SAMPLING IN TIME AND SPACE

Ronaldo Iachan, Research Triangle Institute
P. O. Box 12194, Research Triangle Park, NC 27709

KEY WORDS: Temporal variation, seasonal, autocorrelation.

1. Introduction

This paper discusses several situations where sampling takes place in space and/or time. In these cases, the efficiency of alternative sampling strategies will depend on the underlying spatial or temporal autocorrelation present in the population.

The classical finite population sampling theory considers a population of N units labeled 1,...,N. A survey variable of interest, Y, takes on values Y(1),..., Y(N) that are considered fixed in this approach. Under a superpopulation modeling approach, these values are assumed to arise as realizations of a random variable Y.

A simple superpopulation model appropriate for sampling in one dimension, for example, sampling in time alone (for a fixed network of units), is that the Y values follow a stationary process. Applications of such a stationary model may be found in Cochran (1946) and Iachan (1983). The assumption that the autocorrelation function is non-increasing (and non-negative) seems to arise naturally in this context: observations closer together (in time) are expected to be more strongly correlated than observations further apart (in time).

Following a description of some illustrative surveys in Section 2, Section 3 presents a brief discussion of the precision expected in the one-dimensional case. The equicorrelation case (constant correlation), which provides an upper bound for the variance under the given assumptions, is further discussed.

For sampling in a two-dimensional space (the plane), it is natural to index the population values Y(i,j). A similar stationary model would then consider autocorrelation functions of the form $\rho(u,v)$ between units with coordinates (i,j) and (i+u,i+v). Results for two-dimensional sampling are available in Iachan (1985b) and are summarized in Section 4. This section also discusses the case when time is one of the two dimensions considered. The three-dimensional case is discussed in Section 5.

Section 6 presents a discussion of design issues of practical importance when spatial and time dimensions are considered simultaneously. In particular, stratification issues are discussed and illustrated with the example surveys described in Section 2.

2. Illustrative surveys

Some problems that occur when the time dimension is of relevance in the survey design may be illustrated by a number of surveys in the recent experience of the Research Triangle Institute (RTI) and elsewhere.

Two types of survey in our recent experience illustrate similar design issues related to the existence of successive, regularly spaced measurements on each sample unit. The first type of survey is an electrical load research survey where a total load estimate is desired for each time of the day. For simplicity, we will assume that (24) hourly estimates are desired. The sample measurements are integrated totals or averages (e.g., over one hour) taken on each of n sample households for each of t days. For each of the (24) hourly estimates, one has available (a) household averages based on t daily measurements, and (b) a sample average based on nt measurements.

Integrated measurements are also generally taken for indoor air and personal exposure studies. In both the indoor air and the load research contexts, the autocorrelation between successive measurements on the same unit is expected to be significant. That is, the precision expected from a series of t measurements taken on each of n sample units falls in between the following two extreme scenarios:

(a) the nt measurements are uncorrelated;

(b) the measurements on a same sample unit are perfectly correlated.

A misleading picture may result from assuming either one of the two scenarios at the design phase. The effective sample size is between the "sample size", n, and the total number of measurements, nt. A more detailed discussion is provided in Section 3.

The sample units for the load research study are typically residential customers or housing units. Parameters of interest include the total seasonal load for a given hour of the day (e.g., from 1 to 2 PM) and also end-use specific loads for appliances such as air conditioning and central heating. The fact that such appliance uses are strongly associated with the season guides the design stratification discussed in Section 6.

The load research study also illustrates another problem related to the temporal dimension, namely that of appropriate unit definition. As discussed in Section 6, several factors need to be considered when deciding between units such as households or persons and the corresponding pairs obtained by crossing these units with time occasions. A similar problem arises in surveys of mobile populations, e.g., in the evaluation of the Migrant Education Program and in studies of the homeless population currently undertaken by RTI.

The primary objective of the descriptive study of the Migrant Education Program is characterizing program participants (children of migrant farmworkers and fishermen). The ultimate sample units are participating migrant students; states, local projects and schools are selected at earlier sampling stages. Due to the intrinsic mobility of the migrant population, the same sample student may be captured in several different locations. That is, the sample unit might be best defined as a migrant x time pair; however, to avoid the complexities resulting from this approach, a narrow time window is defined where a population "snapshot" is to be taken.

Migratory problems also occur in surveys of animal populations; capture-recapture methods applicable in those instances are not feasible for migrant students. An example of a non-mobile animal population study is provided in Iachan (1985a). The population considered in the study is the shellfish population along the East Coast of the United States in a given season; the focus of the article is on two-dimensional (spatial) stratification of the study area at a fixed point in time.

