# OPTIMAL PERIODICITY OF A SURVEY: ALTERNATIVES UNDER COST AND POLICY CONSTRAINTS

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

KEY WORDS: Data user needs; Indirect estimators; Loss functions; Probable-error models; Repeated surveys; Small area estimates; Statistical policy issues; Structural time series modeling

This paper is a progress report from a series of ongoing studies related to constrained optimization of the periodicity of school-based surveys -- that is, considering a range of choices of sample size and intersurvey timing intervals subject to a set of external constraints and programmatic goals for the fulfillment of data user needs. Ghosh *et al.* (1994) presented our general approach to these questions via a family of "probable-error" models with joint consideration of sampling error, data deterioration, and cost. There we addressed some of the tradeoffs, under a given multi-year budget for fixed and variable survey costs, between more frequent data collections with smaller sample sizes at each collection and less frequent data collections with larger sample sizes at each collection.

We now give more explicit attention to the statistical policy issues that arise when a set of survey redesign options confronts the policymaker with the possible adoption of "indirect estimation" methods for some subnational or subdomain estimates while, say, retaining "direct estimation" methods for national-level statistics and for the larger states and larger analytic domains of interest. The statistical policy framework we adopt here is in the spirit of the "recommendations and cautions" set forth in *Indirect Estimators in Federal Programs* (Subcommittee on Small Area Estimation, 1993).

## Schools and Staffing Survey

Our work has been specifically directed toward techniques that may lead to future redesign options for the Schools and Staffing Survey (SASS). SASS has been developed and sponsored by the U.S. National Center for Education Statistics (NCES) and is conducted for NCES by the U.S. Bureau of the Census. As stated in Bobbitt *et al.* (1995), "SASS is an integrated survey of public and private schools, school districts, principals, and teachers. It was conducted first during the 1987-88 school year, again in 1990-91 and 1993-94, and will be conducted at five-year intervals thereafter. SASS is a mail survey that collects public and private sector data on the Nation's elementary and secondary teaching force, aspects of teacher supply and demand, teacher workplace conditions, characteristics of school principals, and school policies and practices...."

The shift from three-year intervals to five-year intervals is understood to be the result of current and foreseeable budgetary resource constraints for federally-sponsored education surveys and does not rule out consideration of a range of design or redesign options for SASS in the 21st century. Electronic recordkeeping, new data collection technologies, and near real-time data processing capabilities may well open up new design options for school-based surveys.

## Partial Redesign of a School-Based Survey

The first three rounds of SASS were conducted at three-year intervals and the intention was that each data collection would have a sufficient sample size to permit statistical estimates to be made for most public school variables and school types at the geographic level of individual states. After data collection and analysis of the 1987-88 SASS, it became evident that "(1) state estimates from the states with smaller populations had higher than expected standard errors, (2) state estimates from the states with larger populations had lower than expected standard errors, (3) state elementary and state secondary estimates could not be made except for the largest states, and (4) the overall national estimates had much lower than expected standard errors" (Kaufman and Huang, 1993). In view of these findings, the design for the 1990-91 SASS was changed to reduce the sample sizes for the largest states and increase the sample sizes for the smallest states. The result was that direct estimates for 1990-91 (and 1993-94) are available for individual states for most school and teacher variables for elementary and secondary schools separately -- and, in most cases, for combined public schools (with grade spans of grade 6 or less to more than grade 8). The quality of national-level estimates was not degraded appreciably by these reallocation steps. Producing separate estimates for elementary and secondary schools was a major objective and hence a major change in the sample allocation was felt to be justified.

## Direct and Indirect Estimators

This example serves to illuminate a design and estimation challenge for school-based surveys such as SASS. The present policy-and-practice setting for

SASS is that only "direct estimates" (and their associated estimated standard errors) will be published by NCES in its official publications.

NCES has a broad legislative mandate to "collect, analyze, and disseminate statistics and other data related to education in the United States and other nations." Other federal statistical agencies operate under somewhat different or additional legislative mandates. For example, the Bureau of Labor Statistics (BLS) prepares monthly employment and unemployment estimates for some 5,300 geographic areas, including "...subcounty areas for which data are required by legislation." Since 1989, using data from the Current Population Survey (CPS), BLS has been publishing monthly *direct* sample survey estimates of employment and unemployment for the 11 largest states as well as for Los Angeles and New York City. BLS also publishes monthly *indirect* estimates for the 39 smaller states and the District of Columbia.

