OPTIMAL PERIODICITY OF A SURVEY: EXTENSIONS OF PROBABLE-ERROR MODELS

Wray Smith, Dhiren Ghosh, and Michael Chang, Synectics for Management Decisions
Wray Smith, Synectics for Management Decisions, 3030 Clarendon Blvd #305, Arlington VA 22201

KEY WORDS: Absolute error modeling, Data deterioration, Fixed-and-variable costs, Repeated surveys, Sampling designs

This paper extends prior work on the problem of choosing optimal periodicity (and associated sample sizes) for repeated surveys of public and private schools with joint consideration of data deterioration (resulting from unobserved year-to-year changes in the underlying process variables), sampling error, and cost. The family of "probable-error models" that was first described in Ghosh *et al.* (1994) has been extended and empirical results obtained for state-level as well as national-level estimates using data from three rounds of the Schools and Staffing Survey (SASS). As noted in the 1994 paper, the models provide "a direct approximate method for characterizing the decision problem of making a joint choice of inter-survey intervals and sample sizes under a fixed cost constraint." The extensions reviewed in the present paper assume, for the most part, that conventional direct estimation methods will be used by the data user. In the case of a proposed alternative sampling design suggested by the modeling results, the data user may wish to consider the use of an indirect estimation (time series modeling) approach along the lines discussed in Smith *et al.* (1995).

SASS was conducted at three-year intervals for school years 1987-88, 1990-91, and 1993-94. Future rounds may be conducted at intersurvey intervals of 4, 5, or 6 years. The modeling extensions are illustrated here in a review of two of several new models that were formulated as modifications of the earlier models. The two models provide alternative formulations to account for the approximate average errors incurred by a data user within successive 12-month periods following a SASS data collection and up to the time of the next data collection. Projected absolute errors have been estimated for future national-level and typical state-level data collections for selected policy variables and a range of fixed-to-variable cost ratios for each possible periodicity.

The two illustrative models, denoted as Model 3A and Model 4M, are modifications of Model 3 and Model 4 of the 1994 paper. They combine a sampling absolute error *(s.a.e.)* and a process shift D over time in different ways to obtain, for different periodicities, estimates of the year-by-year projected absolute errors that would be incurred by a data user as well as average projected absolute errors for each multi-year periodicity. Annual dollar resources for SASS are assumed to be fixed. For each of several scenarios this assumption constrains the total annualized cost to a fixed amount and hence determines the sample size for each combination of a periodicity (4, 5, or 6 years) and a fixed-to-variable cost ratio.

We assume that data users will keep on using the data obtained from the most recent past survey until a new survey is undertaken and the newly collected data are processed and released to data users. Thus, if the inter-survey period is long, "deterioration" of the data could affect the quality of decisions made by users. On the other hand, if the survey is undertaken very frequently, the costs of conducting the survey and analyzing the data and the indirect costs of the response burden may be judged to have costs that exceed the benefits achieved in using fresh data. In the context of repeated surveys, it is useful to distinguish both opportunities and problems presented by different designs.

Typical analyses of cost-benefit tradeoffs tend to focus on the best use of a fixed resource amount over a time period that would include two or more survey data collections. The present budgetary restrictions for the 1990s are such that the "fixed" resource amount may be arbitrarily depressed and may overconstrain any realistic formulation of the optimization problem. In fact, the "truly optimal" formulation may be precluded by external constraints.

The usual cost model for a sample survey assumes a start-up cost C_0 and a per unit (ultimate sample unit) cost C_1. Thus, the total cost is represented as $C = C_0 + nC_1$. However, the start-up cost may depend on the periodicity. If so we represent the startu-up cost as $C_{0,k}$ (where k is the periodicity), which may be regarded as increasing with increasing periodicity; that is, the start-up cost may be more if the periodicity is five years compared to the start-up cost for a periodicity of four years and so on. On the other hand, the start-up cost may be considered to be constant; that is, it may not depend on the periodicity of the survey. Further details are given in Ghosh *et al.* (1994).

