Survey Methodology, December 1990
Vol. 16, No. 2, pp. 181–194
Statistics Canada

Sample Maintenance Based on Peano Keys

KIRK M. WOLTER and RACHEL M. HARTER[1]

ABSTRACT

We discuss frame and sample maintenance issues that arise in recurring surveys. A new system is described that meets four objectives. Through time, it maintains (1) the geographical balance of a sample; (2) the sample size; (3) the unbiased character of estimators; and (4) the lack of distortion in estimated trends. The system is based upon the Peano key, which creates a fractal, space-filling curve. An example of the new system is presented using a national survey of establishments in the United States conducted by the A.C. Nielsen Company.

KEY WORDS: Recurring surveys; Sample maintenance; Changing population units; Peano key.

1. INTRODUCTION

We are concerned with recurring surveys conducted over time and the maintenance they require. Let $\mathcal{U}_t$ denote a survey universe at time t, with $t = 0$ denoting the inception of a new survey. We assume a probability sample of units of $\mathcal{U}_0$ has been selected, and thus that it is feasible to construct unbiased (or at least consistent) estimators of the population total and other parameters of interest. As time goes by, we assume the universe is surveyed repeatedly at regular intervals of time, in part to track the "level" of the population, and in part to measure its "trends". A panel or a rotation sampling design is usually employed for this purpose (*e.g.*, see Rao and Graham (1964) and Wolter (1979) and the references cited by those authors). In all such surveys of people or their institutions, which is all we concern ourselves with here, the composition of the universe changes with time as births, deaths, and other changes occur to the status of the units. The survey frame, the sampling design, and the schemes for observing or collecting the survey data must be maintained for such change; otherwise, the sample may become excessively biased and cease to be representative of the universe.

The types of maintenance issues that arise in recurring surveys depend in part on the kind of universe under study, in part on the choice of sampling unit, and in part on the interplay between the sampling unit and the universe elemental units. We shall summarize briefly the issues that arise in four different situations:

(i) establishment surveys with establishment as the sampling unit;
(ii) establishment surveys with company or some similar cluster of establishments as the sampling units;
(iii) surveys of people or households with the address or housing unit as the sampling unit; and
(iv) surveys of people or households with the household or family as the sampling unit.

In this work, we use the words "establishment" and "company" in a generic sense. An establishment may be a retail store, a manufacturing plant, a school, a hospital, a golf course, or any other similar, single-location entity, while the corresponding company would be the corporate, legal entity that owns the retail store, or the school district, and so on. In some cases, of course, the establishment and company will be synonymous, *e.g.*, a single, independent grocery store.

[1] Kirk M. Wolter and Rachel M. Harter, Statistical Research Department, A.C. Nielsen Company, Nielsen Plaza, Northbrook IL 60062, USA.

For case (i), the main universe dynamics include:

- establishments arising from new construction
- reclassified establishments from some out-of-scope category to an in-scope category
- reclassified establishments from one in-scope category to another in-scope category
- reclassified establishments from an in-scope category to an out-of-scope category
- conversion of a structure from residential use to commercial use
- conversion of a structure from commercial use to residential use
- demolition of an existing establishment
- establishment that moves in and out of vacancy status
- changes in the configuration of an establishment, *e.g.*, division into two or more establishments.

Case (ii) is far more complicated than case (i), principally because sampling units are now clusters of elemental units. All of the issues from case (i) apply to single-establishment companies. For multi-establishment companies, we face the following additional dynamics:

- mergers wherein two companies combine to form a new successor company
- acquisitions wherein one company is acquired by another, with the acquiring company as the sole successor company
- joint ventures wherein two companies collaborate to form a new company that may be a subsidiary to both the parent companies
- divestitures wherein a company spins off a new and independent company
- divestitures where a company sells parts of itself to another acquiring company.

In a sense, case (iii) is very similar to case (i) in respect to the kinds of universe dynamics that may arise:

- housing units arising from new construction
- reclassified housing units from some out-of-scope category to an in-scope category
- reclassified housing units from one in-scope category to another
- reclassified housing units from an in-scope category to an out-of-scope category
- conversions from residential to commercial
- conversions from commercial to residential
- demolition of an existing housing unit
- reconfigurations of existing structures, *e.g.*, reconfigurations of apartments within a small multiunit structure.

Note how closely these issues match those for case (i).

Finally, case (iv) is very similar to case (ii) in terms of the composition and complexity of universe change. Maintenance issues include:

- marriage, wherein a new successor family is created, possibly from whole predecessor families or from part families
- new members move into an existing family, either eliminating another family or part of a family
- divorce, wherein successor families may be created from one predecessor family
- family members move away, either to join another existing family or to establish a new family
- births of family members
- deaths of family members
- a whole family moves, thus requiring tracing and perhaps altering field-work assignments.