"The method used to provide [these] monthly state estimates [for the smaller states] is based on the time series approach to sample survey data. Originally suggested by Scott and Smith (1974), this approach treats the population values as stochastic and uses signal extraction techniques developed in the time series literature to improve on the direct survey estimator." ... "The signal is represented by a time series model that incorporates historical relationships in the monthly CPS estimates along with auxiliary data from the Unemployment Insurance (UI) and Current Employment Statistics (CES) programs. The time series model is combined with a noise model that reflects key characteristics of the sampling error to produce estimates of the true labor force values. This estimator has been shown to be design consistent under general conditions by Bell and Hillmer (1990) and is optimal under the model assumptions." See Chapter 5 in (Subcommittee, 1993); also see Tiller (1992).

A similar approach was taken in Ghosh *et al.* (1994) which assumed, for one model, that there is an underlying stochastic process that is observed periodically by the repeated survey data collections and that this process can be modeled as an ARIMA(0,1,1) time series process observed with sampling error. The formulation of the model is based on a general modeling procedure set forth in Smith (1980) and Smith and Barzily (1982) using Kalman filter concepts." Average cost as a function of sample size and intersurvey time interval (in years) is minimized by a numerical search procedure for a hypothetical survey with given cost coefficients and known noise covariances, yielding a jointly optimal solution for sample size and intersurvey interval. Available methods for the analysis of repeated surveys are summarized in Appendix A of the present paper. The Smith-Zalkind-Barzily (S-Z-B) approach is described in Appendix B. An extension of Ghosh's probable-error model paradigm to an assumed random walk process is outlined in Ghosh (1995).

**Possible Enhancement of SASS Estimates**

Assume a simple vector autoregressive process that evolves in discrete time at one-year accounting intervals. The vector process may involve a potentially large number of variables that may be observed through data collections at the level of local public schools. A few core variables are selected for observation through two different series of repeated surveys. The first observation series is assumed to be the ongoing annual data collection that is known as the Common Core of Data (CCD). The CCD system covers all public schools in the U.S. and is carried out within States by State education agencies (SEAs). The second observation series is assumed to be the public school component of SASS, for which three rounds of data have now been collected at three-year intervals. SASS covers a sample of public schools with some overlap schools in successive rounds of the survey. SASS also covers a sample of private schools, but these are not considered here.

Both the CCD and SASS series collect data from individual schools on such school variables as grade-by-grade enrollment, number of teachers, ethnic and gender components of enrollment, and number of students eligible for or receiving free lunches. In addition to such common or "core" variables, SASS collects data on such variables as the number of students served by Chaper 1 services, the number of K-12 (Kindergarten through grade 12) teachers who are new to the school this year, the number of K-12 teachers who left the school between October 1 of last year and October 1 of this school year, and the number of K-12 teachers who have a degree beyond the bachelor's degree.

We are currently exploring the possible dependence of components observed in the SASS series, but unobserved in the CCD series, on the observed components in the CCD series. For this purpose we may fit a set of equations in structural time series form (cf. Harvey, 1989) with a signal modeled with components (possibly time varying) that include a Regressor component, a Trend component, and an Irregular component. There is no Seasonal component since the established accounting period for school-

based reporting is annual. If the explanatory power of the CCD regressor variables turns out to be weak, such a finding would support a more frequent SASS data collection. If the dependence turns out to be nontrivial, this finding would support, within limits, a less frequent SASS data collection.

In the course of this work we expect to apply the estimation methodology for short time series set forth in Anderson (1978) for AR(1) processes and extended by Azzalini (1981), and Shumway (1988). We refer to this foundation as the A-A-S approach and will be attempting to connect it to the S-Z-B approach summarized in Appendix B.

Ghosh *et al.* (1994) demonstrated how to determine the optimum periodicity of a survey if the process model is known and is fairly simple (e.g., AR(1), ARIMA(0,1,1) or the Random Walk model). SASS data has been collected only three times; therefore, it is not feasible to fully determine the process model from the SASS data alone. But CCD, which is collected annually, has been in operation for several years and is a complete census. For selected SASS variables not included in CCD, we intend to develop linear models consisting of CCD variables as the candidate explanatory variables for the selected SASS variables in each year of SASS data collection. Such a linear model is like a newly constructed variable; let us call it $M$. The variable $M$ is constructed entirely of CCD variables and thus is defined for each unit (school) in CCD. We may then use Anderson's method to obtain estimated autocorrelation and partial autocorrelation functions over appropriate subdomains of units of CCD. From these, we can estimate the process model for $M$. We can then use the available SASS data and the model for $M$ to estimate a model equation for SASS. If this model turns out be a simple process we can then apply the following cost/error principles in our search for an optimal periodicity.

