The Case for Simulation Models of Federal Surveys

Lawrence H. Cox
Assistant Director for Official Statistics
National Institute of Statistical Sciences
Durham, NC 27709
cox@niss.org

Introduction

This paper describes benefits to survey methodology, operations and cost control offered by the development of simulation models of survey processes. Simulation models might be focused on operational elements of a survey such as survey costs, field operations including interviewer behavior and effects, or nonresponse, or on statistical elements of a survey such as design, estimation and incorporation of ancillary/exogenous information. Conceivably, the model might combine and integrate many such elements, might operate at both large and small scales, might be agent-based, and might incorporate interactions between agents such as interviewers and subjects and between components such as design and cost.

With the broad objective to initiate understanding and discussion of these possibilities within the survey community, the National Institute of Statistical Sciences (NISS) organized the April 2011 NISS Survey Simulation Workshop under the auspices of the NISS Affiliates Program. A brief summary of the Workshop is provided below. The remainder of the paper deals with simulation, simulation in a survey environment, and practical issues related to development of a prototype survey simulation model.

NISS Survey Simulation Workshop

With the cooperation of the Bureau of Labor Statistics, the National Institute of Statistical Sciences, through its Affiliates Program, sponsored the NISS Workshop on Microsimulation Models for Surveys. The Workshop was held April 7, 2011 at the Bureau of Labor Statistics, Washington, DC, with a follow-on half day April 8 of evaluation and planning involving only Workshop principals.

Simulation models have been widely used in science, engineering and medicine for decades, but have experienced less use in statistical science, despite mutual reliance on randomization and modeling. Consequently, Workshop speakers and topics represent a blend of experiences inside and outside the statistics and survey communities.

The Workshop program is summarized as follows. For Workshop details: http://www.niss.org/events/workshop-microsimulation-models-surveys

- *Overview of Simulation Models and a Simulation Model for NHIS Field Operations and Cost Estimates*
  Bor-Chung Chen, U.S. Department of Transportation
- *Simulation Models to Inform Health Policy*
  Carolyn Rutter, Group Health Research Institute
- *Some Cost-Modeling Topics for Prospective Redesign of the U.S. Consumer Expenditure Surveys*
  John L. Eltinge and Jeffrey M. Gonzalez, U.S. Bureau of Labor Statistics
- *Optimizing and Simulating CATI Call Scheduling*
  M. Hidiroglou, F. Laflamme, H. Choudhry, and Y. Bélanger, Statistics Canada

- *Standard Errors and Uncertainty in Agent-Based Models*
  Gregoriy Bobashev, RTI International
- *Developing a Microsimulation Model of a Federal Survey*
  Ben Klemens, U.S. Census Bureau

Simulation Models: What and Why

We focus on the issues and questions arising from the notion of design, development and deployment of survey simulation models, and not on technical issues or descriptions. From this perspective, we view a simulation model as a computer model of one or more processes or operations, referred to as the *system*. Model outputs are individual states of the *idealized system* represented by the model.

Simulation models are needed when analytical models are unavailable due to complexity or lack of understanding of the process or operation. A simulation model is useful for generating an empirical distribution of (idealized) system states, and for sensitivity analysis. Simulation models support decision making, evaluation or operational control of the system. Computationally, simulation models are driven by mathematical, statistical and logical relationships between appropriate *parameters* and subsystems. Based on initial values or assumptions (typically, distributions) regarding behavior of model parameters, during a single run the model assigns parameter values and computes a *predicted state* of the system. Repeated runs yield an empirical distribution of states conditional on initial values and assumptions regarding (distribution of) parameters.

Statistically, the simplest simulation models assume parameters are mutually independent, whereas complex models incorporate parameter dependencies reflecting relations or interactions between model "agents." A relevant example for survey simulation would be the effects of behavior, training, characteristics, etc. of a particular interviewer on the response propensity of a particular subject to a particular question, or on interview cost. Parameter dependence tracks *model scale*—macro, meso or micro—which roughly equates to reliance on marginal, conditional or joint distributions.

Our ultimate vision is a general purpose Survey Simulation Laboratory, available as a broad-gauge tool and research and development platform for Federal surveys. This vision may or may not be achievable scientifically, may or may not be worthwhile, cost-effective or of high priority to Federal agencies, and may or may not be within the capabilities of our decentralized Federal survey and survey methodology system. Discussion of feasibility, desirability and achievability of this concept—at an appropriate scale-- is the principal goal of this paper.