To handle the universe dynamics listed above, properly reflecting them in the sample, so that sample representativeness is retained over time, the survey organization must design and adopt an explicit system of maintenance. We define a *sample maintenance system* to be a sampling design and a universe updating methodology, possibly specified in the form of simple rules, that permit the statistician to achieve known, nonzero probabilities of inclusion for each of the elemental units in the population for each time period in the recurring survey, or failing that, to weight the survey data properly so as to achieve unbiased or consistent estimators of the population parameters of interest. From cases (i) through (iv) above, it is clear that a maintenance system must perform at least four functions:

- give new elemental units a known, nonzero probability of selection
- account properly for elemental units that may no longer exist in a substantive sense
- not give elemental units multiple chances of selection into the sample; otherwise, if multiple changes are given, the system must appropriately record this information so that adjustments may be made in the estimation procedures
- appropriately update the universe frame so as to facilitate and control the above activities.

A general and necessary rule of thumb for any sample maintenance system is that the system, or the rules that define the system, must treat symmetrically universe changes both within and outside of the sample. If a proposed maintenance rule violates this rule of thumb, then there is risk of bias in estimators of totals and other universe parameters to be estimated. For example, consider two rules that might be used for case (ii) for sampling new companies created as the result of a divestiture. One possibility is to declare the new companies part of the sample *if* their predecessor companies were part of the sample, and otherwise, if their predecessors were not part of the sample, to subject the new companies to a new round of sampling. This rule is seen to give the new companies multiple probabilities of selection, and thus may result in biased estimation unless appropriate adjustments are made in the estimation procedure. (The adjustments we have in mind are related to the multiplicity rules studied by Monroe Sirken (1970) and others.) A second possibility is to declare the new companies part of the sample *if and only if* their predecessor companies were part of the sample. Because this second rule treats symmetrically the universe changes both within and outside of the sample, it is seen to result in unbiased estimation for the survey parameters of interest.

In designing a sample maintenance system, the statistician must be guided not only by the statistical properties of the resulting estimators, but also by the cost, feasibility, and customer acceptance of alternative rules. Some rules may require additional data collection, thus entailing additional cost that must be planned from the inception of a new recurring survey. Certain applications may actually require that additional data be collected retrospectively. This may be impractical, or at the very least, may entail considerable nonsampling error, thus risking bias. Some rules may well be feasible and cost-effective, yet may not satisfy the requirements of the customers or users of the survey data.

Finally, we note that this problem of maintenance is neither new nor newly recognized; for example, maintenance systems have been in place for years in many of the major recurring surveys at Statistics Canada, the United States Bureau of the Census, and the A.C. Nielsen Company. Nevertheless, there is remarkably little literature on this subject. For brief discussions of some maintenance issues, see Wolter *et al.* (1976) for case (ii), Hanson (1978) for case (iii), and Ernst (1989) for case (iv). Also see the broad comments of Duncan and Kalton (1987) on household surveys and Colledge (1989) on business surveys.

In the balance of this article, we focus on case (i), where the establishment is both the sampling and elemental unit. This is the case we face in our establishments surveys at the A.C. Nielsen Company. Section 2 describes one of our major surveys, the Scantrack survey,

and the specific maintenance issues we face in that survey. We also describe some of the key objectives we had in designing a new maintenance system for this survey.

The new maintenance system is based upon a parameter known in mathematics as the Peano key, which creates a fractal, space-filling curve. The Peano key is defined in Section 3, where we also provide several graphical displays for illustration purposes. We close the article in Section 4 by describing the rules that implement our new maintenance system.

2. THE SCANTRACK SURVEY

The Nielsen companies provide information from several marketing surveys. The media surveys, such as Nielsen Television Index and Nielsen Station Index, are based on samples of either housing units or households. Surveys for the packaged goods industry, including Nielsen Food Index, Nielsen Drug Index, and Nielsen Scantrack United States (NSUS), are based on samples of stores. The Single Source service, which ties together consumer purchasing behavior with household television viewing and retail marketing support, is based on both household and store samples. Although sample maintenance is an important issue to each of these surveys, the present discussion will focus on our Scantrack sample of grocery supermarkets, which is the basis for the NSUS service. The Scantrack sample includes 3,000 supermarkets, stratified by 50 metropolitan markets and a remaining United States stratum. Within a market, the sample is further stratified by major chain organizations. The frame is ordered geographically, and a systematic sample is selection within each stratum to achieve proper socio-economic representation. This sample is also representative of store age, store size, and other factors associated with item sales. Although a geographically ordered systematic sample is exceedingly simple and straightforward, the choice of this sample design is justified based on years of experience, as well as the results of empirical studies in which various sample designs were tested on universe data.

Stores in the Scantrack sample are equipped with electronic scanners at the checkout, which read bar codes on packaged goods. Bar codes are called universal product codes or UPC's. When the item is scanned, the transaction is entered into the store's computer where the UPC is matched with the item's price. Each week, the sample stores provide us with total sales movement and price data for every item that is scanned in the store. Since a supermarket typically carries over 10,000 UPC's, we receive and process over 30 million observations per week.

In addition to scanner data, we obtain data on promotion conditions for the items in each of the sample stores, including whether an item was featured in a newspaper advertisement, store display, or store coupon. If an item was featured, we also know the type of newspaper advertisement used and the location of the display within the store.