**Direct and Imputed Costs in Choice of Periodicity**

Any formalization of the problem of seeking an "optimal" choice of survey interval and survey size must account for the fixed and variable costs of operating a system of repeated surveys, such as SASS, as well as imputed costs due to increasing errors in the estimates as sample size is reduced and out-of-date estimates are used. In a recent book on survey errors and survey costs, Groves (1989) provides an up-to-date review of the kinds of considerations which should go into creating cost-and-error models for surveys, with particular emphasis on household surveys. Currently there is no comparable work on cost-and-error modeling for surveys of institutions such as schools.

In the case of SASS, there is an ongoing, more-or-less fixed annual cost of maintaining the core elements of the SASS system whether or not a survey is conducted in a particular year. Some costs might be regarded as either fixed or variable. Among these are the costs of updating list and area frames, with special emphasis on updates immediately preceding each wave of data collection. In this paper we lump such costs with the fixed annual costs of maintaining institutional memory for all aspects of SASS, making evolutionary design changes in coverage and content to be incorporated in the successive waves of data colletion, and conducting ongoing research in support of SASS processing and estimation procedures.

In addition to the directly measurable dollar outlays associated with maintaining and operating the SASS system, it is possible to include imputed dollar costs to represent the loss or penalty which is incurred by public and private users as a result of using outdated survey data. Smith and Zalkind (1978), Smith (1980), and Smith and Barzily (1982) used such an approach, formulating an imputed loss associated with the use of imprecise estimates from an observed economic process where the objective was the allocation of public funds on the basis of such estimates. This approach of Smith, Zalkind, and Barzily involves a framework in which knowledge of the state of a socioeconomic process is characterized as the level of a stock of information (an equivalent sample size on hand). The S-Z-B approach requires a policymaker to select a scale factor or equivalence to characterize the "cost of not knowing" in dollar terms so that the imputed cost or loss can be combined in the same formulas with the dollar outlays. See Appendix B for additional discussion of the S-Z-B approach.

**Appendix A: Methods for Repeated Surveys**

Since the early papers of Scott and Smith (1974) and Scott, Smith, and Jones (1977), there has been a renewed and growing interest in the application of time series methods to survey data. An excellent review article, Binder and Hidiroglou (1988), may be found in volume 6 of the *Handbook of Statistics*. This review and the papers by Bell (1984), Bell and Hillmer (1990), and Tam (1987) provide a balanced account of time series approaches, including state-space modeling and Kalman filter techniques, for use with data from repeated surveys. Although most statisticians are now

aware of the time series methods of Box and Jenkins (1970), who provided an understandable, systematized approach to model identification, estimation, and forecasting, many survey statisticians are still unaware of the potential of the time series methods for improving estimation with survey data in the sense of minimizing mean squared error. The key principle in the time series approach is that there is information in the time series structure of an observed process which may be used to make better estimates by combining information from past data collections with the new information from a current data collection than would be made if the current data were to be used alone.

**Signal Extraction, Kalman Filters, and State Space**

Classical survey estimates are made under assumptions that the observed variables, whether of labor force, or school enrollment, or other socioeconomic phenomena, have values that are fixed but are observed with sampling error (and possibly nonsampling error). The time series approach regards the process variables as stochastically varying over time and the identification problem is to find a parsimonious time series model evolving in discrete time such as one-year intervals, that will capture the main features of the underlying process sufficiently well. For univariate processes it has often been found to be quite satisfactory to fit a model with a very simple structure, such as an autoregressive model of order 1 or 2.

In the case of surveys of schools and similar institutions, the natural accounting period is the school year, so that within-year changes are of secondary interest and seasonal effects do not arise. Linear trends are easily incorporated in Box-Jenkins type models and can, if desired be factored out by taking first differences of successive observations. Thus the trend component may be accounted for separately. One useful model that is related to classical exponential smoothing is the Box-Jenkins ARIMA(0,1,1). It is mildly nonstationary and can "wander" up and down; in one sense the current process value serves as a "local mean" for the process as time moves ahead one step and the process noise term kicks the process up or down a bit. In classical Box-Jenkins modeling it is assumed that the process is observed without observation error.

