional marginals for population and unit respondents, as well as Horvitz–Thompson estimates, plus Hellinger distances (population to sample, unit respondents and Horvitz–Thompson estimates.
- Nationwide costs by category.

```
>>> SOFTWARE
WSSMEngine (Version 1.78; 2012/10/13)


>>> FILES
Parameter file: Experiment7-HotDeck-BaseCase.params (written 2012/10/15 15:16:26)
Location file: Locations.csv (written 2012/03/27 18:25:14)
Population file: Population7.csv (written 2012/10/15 14:56:28)
CATI Interviewer file: CATIInterviewersB.csv (written 2012/04/17 19:23:19)
CAPI Interviewer file: CAPIInterviewersB.csv (written 2012/04/17 19:24:27)


>>> SELECTED PARAMETERS
Sample design: SRS
WEB contact attempts: 1
CATI contact attempts: 2
CAPI contact attempts: 3
Numerical survey variable imputation method: HotDeck
Categorical survey variable imputation method: HotDeck


>>> COUNTS
  Population       Sample     WEB Resp    CATI Resp    CAPI Resp   Total Resp    Resp Rate
      100000         5000          750         1485         1137         3372        0.674


>>> FRAME VARIABLES

ONE-DIMENSIONAL MARGINALS: OMITTED TO SAVE SPACE

HELLINGER DISTANCES
Population to Sample: 0.031136
Population to Respondents: 0.061301


>>> SURVEY VARIABLE ITEM NONRESPONSE
    Variable        Count         Rate
      Income          316        0.094
   Education          111        0.033
     Housing          295        0.087
        Food          242        0.072
      Transp          143        0.042
     Medical          147        0.044
        Race          264        0.078
      HhEdAt          127        0.038
      HhEmSt          172        0.051
     Vehicle          171        0.051
     Student          106        0.031


>>> EDITS AND IMPUTATIONS
Flagged Values: 0
Imputations: 2094 for item nonresponse, 0 from edit rules
```

Figure 11: First excerpt from the WSSM results output file for the base case of the experiment described in §1 and §4.

```
>>> NUMERICAL SURVEY VARIABLES

MEANS

    Variable      Income   Education     Housing        Food      Transp     Medical
  POPULATION     6965.50      401.64      848.38      400.94      701.97      530.29
      SAMPLE     6959.97      399.02      845.36      402.88      700.39      531.04
   UNIT RESP     6954.12      404.32      848.24      403.02      704.60      531.15
     H-T EST     6962.26      407.22      846.15      405.04      698.62      538.87


COVARIANCES

  POPULATION      Income   Education     Housing        Food      Transp     Medical
      Income   185355.97     6088.74    11668.02     5681.93     9815.61     7029.12
   Education     6088.74   227214.49   164567.35    98150.72   110017.42    72256.98
     Housing    11668.02   164567.35   186479.55   101825.78   114163.60    84777.69
        Food     5681.93    98150.72   101825.78   107631.29    81173.22    51236.87
      Transp     9815.61   110017.42   114163.60    81173.22   190321.57    57217.07
     Medical     7029.12    72256.98    84777.69    51236.87    57217.07    64458.80

      SAMPLE: OMITTED TO SAVE SPACE

   UNIT RESP: OMITTED TO SAVE SPACE

     H-T EST      Income   Education     Housing        Food      Transp     Medical
      Income   267383.75    12794.13    23929.63    10291.22    12700.56     6944.63
   Education    12794.13   334870.41   171902.59   105148.50   116444.46    73399.33
     Housing    23929.63   171902.59   281897.39   102273.38   108848.76    86577.38
        Food    10291.22   105148.50   102273.38   188182.26    68154.04    54588.29
      Transp    12700.56   116444.46   108848.76    68154.04   280386.11    58680.42
     Medical     6944.63    73399.33    86577.38    54588.29    58680.42   155156.45


KULLBACK-LIEBLER DIVERGENCES

Sample to Population: 0.003166
Respondents to Population: 0.003717
Responses to Population: 2.599617


>>> CATEGORICAL SURVEY VARIABLES

ONE-DIMENSIONAL MARGINALS:
        Race     Category   Population  Respondents      H-T Est
                        0        49943         1556      46109.1
                        1        19855          664      19693.1
                        2        20191          649      19253.5
                        3         5049          230       6813.8
                        4         4962          273       8130.5
      HhEdAt     Category   Population  Respondents      H-T Est
                        0         4890          258       7636.3
                        1        59947         1830      54313.9
                        2        29996         1001      29661.4
                        3         5167          283       8388.4
      HhEmSt     Category   Population  Respondents      H-T Est
                        0        24680          802      23804.6
                        1        25318          875      25957.4
                        2        24983          821      24317.0
                        3        25019          874      25921.0
     Vehicle: OMITTED TO SAVE SPACE
     Student: OMITTED TO SAVE SPACE


HELLINGER DISTANCES
Population to Sample: 0.047243
Population to Respondents: 0.075675
Population to Final: 0.109713


>>> COSTS
       Contact          Unit        Person     Incentive      OutofLoc          Edit         Total
      $298,201      $102,806      $188,458      $454,995          $916           $0    $1,045,376
```