NSUS reports include estimated sales totals for individual items and aggregates of items for each market and the total United States. A ratio estimator is used, with all-commodity volume as the auxiliary variable. All-commodity volume, or ACV, refers to total sales of all items in a store, usually on an annual basis. ACV tends to be highly correlated with sales of individual items. In addition, the NSUS reports include estimates of sales and sales rates by promotion condition and estimates of year-to-year sales trends.

Continuous maintenance is necessary for the Scantrack sample because the national supermarket universe of approximately 30,500 stores is not static. In a recent 12-month period, approximately 2,200 new supermarkets opened, and 2,450 existing stores went out of business. Another 170 stores were reclassified during the year. Reclassification can result from any of a number of changes. Some smaller grocery stores enter the Scantrack universe when their ACV's surpass the $2-million-per-year threshold which defines a supermarket. A store might

change name or location, or be expanded through remodeling. Some stores change to an extended or economy format, such as a superstore, warehouse store, or other nontraditional supermarket. In 1979, about 3,800 extended and economy stores accounted for 17% of total supermarket sales. By 1988, the number of extended and economy stores had grown to over 9,000, and they accounted for almost 50% of all supermarket sales (*Progressive Grocer* 1989). Sometimes, individual stores or entire chains are acquired by another organization affecting stratum definitions.

In addition to universe changes, missing or faulty data situations arise that require substitution of sample stores. Some selected sample stores do not scan, and some that do have incompatible scanning equipment. If a store is consistently unable to provide us with usable data, it must be dropped from the sample. Sometimes a request for a sample change within an organization comes from the chain itself. Occasionally, a retailer simply refuses to cooperate.

The principal objectives of our maintenance system for the Scantrack sample are:

(1) the sample should maintain geographic balance through time
(2) the system should maintain the sample size through time
(3) the sample should adhere to principles of probability sampling so as to avoid bias in estimators of total sales, and
(4) sample changes should not disturb excessively estimates of year-to-year trends.

Geographic balance is a proxy for socio-economic balance. Because different neighborhoods have different purchasing patterns, geographical balance is important to achieving an efficient sample design (*i.e.*, low sampling variability) over a wide range of products. Furthermore, geographic balance is an important factor in our customers' perception of an appropriate sample.

A sample size decrease would adversely affect the standard errors of the estimators, and a sample size increase would adversely affect our costs. Neither outcome is desirable. Furthermore, contracts with chain organizations specify sample sizes and cooperation payments, and any changes would have to be renegotiated. This too is undesirable.

All applications involving Scantrack data require efficient, unbiased estimators of total sales. Manufacturers and retailers need such data for everyday business decisions, such as how much to produce, how much to ship, how much to keep in inventory, and how to allocate store shelf space.

Clients also require reliable year-to-year trend information for managing their businesses. Trend estimates help manufacturers assess the overall health of their businesses. Both manufacturers and retailers benefit from knowing the longer-term performance of all major brands in all product categories.

We describe the maintenance system that has been developed to meet these objectives in section 4. But first, we describe a new geographic ordering scheme in section 3.

3. PEANO KEYS

The Peano key is a parameter that defines a certain fractal, space-filling curve. It provides a mapping from $\Re^2$ to $\Re^1$ such that points in $\Re^2$ or spatial objects can be arranged in a unique order (Peano order) on a list. In the application we have in mind, the spatial objects are sampling units, and the space $\Re^2$ is represented by earth's geographic coordinate system.

We obtain the Peano key by interleaving bits. See Peano (1908), Laurini (1987) and Saalfeld, Fifield, Broome and Meixler (1988). Let $X = X_k \ldots X_3 X_2 X_1$ and $Y = Y_k \ldots Y_3 Y_2 Y_1$ represent the longitude and latitude of an arbitrary point in k-digit binary form. Then, the corresponding Peano key is $P = X_k Y_k \ldots X_3 Y_3 X_2 Y_2 X_1 Y_1$. Also see figure 1 for an example for the case $k = 4$. Note how simple it is to calculate the value of P.