We assume that the true value of a variable remains constant for a year after the survey date. This is an appropriate assumption for the SASS survey system since nearly all of the observed variables under the various SASS questionnaires have an annual

accounting period and the SASS data user is interested in changes in variables which are specified to change as of some conventional time point. For example, the official figure for enrollment and number of teachers in a public school is the enrollment "on or about October 1" of the school year. The corresponding number of teachers or the full-time equivalent (FTE) number of teachers are counted at about the same point in time. The student enrollment and the teacher count may fluctuate during the academic year, but SASS and the Common Core of Data (CCD) are, in effect, taking snapshots at the same time over a sequence of years. The error committed in using a survey estimate is exactly equal to the difference between the survey estimate and the true value. Within the first twelve-month interval from the survey date any user incurs an error which equals the difference between the true value and the survey estimate. The estimated standard error of the survey estimate provides an indication of this difference.

If one were interested in estimating from SASS data for a survey year the mean of some characteristic for a specified group of schools, such as the average "number of K-12 teachers that are new to the school this year" for all regular public elementary schools in the state of California, then the estimate would be constructed by applying the school weight for each school to the reported number of new teachers for that school, summing the products and dividing by the sum of the school weights. For some of the SASS-based public school statistics published by NCES, such as those in the *Statistical Profiles* for each round of SASS, the NCES publications include tables of state-by-state estimates of the statistics and, for a selected subset of of the state-by-state statistics, they also include tables of the estimated standard errors for those statistics. For example, the publication *Schools and Staffing in the United States: A Statistical Profile, 1990-91* includes for public schools both estimates of the statistics and estimated standard errors for these statistics on a state-by-state basis for (1) Number of public schools and students and average number of students per full-time-equivalent (FTE) teacher, (2) Percentage distribution of public school teachers by sex and race-ethnicity, percent minority teachers, and average teacher age, and (3) Average base salary for full-time public school teachers and average public school principal salary. As stated in the technical notes to that publication, "Standard errors were estimated using a balanced repeated replications procedure that incorporates the design features of this complex survey."

The difference between the true value and the survey estimate is the deviation from the mean m in the normal distribution of the survey estimate x considered as random variables. We denote the average of the absolute deviations as the "sampling absolute error" or *(s.a.e.)*. Assuming a normal distribution, the projected absolute error incurred by a user during the first year after the survey is $0.8\ s\ /\ \mathrm{sqrt}(n)$ where $s\ /\ \mathrm{sqrt}(n)$ is the standard error of the estimate, assuming simple random sampling. At the end of each year we assume that the true value undergoes a change. The magnitude of this change at the end of each year is denoted $|\ D\ |$. The sampling error component is $0.8\ s\ /\ \mathrm{sqrt}(n)$. Thus the expected value of the total error committed by a data user is dependent on *(s.a.e.)* and on $|\ D\ |$. The magnitude of the change at the end of the second year is also $|\ D\ |$, and so on.

In Model 3A, which is a variant of the Model 3 described in Ghosh *et al.* (1994), we assume the year-to-year process disturbance (process error) to be a normal variable with a zero mean. (If needed, a process error with a nonzero mean could be incorporated into the analysis framework.) Since the process error and the sampling error are both assumed to be normally distributed, they can be readily combined. The projected absolute error is then a linear combination of the process absolute error and the sampling absolute error.

In Model 4M, we explicitly assume that the process change which occurs each year (for example, every October) occurs in accordance with a Random Walk process in discrete time. That is,

$$x_t = x_{t-1} + w_t$$

where w_t has mean zero. We then calculate the average error for different possible periodicities of the repeated survey. The optimal intersurvey interval can be determined if the process variance and the sampling variance are known. In a Random Walk model, the current level of the process is the best current forecast for any future year. One assumes that any known trend component has already been subtracted out. In general, data users will typically use the last available survey value as long as no new survey has been conducted. This assumption concerning user behavior is consistent with our assumption of an underlying Random Walk process.

As noted above, Model 3A is a variant of Model 3 of Ghosh *et al.* (1994). The new Model 4M is a modification and replacement for Model 4 of that 1994 paper. In the original Model 4 we introduced the concept of a loss parameter that converted the sampling error together with the unobserved process shift in non-survey years to a loss expressed in

monetary units. The combination of average cost and average error over a period of years was minimized to determine the optimum periodicity. This was a variation on an approach in Smith (1980) and also on an analysis suggested by S. Kaufman. Model 4M, however, does away with the need for a separate loss parameter, thus avoiding the introduction of a subjective judgment on the part of the survey administrator.