The key discussion questions this paper is intended to stimulate are:

Survey simulators—worthwhile or not?

- scale—what sort of pilot study would demonstrate (or refute) proof of concept?
- where--which survey(s) and what operations/components to simulate?
- how--to organize and finance such an effort?
- timeframe—is there an operational target or need to be met by this effort?

Arguments for a Survey Simulation Laboratory

Simulation of one or more elements of a Federal survey would enhance decision making and survey management in any of the several areas below. Whether the Laboratory is realized as a collection of separate simulation models addressing individual areas, a linked set of individual models, or a meta-level model incorporating all survey elements from design through data collection and dissemination and cost modeling and control is likely to remain an open question for some time.

Public policy and administration

A Federal survey is a precision instrument whose accuracy, precision, timeliness, relevance and reliability are of critical national importance. Millions of taxpayer dollars are invested into Federal surveys, and millions more and decisions affecting the U.S. population, society, economy and public health rely critically on estimates and inferences derived from surveys. Federal surveys fuel a large portion of our data-driven society.

Federal surveys are constantly evaluated and improved by survey methodologists and cognitive scientists, again at significant cost. However, the process by which decisions regarding fielding and management of surveys, such as

- how and where to cut sample or content in compliance with budgetary recisions
- deciding if a new survey is needed or if existing data might be combined, and how
- performing data combination or use of exogenous information to evaluate or enhance survey estimates

typically are made more on a subjective basis—expert knowledge and experience—than on a data-driven, quantitative basis. As envisioned, a Federal survey simulation laboratory would incorporate survey data, metadata and paradata and quantitative analysis into survey decision making in a statistically principled manner.

A survey provides a perfect candidate for simulation-- a complex large scale operation whose characteristics and behaviors are well-understood individually, poorly understood in combination, and cannot be modelled analytically.

Survey planning and management

Simulation can be central to decision making, management and control in various aspects of a survey, including

- cost modeling and control
- resource (re)allocation
- monitoring and optimizing interviewer performance

A relevant example in the Federal sector is environmental management and control where decisions are based on predicted pollution and, for individual polluters, predicted contributions to pollution based on computer simulation models combining ambient monitoring, self-reported and other data.

Survey statistical laboratory

Effective survey methodologies for sample design and estimation have existed for more than 60 years and continue to be developed. However, survey data are increasingly expensive to collect and increasingly harder to quality-assure and obtain. Demands for small area and small domain estimation will increase, and sampling alone is silent on how to proceed to estimate or predict in subdomains for which sample data are unavailable. Methods for combining information from multiple sources and for using exogenous data to evaluate, calibrate and enhance survey data are needed. Sensitivity analyses may be difficult to design or perform. A survey simulation laboratory, based on models for combining sample and nonsample data, would address questions in the following areas:

- design, redesign and modification of a design, such as to cut sample to meet budget cuts or to continue estimation in areas affected by survey disruption
- evaluation
- integration
- exploration (what if scenarios)
- understanding (nonresponse, total survey error)

Implementation Issues

Staffing

It should be noted that while simulation rests upon randomization, simulation practitioners are found more in the operations research community than in the statistical community. This observation is important because simulation expertise, and more generally operations research expertise, is sparse in Federal survey organizations due in part to their firm rooting in survey methodology. This creates cultural, staffing and operational impediments to developing a survey simulation laboratory that will have to be addressed.

Scope and scale

Issues of funding aside, where to start? The principal questions are as follows.

Which surveys have overriding needs (nonresponse, quality) that make them natural candidates for simulation models?

Which surveys are on timelines that would garner genuine interest in the success of a proof-of-concept experiment?

Which surveys are of sufficiently simple scope to imply a short and cost effective development process?

Which high priority/high value survey areas are "ripe" for simulation?

- cost modeling and control?
- field operations and interviewer behavior?
- nonresponse prediction and adjustment?
- design and redesign?
- combining survey and exogenous information?

Good choices for the operational environment, scope and objectives of a pilot study are important to its success, understanding and acceptance by the Federal survey community.

Sources of funding

What organizations would be interested, can and will fund a proof-of-concept project?

- NSF?
- foundations? which one(s)?
- ICSP?
- coalition of the willing?
- single champion?
- in-kind contributions of staff and facilities