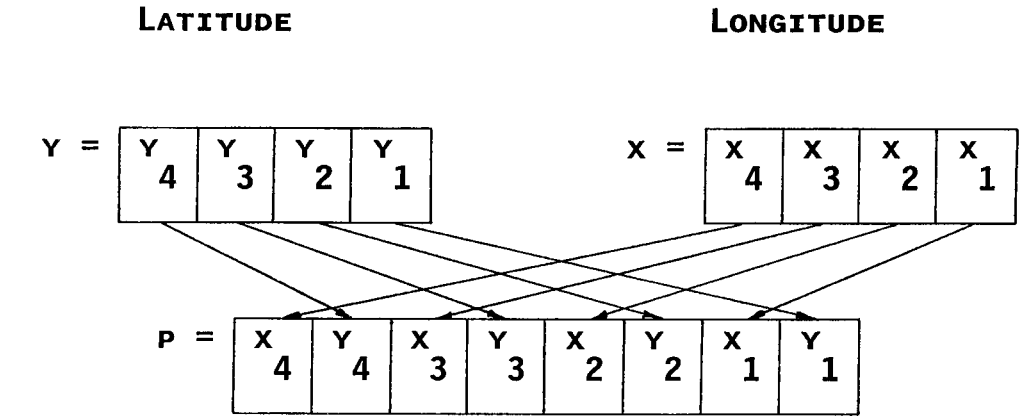

Figure 1. Creating the Peano Key by Bit Interleaving

Given k-digit (for any finite k) latitude longitude coordinates, the spacial "point" represented by the value of P is actually a square in $\Re^2$. As k increases, the sizes of the squares decrease. In fact, as k tends to infinity, the value of P will tend to represent a specific point in $\Re^2$.

The space-filling curve created by the values of the Peano key, P, is in the shape of a recursive N. Figure 2 illustrates the N-curve, using a grid of 1024 points. This figure displays the self-similarity feature of fractal images.

The N-curve passes once and only once through each point in space, points being defined as squares whose size is determined by the number of digits carried in the latitude and longitude coordinates. The order of points on the curve (Peano order) is largely preserving of geographic contiguity. Thus, Peano order facilitates proximity searches. Peano order involves a few geographic discontinuities, such as the jump from point 516 to point 517 in figure 2, as does any mapping from $\Re^2$ to $\Re^1$.

In the specific application we envision here, economic establishments are arranged on a list in Peano order by means of their latitude and longitude coordinates. Probability samples of the establishments may be drawn systematically from the ordered list. Because the earth's coordinate system is stable, there is no ambiguity in determining the list position of new establishments. Thus, they may be subjected to sampling too.

To illustrate this application, see figure 3 which displays a chain of retail establishments in the United States. Each establishment is described by a double-letter code. This code in natural lexicographic order signifies the Peano order of the establishments.

In the next section, we describe a sample maintenance system that is based upon the establishments' Peano order.