Borrowing from the contributions of R. Kalman in the control engineering literature in the 1960s, the Scott-Smith time series approach utilizes a two-equation setup in which there is a process equation which represents the evolution of the underlying (unobserved) process through time. The second equation, the observation equation, consists of the sum of the underlying process variable and an observation noise term. The noise term in some simple models may just represent the sampling error. In other cases it may have a structure of its own. The state of the process may be represented by a vector with two or more components, representing, for example, the levels of two or more process variables such as number of teachers and number of students at a school, or in an aggregate of schools within a state or other subnational grouping.

The classical Kalman approach assumes that the variance-covariance (V-C) matrices are known and time invariant. In real world settings, the V-C matrices will not be known and will have to be estimated from the data. Furthermore, they will not necessarily be time invariant. These complications have led to the formulation of extended Kalman filters which, although theoretically sound, place an estimation burden on the available data and may lead to inconclusive results. Also, it is somewhat awkward to try to accommodate nonlinear features such as the presence of level-dependent variances. For example in a set of elementary schools arranged by size within one state, the variance in enrollment or in number of teachers will typically depend on the size of the school and hence the number of teachers. This is easier to capture using one of the model types known as state-dependent models and in particular with a class of models known as bilinear models (see Smith, 1994).

**Cost Models with Fixed and Variable Costs**

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, if it is of considerable magnitude, could affect the quality of decisions made by users. On the other hand, if the survey is undertaken frequently, the costs of conducting the survey, of analyzing the data, and of response burden may be judged to exceed the benefits achieved in using fresh data.

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.

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 be dependent on the periodicity. We represent it as $C_0^k$ (where $k$ is the periodicity) which may be regarded as increasing with increasing intersurvey interval; i.e., the start-up cost is higher if the interval is three years than if the interval is two years.

We usually assume that total resources for a multi-year time period are fixed. The different possible periodicities spend these total resources in different ways. This assumption then determines the possible sample sizes every time the survey is undertaken corresponding to different periodicities. A modified approach would be to use similar models but to attempt to take explicit account of the fact that total resources may be arbitrarily reduced by external constraints and formulate the decision problem within that framework.

## Appendix B: The S-Z-B Optimization Tools

In Smith (1980) the concept of "equivalent sample size" was adapted to a reformulation of the optimal filter theorem for a scalar (single variable) model of an evolving process observed at discrete points in time. The development was as follows:

Consider a repeated survey system in which the process state is represented by the scalar state variable $x(j)$ evolving as a scalar random walk in discrete time, $x(j) = x(j-1) + w(j)$, where $w(j)$ is the process noise term, with scalar survey measurements $y(k)$ given by $y(k) = x(k) + b(k)$, where $b(k)$ is the measurement noise term and the sample size at each survey time $k$ is the scalar quantity $n(k)$ and the sample noise variance $B(k)$ is given by $B(k) = R / n(k)$ with $R$ as the assumed known constant unit measurement noise variance. The Kalman gain in the optimal filter theorem then becomes

$$K(k) = C(k|k\text{-}t)\,[\,C(k|k\text{-}T) + B(k)\,]^{-1}$$

$$= [C(k\text{-}T)|k\text{-}T) + TQ]\,/\,[C(k\text{-}T|k\text{-}T) + TQ + R/n(k)]\,,$$

which is of the same form as the exponential smoothing parameter in a development due to Harrison; see Harrison and Stevens (1976). The error variance equations in the optimal filter theorem are now of the form

*Between surveys*

$$C(k+j \mid k) = C(k \mid k) + j\,Q\,,$$

*At surveys*

$$C(k \mid k) = [1 - K(k)\,]\,C(k \mid k\text{-}T)\,,$$

where $C(0 \mid 0)$, $Q$, and $R$ are positive scalars and so are $K(k)$, $C(k|k)$, and $C(k+j|k)$. In this development the scalar quantity $n^o(k|k)$ was then defined by $n^o(k|k) = RC^{-1}(k|k)$ and referred to as the updated equivalent sample size after surveying at survey time $k$ with no processing delay. It was further interpreted in inventory terms as the level of a "stock of information" on hand immediately after ordering $n(k)$ additional units (with no leadtime); that is, as an inventory "order level." The scalar quantity $n_r(k+j|k)$ was defined by $n_r(k+j|k) = RC^{-1}(k+j|k)$ and referred to as the equivalent sample size remaining at time $k+j$, $j$ time units after the survey time $k$. It was interpreted in inventory terms as the "stock on hand" $j$ time units after ordering and receiving new stock. For a fixed interval $T$ between surveys, assuming the system is in steady state, $n_r(k \mid k\text{-}T)$ was interpreted in inventory terms as the "reorder point" and $T$ as the "scheduling period." The Kalman gain becomes

$$K(k) = n(k) \,/\, [n_r(k|k\text{-}T) + n(k)]$$

and the updated equivalent sample size becomes

$$n^o(k|k) = n_r(k|k\text{-}T) + n(k)\,.$$

A further interpretation of $n^o(k|k)$ was that it is the size of a survey that would be required to have the same degree of precision as that provided by the combined amount $n_r(k|k\text{-}T) + n(k)$. This development led to a set of equivalent sample size relations in place of the error variance equations in the optimal filter theorem:

*Between surveys*

$$n_r(k+j|k) = n^o(k|k)[1\text{-}jQR^{-1}\,n^o(k|k)\,]^{-1}\,,$$

*At surveys*

$$n^o(k|k) = n(k) + n^o(k\text{-}T|k\text{-}T)\,[1 + TQR^{-1}\,n^o(k\text{-}T|k\text{-}T)\,]^{-1}\,.$$

Smith and Barzily (1982) gave a numerical example for a two-item process assumed to be a vector random walk with scalar sample size $n_d$ and integer sampling interval T (T = 1, 2, ..., 10 years) . With assumed cost coefficients for start-up cost and unit costs of interviewing, they demonstrated that the cost function J was convex in ($n_d$ , T) and found a minimum for J by a numerical search. They noted that a survey administrator who was "concerned that the underlying process parameters may take unexpected jumps or exhibit turning points, which are not modeled by the simple time-invariant random walk models, would presumably opt for sampling more frequently than the optimal interval found by this method."

**References**


Anderson, T.W. (1978), "Repeated Measurements on Autoregressive Processes," *Journal of the American Statistical Association,* 73, 271-278.

Azzalini, A. (1981), "Replicated Observations of Low Order Autoregressive Time Series," *Journal of Time Series Analysis,* 2, 63-70.

Bell, W. (1984), "Signal Extraction for Nonstationary Time Series," *Annals of Statistics,* 12, 646-664.

Bell, W.R. and Hillmer, S.C. (1990), "The Time Series Approach to Estimation for Repeated Surveys," *Survey Methodology,* 16, 195-215.

Binder, D.A. and Hidiroglou, M.A. (1988), "Sampling in Time," in *Handbook of Statistics, Vol. 6,* ed. P.R. Krishnaiah and C.R. Rao, Amsterdam: Elsevier Sci. Publishers, 187-211.

Bobbitt, S.A., Broughman, S.P. and Gruber, K.J. (1995) "Schools and Staffing in the United States: Selected Data for Public and Private Schools," E.D. TABS NCES 95-191, U.S. Dept. of Educ.

Box, G.E.P., and Jenkins, G.M. (1970), *Time Series Analysis: Forecasting and Control,* San Francisco: Holden-Day.

Ghosh, D. (1995), "Periodicity of School Surveys: An Extension of Probable-Error Modeling for the Case of a Random Walk Process," unpublished technical note, Synectics for Management Decisions, Inc.

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

Groves, R.M. (1989), *Survey Errors and Survey Costs,* New York: Wiley.

Harrison, P.J. and Stevens, C.F. (1976), "Bayesian Forecasting" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 38, 205-247.

Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter,* New York NY: Cambridge University Press

Kaufman, S. (1991), "1988 Schools and Staffing Survey Sample Design and Estimation," Technical Report NCES 91-127, U.S. Dept. of Educ.

Kaufman, S. and Huang, H. (1993), "1990-91 Schools and Staffing Survey: Sample Design and Estimation," Technical Report NCES 93-449, U.S. Dept. of Educ.

Scott, A.J. and Smith, T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods," *Journal of the American Statistical Association,* 69, 674-678.

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.

Shumway, R.H. (1988), *Applied Statistical Time Series Analysis*, Englewood Cliffs: Prentice-Hall

Smith, W. (1994), "Nonlinear Modeling for Schools Data with Level-Dependent Variances," unpublished technical note, Synectics for Management Decisions, Inc.

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

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. and Zalkind, D. (1978), "Statistical Decision and Control Approaches for Allocation of Funds," *1978 Proceedings of the ASA Section on Survey Research Methods,* 108-113.

Subcommittee on Small Area Estimation (1993), *Statistical Policy Working Paper 21: Indirect Estimators in Federal Programs.* Washington DC: Statistical Policy Office, OMB

Tam, S.M. (1987), "Analysis of Repeated Surveys Using a Dynamic Linear Model," *International Statistical Review,* 55, 63-73.

Tiller, R. (1992), "Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey," *Journal of Official Statistics*, 8, 149-166.