OPTIMAL PERIODICITY OF A SURVEY: SAMPLING ERROR, DATA DETERIORATION, AND COST

Dhiren Ghosh, Synectics for Management Decisions, Steven F. Kaufman, National Center for Education Statistics, Wray Smith and Michael Chang, Synectics for Management Decisions
Dhiren Ghosh, Synectics Mgmt Decisions, 3030 Clarendon Blvd., Suite 305, Arlington, VA 22201

Key Words: Probable error, Loss function, ARIMA models, Repeated Surveys

Government agencies collect many different kinds of statistical data through sample surveys conducted on a periodic basis (monthly, annually, or at multi-year intervals). When the periodicity is not mandated by law, data deterioration, cost, and sampling error in the data may be considered jointly to determine optimum intersurvey time intervals. In a decision-making process, any loss due to using the survey estimate instead of the true value may be thought of as arising in part from sampling error; also, with the passage of time, the true value evolves and the survey dataset becomes obsolete. In this paper several statistical models of data deterioration are considered jointly with standard cost functions for a survey; that is, "cost-and-error models."

The concept of "probable error" is utilized in three related models in which the additivity of errors over time is assumed. A loss function is minimized in a fourth model along with a procedure for estimating the loss parameter. A fifth model assumes 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 this model is based on a general modeling procedure set forth in Smith (1980) and Smith and Barzily (1982) using Kalman filter concepts. The use of the first three models as decision aids in the choice of optimum intersurvey intervals is illustrated with data from the Schools and Staffing Survey (SASS).

We assume that data users will continue to use the data obtained from the most recent 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 too frequently, the costs of conducting the survey and analyzing the data and the response burden may be judged to outweigh the benefits to be 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 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 $C = C_0 + n\ C_1$. 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 periodicity; i.e., the start-up cost is more if the periodicity is 3 years compared to the start-up cost if the periodicity is 2 years, and so on. On the other hand, the start-up cost may be considered to be constant; i.e., it may not depend on the periodicity of the survey.

In the family of statistical models that we develop below, we assume that the total resources are fixed. The different possible periodicities spend this total resources in different ways. This assumption then determines the possible sample sizes every time the survey is undertaken corresponding to different periodicities. Thus, if we are comparing two possible periodicities, say two years as against three years, we consider a six-year cycle (the least common multiple of the two periodicity numbers). In the six-year cycle, a survey with periodicity two years will be conducted three times while a survey with periodicity three years will be conducted only twice. If C_0^k and C_1 (where C_1 is assumed to be independent of the periodicity of the survey.) are known (whether the start-up cost is constant or increasing) we can calculate the possible sample sizes for these two alternatives where the total measure C is also known.

A Family of Error Models

We assume that the true value of a variable of interest remains constant for a year after the survey date. So the error "committed" in using the survey estimate is exactly equal to the difference between the survey estimate and the true value. So during one year from the survey date any user incurs an error which equals the difference between the true value and the survey estimate. The estimate of the standard error from the survey provides an indication of this difference. The survey estimate is normally distributed around the true value with a standard deviation which is the standard error of the estimate. The difference between the true value and the survey estimate is the deviation from the mean in the normal distribution of

the survey estimates considered as random variables. The average of these deviations is called the probable error. It is calculated as follows for any normal distribution:

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}|x-m|e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \sqrt{\frac{2}{\pi}}\sigma = 0.8\sigma$$

Thus the average error incurred by any user during the first year after the survey is equal $0.8\sigma/\sqrt{n}$ where $\sigma/\sqrt{n}$ is the standard error of the estimate. At the end of one year, we assume that the true value undergoes a change denoted by D_1. So the expected value of the total error committed by all the users is the sum of the probable error and D_1 . Proceeding in the same manner we denote the change in the second year as D_2 and so on.