The following table sets forth the year-by-year evolution of the projected absolute errors for the two models. In Model 3A the evolution is based on $|D|$, the magnitude of the annual change in the true value, and the sampling absolute error, *(s.a.e.)*. For Model 4M, the evolution is based on D^2, which is proportional to the variance of the process disturbance, and the sampling absolute error, *(s.a.e.)*. The *(s.a.e.)* depends on the sample size which, in turn, depends on the chosen periodicity under the constraint of fixed annualized cost.

Projected Absolute Errors for Selected Models

Year	Model 3A	Model 4M
1	$(s.a.e.)$	$(s.a.e.)$
2	$0.8\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
3	$0.8\sqrt{2}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{2D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
4	$0.8\sqrt{3}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{3D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
5	$0.8\sqrt{4}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{4D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
6	$0.8\sqrt{5}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{5D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
Avg	$\frac{0.8}{6}\sum_{i=1}^{6}\sqrt{i-1}\,\lvert D\rvert+(s.a.e.)$	$\frac{0.8}{6}\sum_{i=1}^{6}\sqrt{(i-1)D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$

We applied the models described above using three rounds of SASS data at the national level (U.S.) and at the state level for three selected states (California, Iowa, and New York). The following twelve variables were selected from the School, Administrator, and Teacher questionnaires:

Item 1. Number of students served by chapter 1 services (Schools--public).
Item 4. Number of K-12 teachers that are new to the school this year (Schools--public).
Item 6. Percentage of all schools with minority principals (Adminr--public and private).
Item 7A. Number of students per FTE teacher, by sector (Schools--public).
Item 7B. Number of students per FTE teacher, by sector (Schools--private).
Item 8A. Percentage of schools in which various programs and services were available (Schools--public).
Item 8B. Percentage of schools in which various programs and services were available (Schools--private).
Item 9. Percentage of principals having master's degree (Administrator--public).
Item 10A. Percentage of full time teachers who received various types of compensation (Teacher--public).
Item 10B. Percentage of full time teachers who received various types of compensation (Teacher--private).
Item 11A. Percentage of full time teachers newly hired and were first time teachers (Teacher--public).
Item 11B. Percentage of full time teachers newly hired and who were first time teachers (Teacher--private).

Private school items 7B, 8B, 10B, and 11B were omitted from the state-level computer runs since state-level estimates are not published by NCES for private schools. Item 6, which is based on pooled data for public and private schools combined, was retained in all runs.

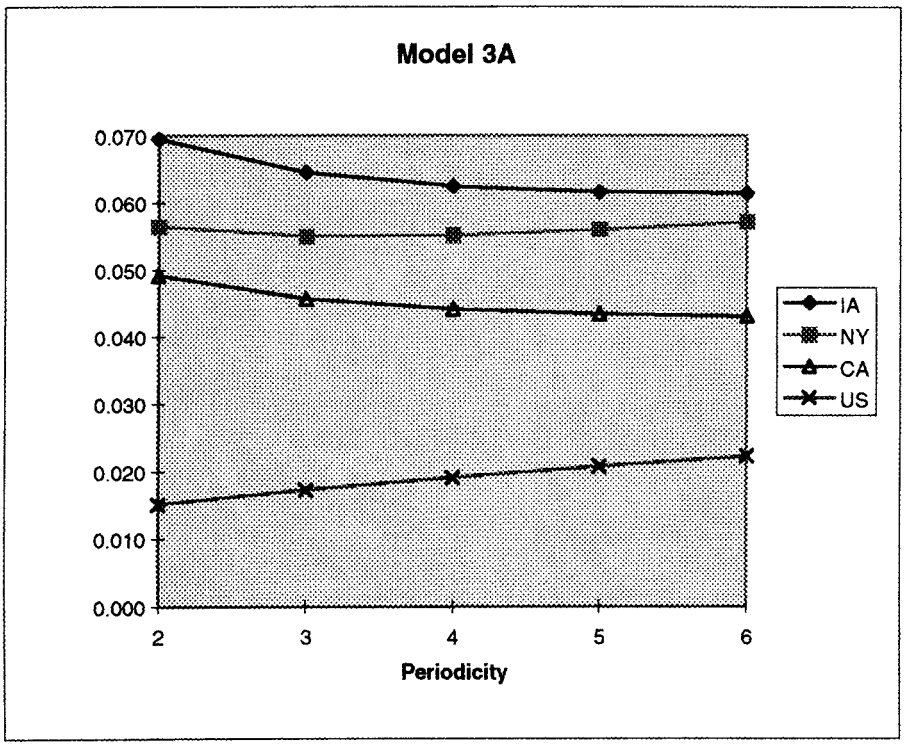
We obtained approximate estimates for the fixed cost and variable cost elements of SASS. We applied the two models for each variable listed above, and computed the projected absolute error for periodicities of four, five, and six years and for specified scenarios of fixed-to-variable cost. The accompanying graphs for Model 3A and Model 4M, respectively, show the average *rel p.a.e.* (where the *rel p.a.e.* for each variable is its *p.a.e.* divided by its mean) for Iowa, New York, California, and the U.S. for a set of eight policy variables with a fixed total cost and a fixed-to-variable cost ratio of 50:50. We see that for the U.S. as a whole, shorter periodicities (even with their smaller sample sizes) result in smaller relative projected absolute errors. For California and Iowa, under both Model 3A and Model 4M, the averages of the relative projected absolute errors are larger for short periodicities and smaller for longer periodicities. For New York, the mean values of the *rel p.a.e.* are essentially flat under Model 3A over the periodicity range from 2 to 6 years but under Model 4M the values decline initially, with a minimum at a periodicity of 4 years, and then rise slightly for periodicities of 5 and 6 years.

