Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

IN SEARCH OF OPTIMAL SURVEY DESIGNS

Paul W. Ludington[1]

ABSTRACT

The Optimal Design Program is an alternative to the current decennial redesign of demographic surveys at the U.S. Census Bureau. The Program seeks to minimize the mean-square error (MSE) of survey estimates by optimizing the selection of annual household samples

Initial research has focused on the use of multi-agent systems (distributed artificial intelligence) to produce optimal annual samples for all demographic surveys. The first multi-agent system optimizes redesign inputs. It represents each household as an autonomous software agent, forecasts historic household data to the current redesign year as probability density functions, and uses the PDFs to simulate current household characteristics that are "optimally" controlled to current survey estimates. The control process must address PDF priorities (in terms of data quality) while minimizing simulation adjustments.

The second multi-agent system selects optimal samples for all demographic surveys Each sample is represented as a software agent. The Bayesian optimization algorithm (BOA) is applied at each design stage to partition the sampling units into sample and non-sample subsets, with no initial stratification of the sampling units required. The samples are "optimal" in that the BOA minimizes a criterion that includes both the mean-square difference of the simulated sample estimates from their population values and the mean-square difference of the sample Bayesian network from its population network.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

KEY WORDS: Optimal Design Program, Multi-Agent Systems, Bayesian Optimization Algorithm, Bayesian Networks, Mean-Square Error

1. INTRODUCTION

The Optimal Design Program addresses three U.S. Census Bureau strategic objectives. The first objective is to improve the efficiency and effectiveness of survey processes. The second objective is to produce accurate and timely statistics by developing new samples that reflect the current characteristics and geographic location of the population. The third objective is to improve the timeliness, accuracy, relevance, accessibility, and cost effectiveness of our surveys and censuses by adopting new applications of technologies and methodologies (U.S. Census Bureau, 2002).

Society has **demographic processes**, like aging and migration, that determine personal characteristics and interactions. These processes are becoming increasingly complex and integrated (such as Internet communication). **Demographic systems** are created to monitor and control demographic processes. Such system (like sample surveys) are becoming increasingly complex, integrated, expensive, automated, and intelligent. One goal is to construct an optimal sampling system. **Demographic data** is produced by demographic systems. This data is increasingly detailed, critical, and confidential (such as data for homeland security). A second goal is to optimize inputs to the survey design process. **Demographic estimates** are computed from demographic data. These estimates must be increasingly detailed (for small areas and domains), integrated, and accurate. The third goal is to select samples that minimize the mean-square error of survey estimates. Demographic applications incorporate accurate demographic estimates. These applications are

[1] U.S. Bureau of the Census, Room 3723-3, Washington, DC 20233

increasingly critical for human survival.

The Optimal Design Program seeks to improve the information input to demographic survey redesigns, the manner in which that information is processed, and the redesign schedule. Current sampling inputs consist primarily of tabulations of census data. The proposed sampling inputs include data from censuses, surveys, and administrative records and involve forecasting, simulation, and tabulation operations. Current sampling frames include the unit, area, group quarters, and new construction frames. The proposed sampling frame is the master address file. The redesign schedule will be annual, instead of decennial, and proposed samples (such as updated primary sampling units) will be implemented only when it is cost-effective to do so. The current estimation areas vary by survey and include the nation or individual regions or states. The proposed estimation area is always the individual state. The current sampling algorithm is stratification and selection. The proposed sampling algorithm, applied at both the primary sampling unit and household stages, is the Bayesian optimization algorithm (BOA). This algorithm minimizes mean-square errors of simulated survey estimates, rather than the current practice of minimizing coefficients of variation on stratification variables. The BOA is a genetic search algorithm, and selects sample units directly, without the need for prior stratification. Current sample weights will reflect the probabilities of selecting optimal samples, rather than the current method of weighting by measures of size and sampling intervals. Variance estimation during sampling is used to compute the mean-square errors of simulated survey estimates, rather than to compute the coefficients of variation on stratification variables.

The Optimal Design Program continuously improves by rapidly incorporating cutting-edge methodologies and technologies, such as multi-agent systems. Such systems use artificial intelligence to enhance the sampling process. Section II overviews multi-agent systems. Section III discusses a multi-agent system for optimizing inputs to the redesign process. Section IV describes a second multi-agent system for selecting optimal samples for all demographic surveys. Section V notes direction for future research in the Optimal Design Program. Table 1 provides motivation for optimal survey design. Table 2 compares optimal design methods with current methodologies. Table 3 outlines the optimal design technologies of multi-agent systems and the Bayesian optimization algorithm.