Figure 12: Second excerpt from the WSSM results output file for the base case of the experiment described in §1 and §4.

	Run				
	1	2	3	4	5
Unit Response Rate	0.664	0.663	0.667	0.662	0.669
HD Frame: Pop. to Respondents	0.0595	0.0612	0.0594	0.0684	0.0650
Mean Income: Population	6968.36	6968.36	6968.36	6968.36	6968.36
HT Estimated Income	6961.44	6956.84	6956.12	6968.86	6960.53
KL Num. Survey: Pop. to Sample	0.0025	0.0043	0.0038	0.0037	0.0035
KL Num. Survey: Pop. to Final	2.5870	2.8459	2.5568	2.7950	2.9970
HD Cat. Survey: Pop. to Sample	0.0517	0.0514	0.0482	0.0496	0.0540
HD Cat. Survey: Pop. to Final	0.1110	0.1164	0.1250	0.1138	0.1125
Total Cost	1,040,487	1,032,590	1,054,045	1,039,245	1,050,439

Table 1: Replicate variability for five runs of `WSSMEngine`, all with the same population. Abbreviations: Cat. = Categorical, HD = Hellinger distance, HT = Horvitz–Thompson, KL = Kullback–Liebler divergence, Num. = Numerical, Pop. = Population.

3.7 Replicate Variability

Because WSSM is a stochastic simulator, it is essential to characterize the extent and nature of replicate variability—how much do the results vary when WSSM is run multiple times from exactly the same parameter file? Table 1 provides some insight. To produce it, `WSSMEngine` was run five times on the parameter file `Experiment7-HotDeck-BaseCase.params` (Figure 2), which corresponds to the base case of the experiment in §2 and 4. The table contains the values of selected outputs for each of the five runs. Replicate variability exists in Table 1, but is less dramatic and more manageable than might have been expected.

4 An Illustrative Experiment

We recall from §1 our "experiment:" what are the effects on data quality (measured by Kullback–Liebler divergences and Hellinger distances) and cost of four strategies, as compared to a "base case" corresponding to the parameter file in Figure 2 and to the output in Figures 11 and 12.

- A 10% increase in sample size, operationalized by changing "`SampleSize = 5000`" to "`SampleSize = 5500`" in the parameter file.
- A 10% decrease in measurement error (for numerical survey variables), implemented by reducing by 10% all of the following parameters: `CAPINoiseStdDevMin`, `CAPINoiseStdDevMax`, `CATINoiseStdDevMin`, `CATINoiseStdDevMax`, and `WEBNoiseStdDev`.

- Imposition of edit rules that replace "erroneous" data values by imputed values. The edit rules imposed, in the syntax described in §3.4, are those shown in Figure 4. These edit rules are identical to the constraint rules used to synthesize the population (Figure 3).

- All of the above.

The base case and the four alternatives were run once on the same population file—`Population7-.csv`. Two pairs of interviewer files were used: one[8] for base case measurement error and the other[9] for decreased measurement error.

Table 2 shows the results. The clearest conclusion is that only the reduction in measurement error makes a substantial reduction to the Kullback–Liebler divergence between the population and the final responses for the numerical survey variables, by approximately 20%. By contrast, the increase in sample size has only modest effect on the population–to–respondent and the population–to–final response data quality measures, and it does increase cost, as makes sense, by approximately 10%.

The effect of the edit rules is more subtle. At first glance, they seem to have almost no effect on either the Kullback–Liebler divergence between the population and the final responses for the numerical survey variables or the Hellinger distance between the population and the final responses for the categorical survey variables. However, this sample is one for which the corresponding population–to–respondent distances are especially high. Table 3 contains the ratios of the population–to–final responses data quality measures the to corresponding population–to–unit respondents measures in Table 2. While it is not certain that ratios the the proper means of comparison, when they are used, the edit rules are as effective as decreased measurement error for numerical survey variables and *more* effective than decreased measurement error for categorical survey variables. This not surprising, since measurement error only affects numerical variables.