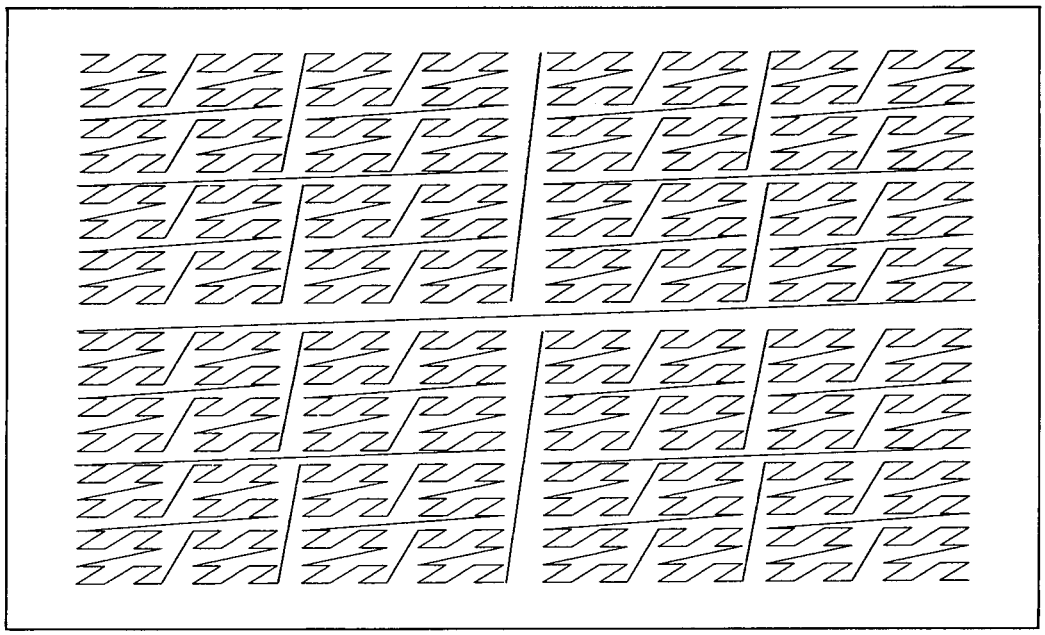

Figure 2. Peano Order Based on 1024 Points

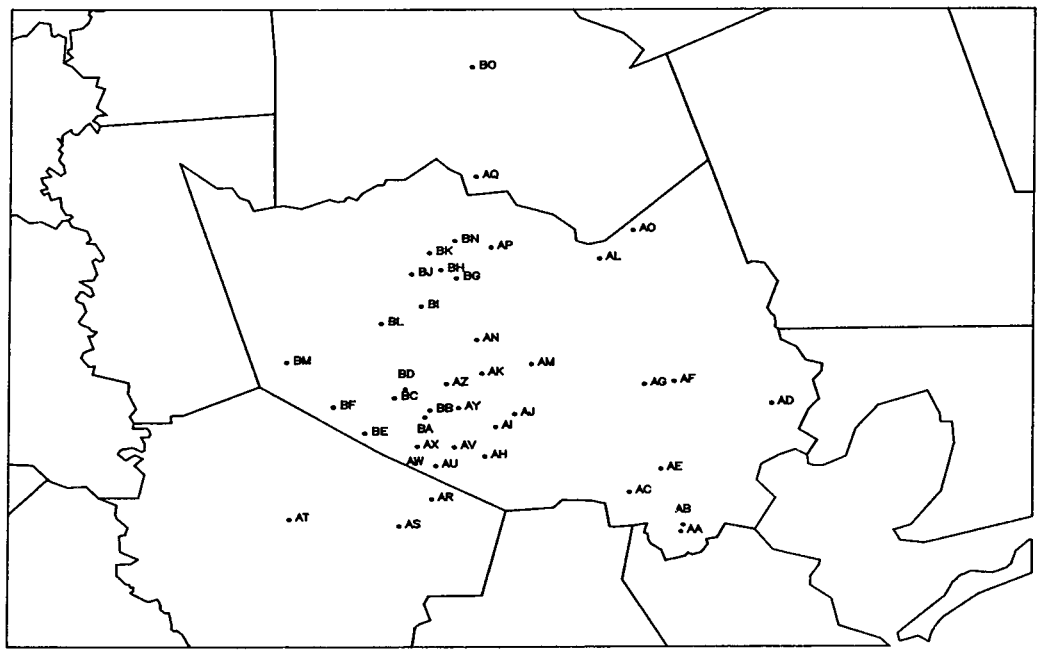

Figure 3. Chain of Retail Establishments in Peano Order

4. RULES FOR MAINTAINING THE SAMPLE

We describe a system for maintaining samples of retail stores, taking proper account of births, deaths, scanning conversions, and other changes in the status of the retail store universe. As stated earlier, we developed the system for applications at the A.C. Nielsen Company.

We consider a given and arbitrary sampling stratum, say of size N, and assume the universe of stores in the stratum is arranged in Peano order. For example, a stratum might include all stores in a given metropolitan market, such as Vancouver or Montreal. Ordering by Peano key values will turn out to be especially well-suited to the maintenance system that follows. Other ordering schemes may be considered for this work so long as they are stable across time and effectively map $\mathfrak{R}^2$ to $\mathfrak{R}^1$ in such fashion as to preserve geographic contiguity and to assign all birth stores a unique position in the ordering.

We assume an original sample is selected systematically with equal probability from the ordered list of stores at time $t = 0$. Let U_{ij} denote the j-th store in the i-th possible systematic sample, for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, where k is the sampling interval and n_i is the size of the i-th possible sample. If $N = nk + r$, $r < k$, then r samples will be of size $n_i = n + 1$ and $k - r$ samples of size $n_i = n$. In what follows, we shall also use the subscript "i" to represent the sample actually selected.

Let P_{ij} denote the Peano key value associated with U_{ij}. Let P_L and P_U denote the smallest and largest possible Peano key values within the market under study. Thus,

$$P_L \leq P_{11} < P_{21} < \ldots < P_{k1} < P_{12} < \ldots < P_{ij} < \ldots < P_{kn_k} \leq P_U.$$

Note that we are assuming each store possesses a unique geographic location and thus a unique Peano key value.

Let Y_{tij} denote the value of some characteristic of U_{ij} at time t. A standard unbiased estimator of the population total, Y_t, is

$$\hat{Y}_{ti} = k \sum_{j=1}^{n_i} y_{tij},$$

while the ratio estimator is given by

$$\hat{Y}_{Rti} = \hat{Y}_{ti} X_t / \hat{X}_{ti},$$

where the X-variable is a measure of size and X_t and $\hat{X}_{ti}$ are analogous to Y_t and $\hat{Y}_{ti}$, respectively.

Define N Peano key segments, S_{ij}, by partitioning the range $[P_L, P_U]$ at the N store values P_{ij}. We let $S_{ij} = [P_{ij}, P_{i+1,j})$, where it will be understood that $P_{k+1,j}$ represents $P_{1,j+1}$. A special definition is needed for the final segment. We define $S_{kn_k} = [P_{kn_k}, P_U] \cup [P_L, P_{11})$ so that the entire Peano range $[P_L, P_U]$ is covered by the N segments. This special definition, which treats the Peano range as if it were on a circle, is needed later to guarantee that all store births are given a nonzero probability of selection. Alternative segmentation schemes may be used without defeating the statistical properties of the maintenance system.

Our maintenance scheme is based upon the Peano key segments. The basic idea is to view the systematic selection process as applying to the segments, with subsampling of stores within the selected segments. Thus, as a formal matter, the segment is the primary sampling unit (PSU), not the store. Of course, as of the time of initial sample selection, there is, by construction, only one store per segment.

4.1 Birth Sampling

At a future point in time, say t', one or more new stores may open for business. Each new store will be assigned its unique Peano key value, and this value will be an element of one and only one Peano key segment. The Peano key permits us to automatically place new stores in their correct and unique positions one the ordered universe list.