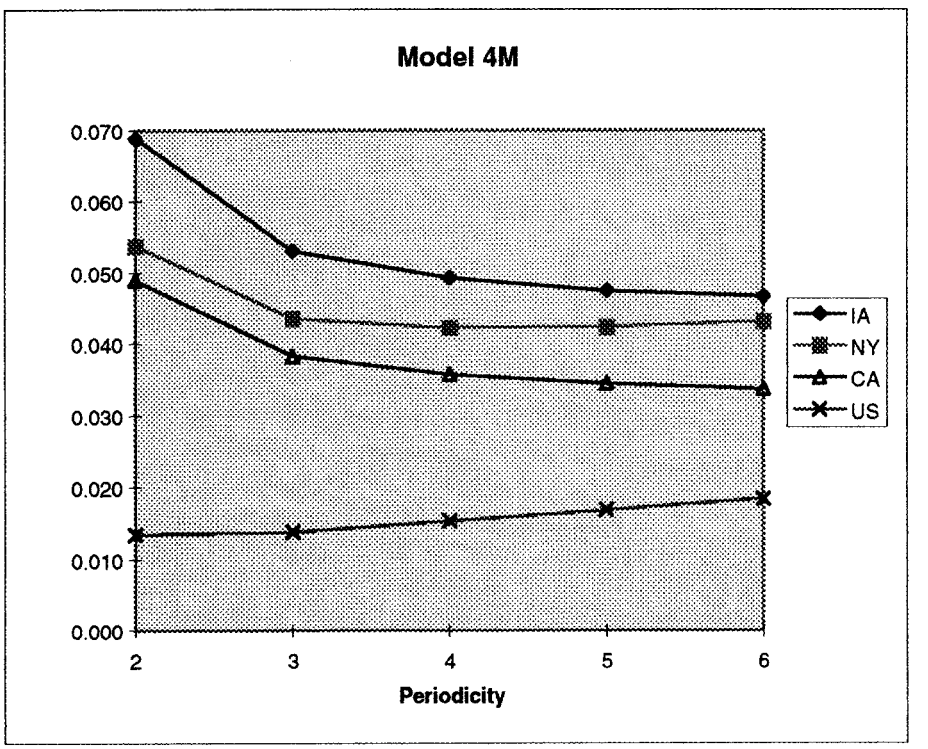
Under the probable-error models the data users who are primarily interested in carrying out analyses for individual States will generally incur smaller errors if they are provided with datasets from longer periodicities and hence larger sample sizes. Data users who are primarily interested in carrying out analyses for the U.S. as a whole will incur smaller errors if they are provided with datasets from shorter periodicities and correspondingly smaller sample sizes.