The "all of the above" strategy is not notably more effective than either the decreased measurement error strategy or the edit rule strategy, except possibly for the categorical survey variables.

Figure 13 highlights the role of measurement error from the perspective of actual and estimated covariance matrices for the numerical survey variables. The matrices for *actual values* of these variables for the population, sample and unit respondents are substantially similar. The matrix labeled `H-T EST` contains Horvitz–Thompson estimates for the finite population covariance matrix, derived from responses that reflect measurement error, item nonresponse, the edit rules and imputation. Most notably, Horvitz–Thompson estimates of variances exceed significantly the true values. This it not surprising, because of the measurement error. Indeed, from Figure 2, measurement error variances (The values in Figure 2 are standard deviations.) have average values of 250,000 for WEB, 62,500 for CAPI and approximately 19,000 for CAPI, the variance inflation is of the order one would expect. Figure 14 confirms this reasoning: it shows populations and Horvitz–Thompson estimated covariance matrices and Kullback–Liebler divergences when mea-

[8] `CATIInterviewersB.csv` and `CAPIInterviewersB.csv`.

[9] `CATIInterviewersB-DecreasedMeasurementError.csv` and `CAPIInterviewersB-Decreased-MeasurementError.csv`.

Measure	Case				
	Base	Sample↑	MeasError↓	EditRules	All
Response Rate	0.674	0.664	0.671	0.654	0.655
HD Frame: Pop to Sample	0.0311	0.0294	0.0301	0.0282	0.0291
HD Frame: Pop to Resp.	0.0613	0.0545	0.0569	0.0624	0.0517
Mean Income: Population	6965.50	6965.50	6965.50	6965.50	6965.50
HT Estimated Income	6962.26	6952.29	6957.12	6959.58	6977.52
KL Num. Survey: Pop. to Resp.	0.0037	0.0022	0.0052	0.0068	0.0035
KL Num. Survey: Pop. to Final	2.5997	2.5895	2.0588	2.6497	1.8485
HD Cat. Survey: Pop. to Resp.	0.0757	0.0729	0.0697	0.0817	0.0680
HD Cat. Survey: Pop. to Final	0.1097	0.1050	0.1139	0.1072	0.0772
Cost	1,045,376	1,168,499	1,053,390	1,047,764	1,154,814

Table 2: Results of the experiment. Abbreviations are the same as in Table 1, with the addition Resp. = Respondents.

Measure	Case				
	Base	Sample↑	MeasError↓	EditRules	All
KL Num. Survey	702.6216	1177.0455	395.9231	389.6618	528.1429
HD Cat. Survey	1.4491	1.4403	1.6341	1.3121	1.1353

Table 3: Ratios of population–to–final responses data quality measures to corresponding population–to–unit respondents measures in Table 2.

surement error is reduced by 90% from the base case values; the improvement is dramatic. On the other hand, the edit rules and hot-deck imputation[10] tend to remove variability from the data.

That a plethora of follow-up experiments can be formulated may already have occurred to the reader. For instance, in this experiment, there was no cost associated with decreased measurement error or edit rules—what if there were, especially for the former? What if the edit rules do not "match" the constraint rules? What if the imputation method were changed? WSSM can produce insight into all of these.

5 Conclusions and Discussion

WSSM is an initial step, and possibly the most salient measure of its success is whether it raises more questions than it answers. NISS plans to release a version for research purposes as soon as feasible, with the goal, *inter alia*, of catalyzing suggestions for new functionality and more detailed

[10]Which resamples from responses that satisfy the edit rules.

POPULATION	Income	Education	Housing	Food	Transp	Medical
Income	185355.97	6088.74	11668.02	5681.93	9815.61	7029.12
Education	6088.74	227214.49	164567.35	98150.72	110017.42	72256.98
Housing	11668.02	164567.35	186479.55	101825.78	114163.60	84777.69
Food	5681.93	98150.72	101825.78	107631.29	81173.22	51236.87
Transp	9815.61	110017.42	114163.60	81173.22	190321.57	57217.07
Medical	7029.12	72256.98	84777.69	51236.87	57217.07	64458.80