The simplest possible rule for sampling births is the following:

> **Rule 1.** A birth store is selected into the sample if and only if its Peano key value is an element of a selected Peano key segment. Birth stores whose Peano key values are elements of nonselected segments are themselves not selected.

Given this rule, a birth store is selected with probability $1/k$. This occurs because its segment, which is unique, is selected with probability $1/k$. Unfortunately, Rule 1 does not provide good control of the sample size over time.

To control the sample size, we advocate some form of subsampling within PSU's. Let $U_{ij1}, U_{ij2}, \ldots, U_{ijB_{ij}}$ denote the stores in segment S_{ij}. The original store is now labeled U_{ij1}, whereas $U_{ij2}, U_{ij3}, \ldots, U_{ijB_{ij}}$ are the birth stores in Peano order. The number, $B_{ij} - 1$, of births in any given segment will be 0, 1,or 2 in most applications. Then we may subsample as described in the following alternative rule.

> **Rule 1A.** A birth store will be subjected to subsampling if and only if its Peano key value is an element of a selected Peano key segment. Associate with $U_{ij1}, U_{ij2}, \ldots, U_{ijB_{ij}}$ the probabilities $p_{ij1}, p_{ij2}, \ldots, p_{ijB_{ij}}$, where $p_{ijb} > 0$ and $\sum p_{ijb} = 1$. Now choose one of the stores according to this probability measure. Subsampling is independant from one selected segment to the next. Birth stores whose Peano key values are elements of nonselected segments are themselves not selected.

The probabilities in Rule 1A may be equal or unequal. If unequal, they may be defined in proportion to some preliminary measures of size, or defined so as to accelerate or retard the replacement of the sample.

We observe that our principal maintenance objectives are well-satisfied by Rule 1A. First, the rule maintains geographic balance over time because there is always one unit selected from each of the originally selected segments, which themselves were geographically balanced by virtue of the systematic sampling design. Second, the rule maintains a constant sample size over time because there is always one and only one store selected from each of the originally selected segments. Third, the rule is in accord with strict principles of probability sampling, whereby probabilities of inclusion are known and nonzero, and thus unbiased estimators of population totals are available. Finally, by appropriate choice of the p_{ijb}, we may control distortion in year-to-year trends.

The unconditional probabilities of selection are given by

$$\pi_{ijb} = k^{-1} p_{ijb}$$

for $b = 1, \ldots, B_{ij}$. That is, π_{ijb} is equal to the probability of selecting the PSU times the conditional probability of selecting the store, given the selected PSU.

Let $Y_{t'ijb}$ denote the value of the unit U_{ijb}, and let $Y_{t'ij+}$ denote the total for the (i,j)-th PSU. Then, the unbiased estimator of the population total $Y_{t'}$ is given by

$$\hat{Y}_{t'i} = \sum_{j=1}^{n_i} y_{t'ijb}/\pi_{ijb},$$

where $y_{t'ijb}$ is the value of the single unit selected from the (i,j)-th selected segment, with variance

$$\text{Var}\{\hat{Y}_{t'i}\} = \frac{1}{k}\sum_{i=1}^{k}\left(k\sum_{j=1}^{n_i} Y_{t'ij+} - Y_{t'}\right)^2 + k\sum_{i=1}^{k}\sum_{j=1}^{n_i}\sigma^2_{t'ij}, \tag{1}$$

where

$$\sigma^2_{t'ij} = \sum_{b=1}^{B_{ij}} p_{ijb}\left(\frac{Y_{t'ijb}}{p_{ijb}} - Y_{t'ij+}\right)^2.$$

The first term on the right side of (1) is the variance due to the sampling of segments. This is the original variance in the sense that it is the variance expression that applied at the time of original sample selection. The second term on the right side is the variance due to subsampling within segments. Note that $\sigma^2_{t'ij}$ vanishes for any segment in which birth subsampling has not occurred. Note also that the subsampling scheme achieves its minimum variance when, for each given i and j, the probabilities p_{ijb} are defined to be proportional to $Y_{t'ijb}$. In this case, the within component of variance vanishes. For any real application, however, this proportionality condition will be satisfied only approximately.

As usual, a first-order Taylor series approximation may be used to discover the variance of the ratio estimator. See Wolter (1986) for appropriate techniques to estimate the variance of both the unbiased estimator, $\hat{Y}_{t'i}$, and the ratio estimator $\hat{Y}_{Rt'i}$.

As time passes, it will be necessary to periodically update the sample to reflect additional births and other changes in the universe. It may be desirable to schedule the updating at regular intervals of time, so as to facilitate management of the work. We will refer to these intervals as update cycles. Such cycles may occur monthly, bimonthly, quarterly, or at whatever interval makes sense in a particular application. Factors to consider in establishing the frequency of the updating cycles include cost of the updating process; desired accuracy of the estimators of level and trend; and perceptions of the customers or users of the data.

Generally speaking, more frequent updating will cost more, achieve greater accuracy, and be perceived better by customers than less frequent updating.