The temporal dimension in radon studies may be "designed away" by considering a fixed time period for measurement. Typically and conservatively, this time period is part of the season when contaminant (e.g., radon) levels are believed to be highest (usually the season when closed house conditions occur most often). In other types of indoor air surveys, day-to-day variability is such an important component of the total variance that the design must also include temporal randomization.

3. The one-dimensional case

This analysis applies not only when there is a meaningful ordering of the N population units but also when successive time measurements are taken on a fixed sample. While the first situation was discussed in Iachan (1983), the second situation is more carefully discussed here: it involves sampling of N units and t (integrated) measurements at regular periods on each sampled unit.

The superpopulation model considered in this case is that similar measurements taken on the same unit at times j and k, Y(j) and Y(k), have the same mean and variance and serial correlation, corr(j,k)= R(k-j), that only depends on the distance (or difference) between the two time periods. A special case considered further is that of constant correlation, R(k-j)=r, referred to as the equicorrelation case.

Under this model, the variance of a sample mean based on t successive measurements may be written as

$$\mathrm{Var}\,(\bar{Y}) = \frac{V}{t}\left(1 + \frac{2}{t}\sum_{j<k}^{t}\sum^{t} R(k-j)\right) \qquad (3.1)$$

where

$$\bar{Y} = \frac{1}{t}\sum_{1}^{t} \bar{Y}\,(j)$$

is the mean over time of similar measurements $\bar{Y}(j)$, and V is the sampling variance of each $\bar{Y}(j)$. Note that $\bar{Y}(j)$ may be the measurement for a fixed sample unit or averaged over the sample.

In the equicorrelation case, R(k-j) =r, all j, k, and the variance (3.1) simplifies to

$$\mathrm{Var}(\bar{Y}) = \frac{V}{t}\,(1 + (t-1)r) \qquad (3.2)$$

Note that the relative increase in variance due to the autocorrelation is

$$\Delta = [\mathrm{Var}(\bar{Y}) - \frac{V}{t}\,] \div \frac{V}{t} = (t-1)r\ .$$

The variance V/t would be obtained if successive measurements on a given unit were uncorrelated.

The models considered in the literature (e.g., Cochran, 1946, Iachan, 1983) assume that the autocorrelation function R(.) is non-increasing (and non-negative). Under this model, the simple equicorrelation model variance (3.2) provides an upper-bound for the variance. Consequently, conservative sample sizes to achieve a specified precision may be based on the equicorrelation model.

In the load research context, for instance, $\bar{Y}(j)$ is the sample mean load for the j^{th} successive measurement for a given time (e.g., hour) of the day, possibly for a specific end-use (or appliance). For each end-use, if hourly (integrated) total loads are considered, there are 24 such estimates, one for each hour-block of the day.

In the indoor air study example, the number of successive measurements will of course depend on (a) the total length of the monitoring period, and (b) the time period over which measurements are integrated. Both (a) and (b) are contingent upon the type of monitoring device employed in the study.

4. The two-dimensional case

This section considers a population of $N = N_1N_2$ units arranged in a grid of $N_1 = n_1k_1$ rows and $N_2 = n_2k_2$ columns. Rectangular strata with k_1k_2 units labeled (i,j) are then defined, and equal n_1n_2 in number.

In the two-dimensional context, models assumed by Quenouille (1949), Das (1950) and Bellhouse (1977) all have the form

$$E(Y_{ij}) = \mu,\ E\{(Y_{i,j} - \mu)(Y_{i+u,j+v} - \mu)\} = \sigma^2\rho(|u|,\ |v|) \qquad (4.1)$$

and $\rho(0,0)=1$. Here, the expectations (E) are taken under the model distribution. In other words, the population values $Y_{i,j}$ are assumed to arise as a realization of a spatial process, stationary in the wide sense with autocorrelation function ρ. We will assume that ρ is nonnegative and nonincreasing in both arguments.

A variety of natural processes with the above structure are examined by Matern (1960). Two examples are given by Das(1950),

$$\rho(|u|,|v|) = \sum_{j=1}^{G} a_j \exp(-\gamma_j|u| - \delta_j|v|), \qquad (4.2)$$

and by Matern (1947),

$$\rho(|u|,\ |v|) = \exp\{-\lambda(|u|^2 + |v|^2)^{1/2}\} \qquad (4.3)$$

Notice that isotropic correlation functions are of the form (4.1).

Quenouille (1949) introduced aligned and unaligned sample designs in the plane. A sample aligned in both directions, $s=s_1xs_2$ of size $n=n_1n_2$, is simply the cartesian product of a sample (s_1) of n_1 row labels and a sample (s_2) of n_2 column labels. Aligned simple random samples, stratified random samples and systematic samples are thus obtained when both

s_1, s_2 are selected with the respective one-dimensional sampling schemes. Of course different sampling methods may be used along the rows and the columns.