SAMPLE	Income	Education	Housing	Food	Transp	Medical
Income	188975.00	5791.62	12778.18	5183.57	8423.95	5856.54
Education	5791.62	223935.76	163445.95	99246.46	112719.19	72653.99
Housing	12778.18	163445.95	183682.96	102631.44	115468.54	84358.99
Food	5183.57	99246.46	102631.44	107034.69	82093.41	51400.09
Transp	8423.95	112719.19	115468.54	82093.41	193730.98	58317.77
Medical	5856.54	72653.99	84358.99	51400.09	58317.77	64397.93

UNIT RESP	Income	Education	Housing	Food	Transp	Medical
Income	188626.63	8009.35	15347.58	7671.71	9981.91	7828.95
Education	8009.35	232189.08	167857.06	101992.91	114990.74	72913.05
Housing	15347.58	167857.06	185167.79	103911.63	115098.15	84086.22
Food	7671.71	101992.91	103911.63	109097.65	83945.25	51094.94
Transp	9981.91	114990.74	115098.15	83945.25	193907.80	58340.53
Medical	7828.95	72913.05	84086.22	51094.94	58340.53	64637.61

H-T EST	Income	Education	Housing	Food	Transp	Medical
Income	259222.37	8829.31	13765.19	11962.97	9541.88	4130.38
Education	8829.31	303304.87	160330.29	103465.76	110976.96	74145.52
Housing	13765.19	160330.29	250115.44	104368.98	106606.97	84554.30
Food	11962.97	103465.76	104368.98	177420.02	78879.92	49963.64
Transp	9541.88	110976.96	106606.97	78879.92	252958.25	53433.30
Medical	4130.38	74145.52	84554.30	49963.64	53433.30	136407.53

Figure 13: Actual and estimated covariances for the “all of the above” strategy in the experiment.

COVARIANCES

POPULATION	Income	Education	Housing	Food	Transp	Medical
Income	185606.86	4399.78	9893.67	4699.38	8916.46	6292.32
Education	4399.78	227909.90	163853.07	97658.64	108754.19	71701.31
Housing	9893.67	163853.07	185468.38	101264.38	113686.77	84100.57
Food	4699.38	97658.64	101264.38	107736.80	80507.30	50831.89
Transp	8916.46	108754.19	113686.77	80507.30	188544.91	56731.28
Medical	6292.32	71701.31	84100.57	50831.89	56731.28	63984.73

[...]

H-T EST	Income	Education	Housing	Food	Transp	Medical
Income	180259.30	4286.91	5301.14	46.77	3565.06	4455.89
Education	4286.91	224122.41	160788.48	95796.49	106854.71	71586.49
Housing	5301.14	160788.48	183528.91	99606.66	110794.56	84677.87
Food	46.77	95796.49	99606.66	107092.91	78358.88	50418.99
Transp	3565.06	106854.71	110794.56	78358.88	189797.38	55917.03
Medical	4455.89	71586.49	84677.87	50418.99	55917.03	66922.00

KULLBACK-LIEBLER DIVERGENCES

Sample to Population: 0.002736
Respondents to Population: 0.003252
Responses to Population: 0.004684

Figure 14: Estimated covariances and Kullback–Liebler divergences when measurement error is reduced by 90% from base case values.

modeling of particular aspects of the survey process.

We highlight three modeling issues. To us, the most glaring shortcoming is that WSSM lacks true dynamics, and is therefore incapable of representing adaptive (Wagner, 2008) or responsive (Groves and Heeringa, 2006) designs. Also, WSSM does not include any statistical disclosure limitation (SDL), although adding at least some forms is on the list of planned, short-term modifications. Additive noise would be especially straightforward, and is attractive conceptually because it is in effect deliberate—as opposed to uncontrollable—measurement error.[11] If SDL is added, measures of disclosure risk are also necessary (Cox et al., 2011). Finally, the current treatment of costs in WSSM is too simplistic, both with respect to nature of the costs and the well-known lack of credible (in some instances, any) cost paradata (Groves, 2004a; Karr and Last, 2006).

We conclude with some thoughts about three central questions. First, what are the uses of any survey microsimulator? We believe that three uses are promising: *education*—WSSM would be a valuable component of any course on survey methods; *evaluation of theory and methodology*—for instance, WSSM could be modified to incorporate Bayesian methods for imputation; and *planning*—the kinds of analyses associated with the experiment in §4 are of most value when done prospectively. For instance, we plan to use WSSM to evaluate the as-yet unproven concept of TSE-aware SDL: given what is known about other sources of error, can SDL be targeted to reduce risk substantially but reduce quality only incrementally?