These observations have led to an alternating large-and-small-sample scenario which was formulated as follows: Assume the same fixed annualized resource budget that would otherwise support the large-sample scenarios with a periodicity of five years over a range of cost ratios (0.3, 0.4, 0.5, 0.6, or 0.7). Then the assumed sample sizes for the U.S. at the mid-point cost ratio p=0.5 were 9,000 public schools and 48,000 teachers. These sizes were proportionally smaller or larger for smaller or larger cost ratios. Assign these sample sizes to a periodicity of six years instead of to a periodicity of five years. This results in a "cost dividend" of 20 per cent which can be invested in a one-fifth U.S. sample of 1,800 public schools and 9,600 teachers for a data collection which can be conducted at the halfway point between two full-sample data collections; namely, three years after the previous large data collection. Assume that there is no processing delay. For simplicity, assume that the schools in the one-fifth sample are nonoverlapping with the schools in the full sample. Further assume that for the U.S. as whole only direct estimates will be of interest and, hence, the two independent samples (the full sample and the one-fifth sample) will be treated as independent cross-section surveys three years apart.

Now consider the projected absolute errors that will be incurred by a data user over a six-year period.

Average *rel p.a.e.* for Iowa, New York, California, and U.S. for Eight Policy Variables with Fixed Total Cost and p=0.5





During the first, second, and third years after a full-sample data collection a data user who is interested in national data will continue to use that sample. Under either Model 3A or Model 4M the projected absolute errors will increase each year. In the fourth year after the large sample data collection a new dataset from the one-fifth size national sample will become available. The data user then disregards the data in the old large sample and begins to use the data from the new one-fifth sample and continues to use it until data from the next full sample becomes available in the seventh year. For the U.S., the sample sizes in the one-fifth national sample are large enough that quite good estimates may be made. That is the user is not heavily penalized in shifting every three years between the full national sample and the one-fifth national sample.

For the cost ratio p=0.5 the average rel p.a.e. values for an alternating large-and-small sample design with large sample periodicity of six years are less than or equal to the average rel p.a.e. values for the single large-sample scenario with periodicity of five years. Furthermore, the user of U.S.-level data will be receiving the benefits associated with the receipt of fresh data every three years instead of every five years. Related numerical results will be found in Smith, Ghosh, and Chang (1996).

Our main conclusion from the present study is that the National Center for Education Statistics should consider adopting an alternating large-and-small-sample design for SASS with an appropriate full-sample periodicity together with a mid-period fractional-sample to provide a periodic update at the national level and for larger States.

References

Abramson, R. *et al.* (1996), "1993-94 Schools and Staffing Survey: Sample Design and Estimation," Technical Report NCES 96-089, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics (forthcoming).

Ghosh, D., Kaufman, S., Smith, W. and Chang, M. (1994), "Optimal Periodicity of a Survey: Sampling Error, Data Deterioration, and Cost," *1994 Proceedings of the ASA Section on Survey Research Methods,* 1122-1127.

Kaufman, S. (1991), "1988 Schools and Staffing Survey Sample Design and Estimation," Technical Report NCES 91-127, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Kaufman, S. and Huang, H. (1993), "1990-91 Schools and Staffing Survey: Sample Design and Estimation," Technical Report NCES 93-449, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," *International Statistical Review,* 45, 13-28.

Smith, W. (1980), "Sample Size and Timing Decisions for Repeated Socioeconomic Surveys," unpublished D.Sc. dissertation, The George Washington University, School of Engineering and Applied Science.

Smith, W. and Barzily, Z. (1982), "Kalman Filter Techniques for Control of Repeated Economic Surveys," *Journal of Economic Dynamics and Control,* 4, 261-279.

Smith, W., Ghosh, D. and Chang, M. (1995), "Optimal Periodicity of a Survey: Alternatives Under Cost and Policy Constraints," *1995 Proceedings of the ASA Section on Survey Research Methods.*

Smith, W., Ghosh, D. and Chang, M. (1996), "Optimizing the Periodicity of the Schools and Staffing Survey: An Updated Assessment Based on Three Rounds of SASS Data," Technical Report, Synectics for Management Decisions, Inc., Arlington, VA (forthcoming).