Unaligned simple random samples are formed by selecting $n=n_1n_2$ labels at random without replacement from the $N=N_1N_2$ labels. Unaligned stratified random samples are obtained by selecting at random without replacement within the defined rectangular strata. An unaligned systematic sample may be described as follows. Draw at random without replacement numbers

$a_1,\ldots,a_{n_2}$ between 1 and k_1, and numbers

$b_1,\ldots,b_{n_1}$ between 1 and k_2. The sample

consists of units labelled

$$\{(a_j + k_1(i-1),\ b_i + k_2(j-1)):i=1,\ldots,n_1;$$

$$j=1,\ldots,n_2\}.$$

Assuming model (4.1) with

$$\sum_u \sum_v |\rho(|u|, |v|)| < \infty,$$

it is shown in Iachan (1985b) that the limiting expected variances for the three sample designs in each class (aligned and unaligned) follow the hierarchy

$$\sigma^2_{sy(a)} \leq \sigma^2_{st(a)} \leq \sigma^2_{sr(a)}$$

and

$$\sigma^2_{sy(u)} \leq \sigma^2_{st(u)} \leq \sigma^2_{sr(u)}.$$

Hajek (1959) has shown the optimality of (linear) systematic sampling in the sense of minimum expected variance, among all designs with the same probabilities of inclusion. This result cannot be extended to plane sampling, since Bellhouse (1977) has shown the nonexistence of an optimal sampling design in two-dimensions for a general class of correlation functions. An optimum does exist, however, for a subfamily of the class of isotropic convex correlation functions (Dalenius et al., 1960). Under more restrictive assumptions, optimality of the appropriate systematic sampling designs is shown by Bellhouse (1977) in three different classes that include designs aligned in one or both directions, and unaligned designs.

These two-dimensional model-based results are useful in the design of surveys of spatially correlated (non-mobile) populations such as forest trees (Matern, 1960) or shellfish (Iachan, 1985a). As discussed in the next section, if the study (planar) population to be sampled in time also presents temporal autocorrelations, the models need to be extended to three dimensions.

Another type of two-dimensional population is obtained by adding a temporal dimension to the usual kind of (linear) finite population. In this case, the population has labels (i,t): i=1,...,N; t=1,...,T. The autocorrelation functions of interest have (one-dimensional) "marginals" $r_1(j)$ and $r_2(t)$ of the form considered in the previous section (non-increasing and non-negative).

Optimal sample designs for sampling in time and on the line will then be obtained by crossing the two "marginal" sample designs, i.e., will be in the aligned class. If both r_1 and r_2 satisfy some additional assumptions (summability or convexity), then aligned systematic sampling will be optimal.

This type of design has been used in asbestos, well-water and personal exposure studies, where either (a) a sample of units is selected independently for each time period, or (b) a time period is randomly assigned to each selected unit.

5. The three-dimensional case

As mentioned in the previous section, it is possible to generalize many of the earlier results to the three-dimensional case. The three-dimensional case arises in two distinct situations of particular concern:

(a) when the sampling units are distributed in the three-dimensional space, or

(b) when the sampling units are in the plane and have changing characteristics in time.

The first case may occur for a variety of natural populations: fish in a lake, birds or insects in the air, cells in a human organ, and so on. The types of autocorrelation functions expected in these cases are simple extensions of the functions considered in Section 4. The results obtained in Iachan (1985b) may thus be extended to this case. In particular, systematic sampling still outperforms (in the sense of smaller limiting expected variance) its two most common competitors.

The second case has greater practical importance. It seems reasonable to assume that the three-dimensional autocorrelation function, R(t,u,v), has marginals r(t) and R(u,v) satisfying the assumptions stated in Section 3 and Section 4, respectively. It is an area for further research to extend the previous results to this case.

Some comparisons have been performed by Kalsbeek (1988) when the focus is on the estimation of a mobile population size and no structure is imposed on the population. Using ANOVA-type decompositions and linear cost models, the paper investigated the efficiency of some sample designs in the two-dimensional classes discussed in Section 4.

6. Stratification

This section will illustrate several issues related to stratification when time plays a role in the design of the survey. The surveys described in Section 2 will be used as illustrative examples.

Stratification is inevitably connected to the choice of sampling unit. It may be difficult in practice to stratify sampling units defined by crossing units and time periods. For example,

electrical load surveys may consider sampling units of the form households x times. If seasonal strata are defined, this approach may involve frequent switching of households from stratum to stratum.

In a recent load research survey design, we have defined four basic strata based on summer and winter electrical usage:

(a) low-summer, low-winter;
(b) low-summer, high-winter;
(c) high-summer, low-winter;
(d) high-summer, high-winter.

Exhibit 1 illustrates how appliance use patterns are related to these four primary strata. These strata were further partitioned to yield substrata of approximately equal total (kWh) consumption.