2. MULTI-AGENT SYSTEMS

A multi-agent system (MAS) is a collection of intelligent and autonomous software agents. An "agent" is a software entity that possesses knowledge and goals and which acts upon its environment to achieve those goals. This section overviews multi-agent systems (Sen , 2000) (Wooldridge, 2002). Agent architectures typically include sensors to input environmental information, effectors to operate upon that environment, a model of other agents in the system, and modules for coordination, learning, planning, and inference. Multi-agent systems are classified by the relationships among their agents. Cooperative agents jointly pursue common goals. Each goal is decomposed into sub-goals that are assigned to individual agents based upon their capabilities and access to resources. Cooperative agents organize into authority hierarchies and use protocols to share information.. Non-cooperative agents are either adversaries or merely indifferent to each other. The multi-agent systems described in this paper feature cooperative agents.

The field of multi-agent systems is related to other areas of artificial intelligence, such as knowledge representation, search, and machine learning. Also, the field utilizes techniques and results from sociology, cognitive science, economics, and management science. Multi-agent systems have four major advantages. First, they are useful for modeling problem domains that are inherently distributed in nature, without a high degree of central control, such as demography. Second, multi-agent systems are modular, with functional agent designs. This is useful in the study of complex demographic phenomena.. Third, multi-agent systems are adaptive, making them more robust than centralized systems for forecasting demographic values for dynamic populations. Fourth, multi-agent systems model the social aspects of intelligent behavior, such as demographic interactions.

The Optimal Design Program has two multi-agent systems. The first system optimizes redesign inputs as described in section 3. The second system optimizes household samples as described in Section 4.

3. OPTIMAL REDESIGN INPUTS

The first multi-agent system for the Optimal Design Program inputs data from surveys, censuses, and administrative records and simulates a demographic "state" for each household in the current master address file. The word "state" in this context refers to a set of values for (potentially many) demographic variables associated with each household. Simulated states are actual data values, if these values are known. Otherwise, they reflect a "best guess" of that data.

An autonomous software agent represents each household, and each household agent is associated with a particular node in a Bayesian network. A Bayesian network is a graphical model consisting of nodes linked together by arcs and having a conditional probability associated with each arc. Thus, each network node is associated with a household agent in a particular demographic state. Each network arc is associated with a household interaction, coded in an agent message, that is passed with a specified probability.

A probability density function (PDF) is estimated for each household agent (in each design replicate) which reflects the probabilities that household occupies alternative demographic states. The PDF is estimated from historic data known about that household and from the probabilities of demographic changes and interactions for similar households. Generating a random number for each household agent (in each replicate) and comparing that random number to cumulative PDF probabilities permits a current demographic state to be simulated for that household.

The values of demographic variables in each state (such as the number of persons in the household), can be aggregated to higher levels of geography. In particular, aggregated values can be controlled during simulation to recent demographic survey estimates. This control is "optimal" in that household agent priorities reflect the quality of data supporting their PDFs, and control adjustments are made first to lower-priority agents. The final simulated household values are tabulated as necessary (such as by PSU) to support the selection of optimal demographic samples in section 4.

4. OPTIMAL HOUSEHOLD SAMPLES

The second multi-agent system for the Optimal Design Program selects annual household samples for all demographic surveys. An autonomous software agent represents each estimation area for each survey. Other agents represent the households in each estimation area. The system inputs the simulated data for all households, along with the number of sample PSUs and sample households in each estimation area. A sample-selection algorithm then selects an optimal set of sample households for each survey and estimation area. Embedded in the sample-selection algorithm is the Bayesian optimization algorithm (Pelikan, 1999).

The BOA is a genetic search algorithm which selects optimal sets of households within optimal sets of PSUs.. At each iteration in a sequence, it constructs a Bayesian network from the simulated data for the current set of sample units. Through probabilistic inference on the network, it identifies non-sample units to exchange with sample units to improve the overall sample. By applying weighting and variance estimation to the sample data at each iteration, it computes the mean-square differences of the simulated sample estimates from their population values. It also computes the mean-square difference of the sample Bayesian network from its population network. Then it minimizes a criterion of these mean-square differences.

The entire sample-selection algorithm consists of the following steps for each survey.

(1) Select an estimation area.

(2) Input the current demographic estimates for that area (EST1) and simulated data for all its households.

(3) Build a Bayesian network for the area (BN1) from its simulated household data.

(4) Randomly select an initial set of sample PSUs.

(5) Build the sample Bayesian network (BN2).

(6) Update the sample PSUs

- Exchange non-sample PSUs for sample PSUs after inference on BN2..

(7) Compute sample estimates (EST2) from household data for the sample PSUs.

(8) Update the sample Bayesian network (BN2).

(9) Compare EST2 to EST1 and BN2 to BN1:

- If not close enough, then return to (6).
- If close enough, then continue

(10) Randomly select initial sample households for all sample PSUs.

(11) Build a sample Bayesian network (BN3) for each sample PSU.