For an update cycle at any future time t', Rules 1 or 1A may be used to maintain the sample. New stores are always placed automatically in their correct segment, by their Peano key values, and the subscript b reflects this order at each cycle. To explicitly reflect these ideas, we should have further subscripted the $U's, B's, p's$, and $\pi's$ by time, but we avoided doing so as a notational convenience. The expressions for the estimators of total, $\hat{Y}_{t'i}$ and $\hat{Y}_{Rt'i}$, and their variances remain valid for each t'.

4.2 Updating for Deaths

Rules for maintaining a sample over time must obey an important general principle. They must treat equally both selected and nonselected units. In the case of deaths, this principle implies that all deaths, both those in and out of the sample, must be handled in the same fashion in any sample updating process. If this principle is not followed, the resulting estimators will be biased, and the bias may accumulate over time.

In what follows, we describe procedures for death updating that follow this essential principle. There are two cases to consider: (i) deaths are not known on a universe basis, (ii) deaths are known on a universe basis.

For case (i), we suggest Rule 2.

Rule 2. All deaths in the sample will be known. They should remain in the sample but be set to zero (*i.e.*, $y = 0$) at the time of an update cycle.

This rule permits unbiased estimation of the universe population totals. Deaths cause the estimator variances to increase, and estimators of variance will properly reflect this increase, provided the deaths are retained in the sample with zero values.

For case (ii), we suggest Rule 3.

Rule 3. Remove all deaths from the universe at the time of the next update cycle. Subject only the remaining live cases to sampling, including births.

Rule 3 will cause the store count B_{ij} to change in segments where deaths have occurred, unless births exactly offset deaths. A replacement store will necessarily be selected within a given segment whenever the sample store from the segment has died -- except when there is a death but no birth and $B_{ij} = 0$ -- and a replacement store may be selected even when the sample store is alive and well.

In the exceptional case, where $B_{ij} = 0$, the sample size drops by 1. An interesting problem for future research is to investigate the mean square error of this rule versus that of an alternative rule which selects a replacement store from the same zone of k stores, instead of permitting the sample size to drop by 1. This alternative is conditionally unbiased but unconditionally biased.

Two additional issues must be addressed in handling deaths. The first issue concerns the coordination of birth and death updating. Store births and deaths will occur naturally at irregular intervals, depending upon business conditions and population growth. In some time periods, neither births nor deaths will occur. In other time periods, births may occur but not deaths, or vice versa. While in other periods, both deaths and births will occur. In theory, it would be possible to employ different update cycles for store births and deaths. For example, one might update bimonthly for both births and deaths, but in alternating months. This approach may have advantage in leveling the work load over time. On the other hand, alternating cycles may tend to defeat the ability of the sample to properly measure trends, creating a sawtooth pattern in the store time series as first births are introduced, then deaths dropped, then births, deaths, and so on. On balance, we recommend coincident sample updating for births and deaths so as to preserve trends.

The second issue concerns the handling of deaths during the period from their actual occurrence until the next update cycle. This issue arises only if the frequency of the updating process is less than that of the data-collection process. If the two processes are coincident, then there are no new problems. If updating is the less frequent, then there are two alternatives:

a) drop the deaths from the sample as soon as they are known to us (to be more precise statistically, this means the deaths are included in the sample with a value of zero)

b) continue the deaths in the sample by imputing for them until the time of the next update cycle.

Alternative a) is the simplest, cleanest way of proceeding. Aside from the problem of births, it is unbiased and permits correct variance estimators. Because of the birth problem, however, this alternative may have a negative effect on the ability of the sample to properly measure trends. As deaths occur during the first weeks of an update cycle, one can imagine a slight decline in the store time series, not because of fundamental change in economic conditions, but simply because the sample reflects deaths and not births. Alternative b) provides a short term fix to the problem of properly measuring trends. The essential notion here is that by imputing for

deaths, we implicitly make a correction for any births that have occurred since the last update cycle. This fix is not particularly elegant, and it is difficult to frame a rigorous, unassailable technical justification for it. On the other hand, history has shown that populations of economic establishments tend to be stable in the short run. Deaths are often associated with or are compensated by births, with the net size of the population remaining approximately level in the short run. The United States Bureau of the Census has used this alternative in its wholesale trade survey, with quarterly update cycles and monthly data collection. See Wolter *et al.* (1976).

4.3 Chronically Nonusable Stores or Scanning Conversions

In this final subsection, we present sample maintenance rules for handling stores that are chronically nonusable, such as stores that do not scan; do scan but with such poor discipline as to render their data faulty and nonusable; or refuse to participate in the survey. We shall explicitly discuss nonscanning stores and sample maintenance rules for handling conversions from nonscanning to scanning and vice versa, although the material that follows may be seen to apply more generally to all conditions of chronic nonusability. We shall let A denote the set of scanning stores and B the set of nonscanning stores, where $A \cup B$ spans the entire universe.

First, we treat conversions to scanning. There are two principal cases to consider: (i) scanning status is known for all stores prior to sampling; (ii) scanning status is not known prior to sampling, but is observed after sampling for the selected stores only.