Stratum boundaries in the two-dimensional plane were constructed based on summer and winter average daily (kWh) consumption as illustrated in Exhibit 2. Note that the two-dimensional nature of the stratification arises not strictly from the temporal component, but from the combined use of two stratification variables.

The two-dimensional stratification shown in Exhibit 2 is obtained by crossing the two marginal, one-dimensional stratifications. Other, more general methods of multivariate stratification are considered in Iachan (1985a). The methods include the possible use of one composite (one-dimensional) variable to reduce the problem's dimensionality. In the two-dimensional example discussed in the paper, ocean depth contours provide efficient strata for estimating shellfish abundance. In state radon surveys, a composite radon potential index has also been suggested for stratification (where strata are groups of counties).

Two-dimensional stratification is often encountered in national surveys where the U.S. territory is divided into compact regions. Regional strata may be formed by clustering geographic units that are not necessarily contiguous (Iachan, 1987). In indoor air and radon surveys, the temporal dimension plays a role through seasonal/climatic effects (Iachan, 1988). Regional strata for radon surveys should also take geological factors into consideration. As a result, strata for these surveys are typically non-compact, i.e., comprised of several disjoint areas.

References

Bellhouse, D.R. (1977). "Optimal designs for sampling in two dimensions." Biometrika 64, 605-611.

Dalenius, T., J. Hajek and S. Zubrzycki (1960). "On plane sampling and related geometrical problems." Proc. 4th Berkeley Symp. 1, 125-150.

Das, A.C., (1950). "Two-dimensional systematic sampling and the associated stratified and random sampling." Sankhyā 10, 95-108.

Hajek, J. (1959). "Optimum strategy and other problems in probability sampling." Casopis pro Pestovani Matematiky 84, 387-420.

Iachan, R. (1983). "An asymptotic theory of systematic sampling." Ann. Statist. 11, 959-69.

Iachan, R. (1985a). "Optimum Stratum Boundaries for Shellfish Surveys." Biometrics 41, 1053-1062.

Iachan, R. (1985b). "Plane Sampling." Statistics and Probability Letters 3, 151-159.

Iachan, R. (1987). "Clustering Procedures in Survey Design." Proceedings of the Social Statistics Section of the American Statistical Association, pp. 368-372.

Iachan, R. (1988). "Issues in Environmental Survey Design." Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 144-152.

Kalsbeek, W.D. (1988). "Design Strategies for Non-sedentary Populations." Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 28-37.

Matern, B. (1947). "Methods of estimating the accuracy of line and sample plot surveys." Meed. fr. Skogsforsknings Inst. 36, 1-138.

Matern, B. (1960). "Spatial variation." Meed. fr. Statens Skogsforsknings Inst.. 49, 1-149.

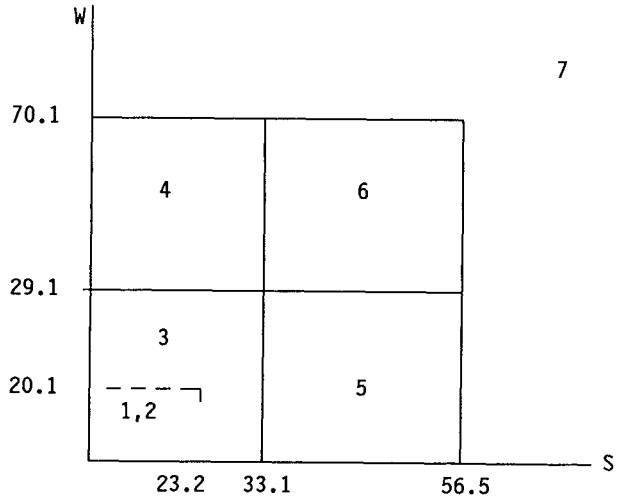
Quenouille, M.H. (1949). "Problems in plane sampling." Ann. Math. Statist. 20, 335-375.

Exhibit 1. End-use saturations in initial size strata for residential sample*

Appliance	Stratum S-Hi W-Hi	S-Lo W-Hi	S-Hi W-Lo	S-Lo W-Lo
A/C	H	L	H	L
Electric Heating	H	H	L	L
Electric Water Heater	H	H	L	L

* Saturations for each end-use in the table are labeled H(high) or L(low) according to whether they are above or below the overall (population) saturation. (Saturations are the percents of population units with the given appliance.)

Exhibit 2. Usage stratum boundaries based on (S,W) space for residential sample



* S = daily average Summer KWh consumption,

W = daily average Winter KWh consumption.

Stratum labels are:

1 = very small (gas),
2 = very small (non-gas),
3 = small (low/low),
4 = low summer/high winter,
5 = high summer/low winter,
6 = large (high/high),
7 = huge.