Other uses are much more challenging. One of these is operational decision-making. In part because it lacks dynamics, WSSM cannot plausibly model operational decisions such as assignment of interviewers to cases on the basis of propensity-to-respond (Groves, 2004b) or on the basis of detailed geography. Other key questions tied to operational decision making are cost modeling, quantifying interviewer effects, and assessing the sensitivity of decisions to small changes in survey operations or conditions.

And, of course, can a survey microsimulator ever be trusted enough to really support informed cost–data quality (or, as articulated in Karr (2012), cost–decision quality) tradeoffs?

The second question is whether WSSM or any other survey microsimulator scales to real problems, where populations are of the order 10^8 or 10^9 and sample sizes of the order to 10^6. We believe that the answer is yes, provided that question is posed as "can be made to scale." For instance, WSSM now loads the entire population into memory, which is not necessary. Nor are the apparent ways to "parallelize" the code exploited. And in the short run, we contend that the sizes in the experiment in §4 are big enough to be insightful. As a point of reference, WSSM in its current version runs on a population of 1,000,000 with a sample size of 20,000 in approximately 3 minutes.

The third question is how—or possibly even whether—WSSM or any other survey microsimulator would be validated. The literature on validation of agent-based models is immature but growing. Some issues have been articulated, and approaches to them have been proposed (Brown et al., 2005; Moss, 2008; Windrum et al., 2007), but it is clear that contextual and situational aspects are dominant. Although the impact of validation focuses on *pre*diction, the path to validation

[11] The authors have argued for some time for inclusion of SDL in the total survey error (TSE) framework.

is *post*diction: can WSSM model past surveys? At this time, it is premature to attempt to answer this question, but essential to keep it in mind.

Acknowledgements

This research was supported by NSF grant SES–1131897 to Duke University and the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank David Banks, John Eltinge, Satkartar Kinney and Jerome Reiter for numerous insightful and challenging discussions.

References

Brown, D., Page, S. E., Riolo, R. L., Zellner, M., and Rand, W. (2005). Path dependence and the validation of agent-based spatial models of land-use. *Internat. J. Geographical Information Sci.*, 19(2):153–174.

Chen, B.-C. (2008). Simulation modeling of field operations in surveys. In *Proceedings of the Survey Research Methods Section*, Alexandria, VA. American Statistical Association.

Chen, B.-C. (2012). Simulating NHIS field operations. *Proc. 2012 Federal Committee on Statistical Methodology (FCSM) Research Conference*. Available online at `http://www.fcsm.gov/12papers/Chen_2012FCSM_II-A.pdf`.

Cox, L. H. (2012). The case for simulation models of federal surveys. *Proc. 2012 Federal Committee on Statistical Methodology (FCSM) Research Conference*. Available online at `http://www.fcsm.gov/12papers/Cox_2012FCSM_II-A.pdf`.

Cox, L. H. (2013). Microsimulation of a government survey. In *Proceedings of the Fourth International Conference on Establishment Surveys–ICES IV*, Alexandra, VA. American Statistical Association. To appear.

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.

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

Groves, R. M. (2004b). Using response propensity models to guide survey administration. Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Groves, R. M. and Heeringa, S. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. Royal Statist. Soc. Series A: Statistics in Society*, 169 (3):439–457.

Horvitz, D. G.; Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 7:663–685.

Karr, A. F. (2011). National Institute of Statistical Sciences Configuration and Data Integration Technical Panel. Final Report NCES 2011-607, National Center for Education Statistics, Washington.

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

Karr, A. F., Cox, L. H., and Kinney, S. K. (2012). The World's Simplest Survey Microsimulator (WSSM). *Proc. 2012 Federal Committee on Statistical Methodology (FCSM) Research Conference*. Available online at `http://www.fcsm.gov/12papers/Karr_2012FCSM_II-A.pdf`.

Karr, A. F., Fulp, W. J., Lin, X., Reiter, J. P., Vera, F., and Young, S. S. (2007). Secure, privacy-preserving analysis of distributed databases. *Technometrics*, 49(3):335–345.

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

Moss, S. (2008). Alternative approaches to the empirical validation of agent-based models. *J. Artificial Societies and Social Simulation*, 11(1):5.

Wagner, J. (2008). *Adaptive survey design to reduce nonresponse bias*. PhD thesis, University of Michigan.

Windrum, P., Fagiolo, G., and Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *J. Artificial Societies and Social Simulation*, 10(2):8.