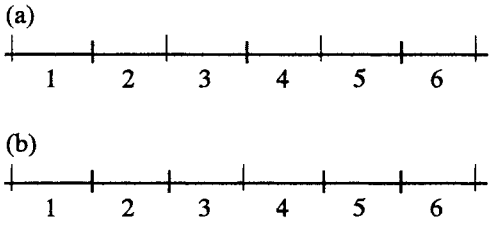
In **Model 1** we ignore the direction of the change in the true value and just add the probable error to the sampling error for the change in the true value.

In **Model 2** we do not ignore the direction of the change. If the change occurs in the same direction as the survey estimate, we ignore the diminution in the shift due to the survey estimate already being in the same direction. If the shift occurs in the opposite direction the total error due to using the old survey estimate can be denoted as D_1 + probable error. Taking the average of the two possibilities we denote the expected error as D_1 + ½(probable error). Here the error terms D_1 and D_2 are treated as if they were random variables. Proceeding in the same manner we denote the change in the third year as D_3 and calculate the expected error as above.

In **Model 3** we add the square of the change to sampling error to denote the total error after the first year. We further assume that the change is normally distributed so the sum of the sampling error and the change is also normally distributed. This enables us to calculate the probable error of the normal distribution.

Determination of Periodicity of a Survey

We start with the assumption that the total resources are fixed and the problem is to determine the best periodicity of a survey. We illustrate the solution of this problem for the special case when the alternatives are: (a) every two years (biennial), or (b) every three years (triennial). We consider a cycle of six years with the survey taken at the starting point.



For a six year cycle, the biennial survey is conducted three times and the triennial survey is conducted twice. We do not take into account the survey after six years since a new cycle starts after the sixth year. We further assume that the true unobserved value remains unchanged for a year after the survey is completed. At the end of a year, the value changes by an amount D_1 and at the end of two years, the value changes again by an amount D_2 . These D_1 and D_2 denote the shift in the true values. If the standard error of a variable in a survey (assuming SRS) is $\sigma/(n^{½})$ where σ is the standard deviation and n is the sample size, the average error or probable error of the estimate is $0.8^{*}\sigma/(n^{½})$. That is, every time the estimated value is used (since the true value is unknown) an *error is committed*; the expected value of this error is $0.8\sigma/(n^{½})$. During the year after the survey, the survey value will be used for any decision, so the average error committed during the year is $0.8\sigma/(n^{½})$. When a year elapses the shift in the true value is added to the expected error to obtain the expected error committed during the second year and so on.

Let us examine the error committed for every year following the survey. These errors over the years are assumed to be additive. Let n_a and n_b be the sample sizes for the biennial and the triennial surveys respectively with simple random sampling. We further assume that the standard deviation in the population for the variable of interest remains unchanged during the whole cycle.

Model 1.

(a)

Year (Ordinal)	Average Error Committed
1	$0.8\sigma/(n_a^{1/2})$
2	$D_1 + 0.8\sigma/(n_a^{1/2})$
3	$0.8\sigma/(n_a^{1/2})$
4	$D_1 + 0.8\sigma/(n_a^{1/2})$
5	$0.8\sigma/(n_a^{1/2})$
6	$D_1 + 0.8\sigma/(n_a^{1/2})$
Average Total Error Committed (in six years)	$3D_1 + 4.8\sigma/(n_a^{1/2})$

(b)

Year (Ordinal)	Average Error Committed
1	$0.8\sigma/(n_b^{1/2})$
2	$D_1 + 0.8\sigma/(n_b^{1/2})$
3	$D_1 + D_2 + 0.8\sigma/(n_b^{1/2})$
4	$0.8\sigma/(n_b^{1/2})$
5	$D_1 + 0.8\sigma/(n_b^{1/2})$
6	$D_1 + D_2 + 0.8\sigma/(n_b^{1/2})$
Average Total Error Committed (in six years)	$4D_1 + 2D_2 + 4.8\sigma/(n_b^{1/2})$