Case (i) is relatively easy to handle. Here is a natural rule:

> **Rule 4.** Do not subject nonscanning stores B to sampling. Sample only from the subuniverse of scanning stores A. As a given nonscanning store converts to scanning, then treat it as a birth, subjecting it to birth sampling. Prior to conversion, nonscanning stores B shall be represented in the universe by utilizing imputation or other missing data techniques.

Given this rule and the prior data (*i.e.*, scanning status) it assumes, the entire survey budget may be allocated to the sample of scanning stores. None of the sample resources need to be committed to nonscanning stores.

To address case (ii), let s denote the selected sample of stores, and let $s_A = s \cap A$ and $s_B = s \cap B$. By assumption, s_A and s_B are not observed until after initial field work is completed. Obviously, all of these sets vary with time, but we suppress explicit time subscripts to simplify the notation.

Sample s_A should be maintained by rules presented elsewhere in this paper for births and deaths. New rules are required to handle s_B. Here is an illustrative rule that treats the stores in s_B as nonrespondents.

> **Rule 5.** At time t, impute for store $U_{ijb} \in s_B$ the value $\hat{y}_{tijb} = x_{tijb} y_{At}/x_{At}$, where x_{tijb} is the value of an auxiliary variable for store U_{ijb}, y_{At} is the sample s_A total for the estimation variable, and x_{At} is the corresponding total for the auxiliary variable. Alternatively, imputation may occur by means of substitution, hot deck/matching, or other means. Now, act as if the data set is complete, applying standard estimators of the survey parameters of interest. At the time U_{ijb} converts to scanning, it shall be deleted from s_B and joined to s_A, and the estimation shall still be performed by means of the standard estimators applied to the completed data set.

Given Rule 5, the effective sample size is reduced because of imputation variance associated with the $\hat{y}_{tijb}$. Substitution maintains a larger effective sample size than the other rules, but is clearly the most expensive to implement. All rules require limited field work on a continuous basis to monitor the scanning status of $U_{ijb} \in s_B$.

As an alternative to missing data techniques, we may observe the nonscanning stores using an alternative mode of data collection. Depending upon the data to be collected, this could involve a store audit or an interview conducted with store personnel by telephone, mail, or in person. This alternative would likely be more accurate than the imputation-based methods, yet additional cost and time may be involved, as well as burden associated with the management and control of two data collection methodologies.

Finally, we treat conversions of sample stores from scanning to nonscanning. Such conversions are likely to be relatively small in number and are treated here only for completeness. Let $U_{ijb} \in s_A$, *i.e.*, i is a scanning store in the sample. Note that U_{ijb} may be either a store that has scanned since being selected into the sample, or a store that converted to scanning after originally entering the sample as a nonscanner under Rule 5.

> **Rule 6.** At the time U_{ijb} converts to nonscanning, it shall be deleted from s_A, joined to s_B, and subsequently handled by missing data techniques, as in Rule 5. Standard formulae shall be applied to the completed data set. To simplify processing and field work, the method selected shall be identical to the method selected to handle conversions from nonscanning to scanning.

In the bizarre instance in which a store flip-flops repeatedly between scanning and nonscanning, one may handle the store by sequentially applying Rule 5 or 6, as the case may be, each time updating the sets s_A and s_B.

REFERENCES

COLLEDGE, M.J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley & Sons.

DUNCAN, G.J., and KALTON, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97-117.

ERNST, L. (1989). Weighting Issues for Longitudinal Household and Family Estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley & Sons.

HANSON, R.H. (1978). *The Current Population Survey: Design and Methodology, Technical Paper 40.* United States Bureau of the Census. Washington, DC.

LAURINI, R. (1987). Manipulation of Spatial Objects by a Peano Tuple Algebra, University of Maryland Technical Report CS-TR-1893, College Park, MD.

PEANO, G. (1908). La Curva di Peano nel Formulario Mathematico. In Opere Scelte di G. Peano, 115-116, Vol. I. Edizioni Cremonesi, Roma, 1957.

PROGRESSIVE GROCER (1989). 56th Annual Report of the Grocery Industry 1989, Vol. 68, No. 4, Part 2, Stamford CT.

RAO, J.N.K., and GRAHAM, J.R. (1964). Rotation Designs for Sampling on Repeated Occasions. *Journal of the American Statistical Association*, 59, 492-509.

SAALFELD, A., FIFIELD, S., BROOME, F., and MEIXLER, D. (1988). Area Sampling Strategies and Payoffs using Modern Geographic Information System Technology. Unpublished paper, United States Bureau of the Census, Washington, DC.

SIRKEN, M. (1970). Household Surveys with Multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

WOLTER, K.M. (1979). Composite Estimation in Finite Population. *Journal of the American Statistical Association*, 74, 604-613.

WOLTER, K.M. (1986). *Introduction to Variance Estimation*. New York: Springer Verlag.

WOLTER, K.M. *et al.* (1976). Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys. *Proceedings of the Business and Economic Statistics Section, American Statistical Association,* Alexandria, *VA*.