(12) Update sample households for all sample PSUs.

- Exchange non-sample households for sample households after inference on each BN3..

(13) Compute sample estimates (EST3) from sample household data for each sample PSU.

(14) Update the sample Bayesian network (BN3) for each sample PSU.

(15) Compare EST3 to EST2 and BN3 to BN2:

- If not close enough, then return to (12).
- If close enough for final estimation area, then stop.
- If close enough for non-final estimation area, then return to (1)...

Annual samples for demographic surveys should be implemented only when it is cost-effective to do so. This may require retaining the sample PSUs from the preceding year, and selecting only sample households in the current year.

5. DIRECTIONS FOR FUTURE RESEARCH

A paper entitled “Mathematical Analysis and Simulation of Multi-Agent Systems for Optimal Survey Design” will be presented at the Fall 2003 Research Conference of the Federal Committee on Statistical Methodology in Washington, DC.

A paper entitled “Implementing Optimal Sampling Programs in Intelligent Organizations” will model the U.S. Census Bureau as a complex, composite agent that implements the Optimal Design Program and generates optimal samples for all demographic surveys. The field of organizational intelligence integrates artificial and human agents and considers such matters as organizational memory and learning capabilities.

To support the above research, small multi-agent systems will be specified and prototyped to analyze their control algorithms, Web databases, agent architectures, and communication protocols. These prototype systems will also assist evaluation of the benefits and costs of the Optimal Design Program.

Table 1
Motivations for Optimal Survey Design

Demographic Processes

- Determine personal characteristics and interactions
- Increasingly complex and integrated

Demographic Systems

- Monitor and control demographic processes
- Increasingly complex, integrated, expensive, automated, and intelligent
- **Goal:** Construct an optimal sampling system

Demographic Data

- Produced by demographic systems
- Increasingly detailed, critical, and confidential
- **Goal:** Optimize inputs to survey design process

Demographic Estimates

- Computed from demographic data
- Increasingly detailed (small areas and domains), integrated, and accurate
- **Goal:** Select samples that minimize MSE of survey estimates

Demographic Applications

- Incorporate accurate demographic estimates
- Increasingly critical for human survival

Table 2
Methods for Optimal Survey Design

Item	Proposed	Current
1. Redesign schedule	Annual	Decennial
2. Estimation area	State	Nation, region, state
3. Estimates, design variables, sample sizes	Sponsor selected	Sponsor selected
4. Sampling frames	Master address file	Unit, Area, GQ, NC
5. Sampling stages	PSU, household	PSU, household
6. Sampling inputs	Census data, survey data and estimates, administrative records	Census data

7. Input processing	Forecasting, simulation, tabulation	Tabulation
8. Sampling algorithm	Bayesian optimization algorithm (BOA)	Stratification and selection
9. Search criterion	Minimize MSEs on simulated estimates	Minimize CVs on stratification variables
10. Search space	All Bayesian networks	All stratifications
11. Search method	Genetic algorithm	Hill climbing
12. Sampling intervals	Not applicable	Used to sample households
13. Sample weights	Reflect probabilities of selecting optimal samples	Reflect measures of size and sampling intervals
14. Variance estimation	Used to compute MSEs on simulated estimates	Use to compute CVs on stratification variables

Table 3
Technologies for Optimal Survey Design

Multi-Agent Systems

- “Distributed Artificial Intelligence”
- Autonomous software agents share goals, possess knowledge, make decisions, and communicate together.
- Agents observe and modify their external environment
- Used to model distributed problem domains that lack central control (such as demography), complex and dynamic phenomena, and intelligent social behavior

Bayesian Optimization Algorithm (BOA)

- Genetic algorithm developed at the University of Illinois in the late 1990s
- Builds a sequence of Bayesian networks and uses inference on them to improve a “population” of entities.
- Used to model relationships among demographic variables for all surveys
- Used to partition primary sampling units and households into optimal sample and non-sample subsets.

REFERENCES

Pelikan, M., D.E. Goldberg, and E. Cantu-Paz (1999), "BOA: The Bayesian Optimization Algorithm," Illinois Genetic Algorithms Laboratory, Urbana, IL: University of Illinois at Urbana-Champaign

Pelikan, M., D.E. Goldberg, and F. Lobo (1999), "A Survey of Optimization by Building and Using Probabilistic Models," Illinois Genetic Algorithms Laboratory Report No. 99018, Urbana, IL: University of Illinois at Urbana-Champaign

Sen, S. (2000), "Multiagent Systems: Milestones and New Horizons", Tulsa, OK: University of Tulsa.

U.S. Census Bureau (2002), "U.S. Census Bureau Strategic Plan FY 2004-2008", unpublished report, Washington DC:

Wooldridge, M (2002), *An Introduction to MultiAgent Systems*, West Sussex, England: Wiley.