Thus (a) is preferable if

$$3D_1 + 4.8\sigma/(n_a^{1/2}) < 4D_1 + 2D_2 + 4.8\sigma/(n_b^{1/2})$$

or

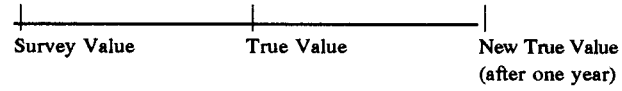
$$4.8\sigma[(n_a^{1/2}) - (n_b^{1/2})] < D_1 + 2D_2$$

and (b) is preferable if

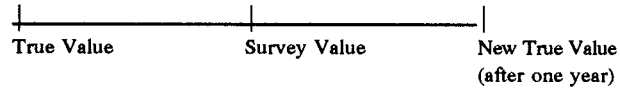
$$4.8\sigma[(n_a^{1/2}) - (n_b^{1/2})] > D_1 + 2D_2$$

Model 2

In Model 1, we assumed that the expected error and the shift in the value are additive for estimating the error in the second or the third year. Examine the following hypothetical case: In this case the addition of the errors seems reasonable.



Alternatively, examine the following case: In such a case, the average error in using the survey value after a year is definitely not $D_1 + 0.8\sigma/(n_b^{1/2})$, it is $D_1 - 0.8\sigma/(n_b^{1/2})$.



If we ignore this contribution of the survey error toward a diminution of the effect of the shift in the true value, the estimate of the average error committed after the first year is $D_1 + 0.4\sigma/(n_b^{1/2})$, and so on. So the errors look as follows:

Year (Ordinal)	(a)	(b)
1	$0.8\sigma/(n_a^{1/2})$	$0.8\sigma/(n_b^{1/2})$
2	$D_1 + 0.4\sigma/(n_a^{1/2})$	$D_1 + 0.4\sigma/(n_b^{1/2})$
3	$0.8\sigma/(n_a^{1/2})$	$D_1 + D_2 + 0.4\sigma/(n_b^{1/2})$
4	$D_1 + 0.4\sigma/(n_a^{1/2})$	$0.8\sigma/(n_b^{1/2})$
5	$0.8\sigma/(n_a^{1/2})$	$D_1 + 0.4\sigma/(n_b^{1/2})$
6	$D_1 + 0.4\sigma/(n_a^{1/2})$	$D_1 + D_2 + 0.4\sigma/(n_b^{1/2})$
Average Total Error Committed (in six years)	$3D_1 + 3.6\sigma/(n_a^{1/2})$	$4D_1 + 2D_2 + 3.2\sigma/(n_b^{1/2})$

Thus (a) is preferable if

$$3.6\sigma/(n_a^{1/2}) - 3.2\sigma/(n_b^{1/2}) < D_1 + 2D_2$$

and (b) is preferable if

$$3.6\sigma/(n_a^{1/2}) - 3.2\sigma/(n_b^{1/2}) > D_1 + 2D_2$$

Model 3

Let us assume that x_j is the value for the j^{th} year and

$$x_{j+1} - x_j = d_j$$

Let the variance of d_j's over the years be $D^2(1)$. For a Random Walk stochastic process, the d_j's are not normally distributed. Similarly, let $D^2(2)$ be the variance of differences over 2 years. For a Random Walk process, $D^2(2) = 2D^2(1)$. But, in general, this relation may not hold because of the autocorrelation of the changes between consecutive years. In general, $D^2(2)$ or $D^2(1)$ is not normally distributed. Never the less, we assume that the probable error from this process is 0.8D(1) or 0.8D(2), as in the case of normal distribution. Under the assumptions, the error looks as follows:

Year (Ordinal)	(a)	(b)
1	$0.8\sigma/(n_a^{1/4})$	$0.8\sigma/(n_b^{1/4})$
2	$0.8D(1) + 0.8\sigma/(n_a^{1/4})$	$0.8D(1) + 0.8\sigma/(n_b^{1/4})$
3	$0.8\sigma/(n_a^{1/4})$	$0.8D(2) + 0.8\sigma/(n_b^{1/4})$
4	$0.8D(1) + 0.8\sigma/(n_a^{1/4})$	$0.8\sigma/(n_b^{1/4})$
5	$0.8\sigma/(n_a^{1/4})$	$0.8D(1) + 0.8\sigma/(n_b^{1/4})$
6	$0.8D(1) + 0.8\sigma/(n_a^{1/4})$	$0.8D(2) + 0.8\sigma/(n_b^{1/4})$
Average Total Error Committed (in six years)	$2.4D(1) + 4.8\sigma/(n_a^{1/4})$	$1.6D(1) + 1.6D(2) + 4.8\sigma/(n_b^{1/4})$

Thus (a) is preferable if

$$4.8[\sigma/(n_a^{1/2}) - \sigma/(n_b^{1/2})] < 1.6D(2) - 0.8D(1)$$

and (b) is preferable if

$$4.8[\sigma/(n_a^{1/2}) - \sigma/(n_b^{1/2})] > 1.6D(2) - 0.8D(1).$$

Model 4

In Model 4 we introduce the concept of a loss parameter that converts the error whether sampling error alone is coupled with the shift over time. This converts the error into loss expressed as monetary units. The sum of average cost and average error over a period of years is minimized to determine the optimum periodicity. We present below the operation of each of these four models.

Let X_k be the true value of variable in the k^{th} year and

$\hat{X}_K$ be the survey value

$$\hat{X}_K = X_K + e_K,\ E(e_K) = 0$$

$$E(\hat{X}_k - X_{k+T-1})^2$$
$$= E(\hat{X}_k - X_K + X_K - X_{K+1} + X_{K+1} - \dots - X_{K+T-1})$$
$$= E(e_K + (T-1)d)^2, \text{ under the Random Walk Model}$$
$$= E(e_K^2) + E((T-1)d^2)$$
$$= E(e_K^2) + (T-1)E(d^2)$$
$$= V(e_K) + (T-1)E(d^2)$$

If $\hat{X}_b$ and $\hat{X}_{b+p}$ are two survey estimates p years apart, let

$$M = \frac{(\hat{X}_b - \hat{X}_{b+p})^2}{b},\ M \text{ is an estimate of } E(d^2)$$

The total error in T years is the following:

Year (Ordinal)	Error
1	$S^2/n + 0 \cdot M$
2	$S^2/n + 1 \cdot M$
⋮	⋮
T	$S^2/n + (T-1)M$
Total Error (in T years)	$T(S^2/n) + \frac{1}{2}T(T-1)M$
Average Error Per Year (in a cycle of T years)	$S^2/n + \frac{1}{2}(T-1)M$

Let α be a weighting factor that converts error into cost or loss. Then

$$\text{Average Cost Per Year} = J = \frac{C_0 + nC_1}{T} + \alpha\left(\frac{S^2}{n} + \frac{T-1}{2}M\right)$$

$$\frac{\partial J}{\partial n} = \frac{C_1}{T} - \alpha\frac{S^2}{n^2} = 0, \text{ this gives } n = \sqrt{\frac{\alpha S^2 T}{C_1}}$$

$$\text{Average Cost} = J = C_0 + \frac{2\sqrt{C_1\alpha}S}{\sqrt{T}} + \alpha\frac{T-1}{2}M, \text{ for } T = 1, 2, 3, \ldots$$

The optimum T is the one for which the average cost is the minimum.

Model 5

In the above four models we have not assumed any underlying stochastic process for the variables that are measured in the surveys. In Model 5 we assume that the underlying process is consistent with an ARIMA (0,1,1) time series model. Consequently data users would be using a minimum mean square error forecast from the past data instead of the data of the last survey after the lapse of one or more intersurvey time intervals .

In this setup, let $e_k(j)$ be the j-step ahead forecast error based on data through time k. The mean square error is

$$E\left(e_{k-T}^2(T)\right) = M_{k-T}(0) + T\cdot E(d^2)$$

where $M_{k-t}(0)$ is the mean square error of the state estimate at the time k-T based on all data through time k-T.

If we assume that the survey system is in a steady state in the sense that

$$M_k(0) = M_{k+T}(0) = M$$

as a result of conducting surveys of constant sample size n every T periods. It can be shown from standard time series analysis techniques that

$$M = \left[\frac{T\cdot E(d^2)}{2}\right]\left[-1 + \frac{(1+4S^2)}{T\cdot n\cdot E(d^2)}\right]^{\frac{1}{2}}$$

We define the average cost per year as in Model 4

$$J = \frac{C_0+C_1 n}{T} + \frac{\alpha}{T}\sum_{j=0}^{T-1} M_k(j)$$

where $M_k(j)$ is the j-step ahead mean square error.

$$= \frac{C_0+C_1 n}{T} + \frac{\alpha}{T}\sum_{j=0}^{T-1}(M + j\cdot E(d^2))$$

$$= \frac{1}{T}[C_0 + C_1 n] + \alpha\left[-\frac{E(d^2)}{2} + E(d^2)\left[\frac{T\cdot S^2}{n\cdot E(d^2)} + \frac{T^2}{4}\right]^{\frac{1}{2}}\right]$$

Average cost J as a function of n and T can be minimized by solving formula for each T in a specified allowable set $T=\{1, 2 \cdots T_{max}\}$ and adopting the n, T) for which J is minimized.

A Note on the Determination of α, the Weighting Factor

One procedure is to assign a value for α strictly based on judgment. If we want to develop a more sophisticated approach for determining a value for α we may argue as follows:

If $C_0 + C_1 n$ is the cost of implementing a survey and it results in sampling error of S^2/n for one variable, the total cost is

$$C_0 + C_1 n + \alpha\frac{S^2}{n}$$

Differentiating with respect to n and equating to zero, we get:

$$C_1 - \alpha\frac{S^2}{n^2} = 0$$

or

$$n = \sqrt{\alpha\frac{S^2}{C_1}} = \sqrt{\frac{\alpha}{C_1}}S, \text{ thus } \alpha = \frac{n^2 C_1}{S^2}$$

We note that the marginal gain from increasing the sample size from n to n+1 is

$S^2/n - \alpha S^2/(n+1)$. The sample size is optimum when the marginal cost equals marginal gain.

$$or\ C_1 = \alpha\frac{S^2}{n} - \alpha\frac{S^2}{n+1}$$

$$or\ C_1 = \alpha S^2\left(\frac{1}{n(n+1)}\right)$$

$$or\ n^2 + n - \alpha\frac{S^2}{C_1} = 0$$

$$or\ n = \frac{-1 + \sqrt{1 + \frac{4\alpha S^2}{C_1}}}{2},\ \textit{disregarding the other root}$$

$$or\ (2n+1)^2 = 1 + \frac{4\alpha S^2}{C_1}$$

$$or\ \frac{C_1(2n+1)^2 - 1}{4S^2} = \alpha$$

It can be seen that the two values of α are close to each other. If we look at the sample sizes employed in previous surveys and construct the cost function, we can get a value for α that has an objective basis.

Conclusion

These models have provided 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.

References

Binder, D.A. and Dick, J.P. (1989), "Modeling and Estimation for Repeated Surveys," *Survey Methodology*, 15, 29-45.

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 Science Publishers, 187-211.

Blight, B.J.N. and Scott, A.J. (1973), "A Stochastic Model for Repeated Surveys," *Journal of the Royal Statistical Society*, Ser. B, 35, 61-68.

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

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

Harrison, P.J. (1967), "Exponential Smoothing and Short-Term Sales Forecasting," *Management Science*, 13, 821-842.

Pfefferman, D. and Burck, L. (1990), "Robust Small Area Estimation Combining Time Series and Cross-Sectional Data," *Survey Methodology*, 16, 217-237.

Rao, J.N.K., Srinath, K.P. and Quenneville, B. (1989), "Estimation of Level and Change Using Current Preliminary Data," in *Panel Surveys*, ed. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh, pp. 457-479. New York: Wiley.

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.