Using Sampling Theory as the Basis for a Conceptual Data Model

Fred C. Martin, Tonya Baggett, Tom Wolfe, and Roy Mita

Abstract.—Greater demands on forest resources require that larger amounts of information be readily available to decisionmakers. To provide more information faster, databases must be developed that are more comprehensive and easier to use. Data modeling is a process for building more complete and flexible databases by emphasizing fundamental relationships over existing or traditional business operations. Data modeling uses a hierarchical series of models beginning with a conceptual model of the activity of interest. From the conceptual model, a logical model is derived that captures more detail, but in an implementation-independent way. Finally, the logical model is transformed into a physical data model by means of application software. We show how sampling theory was used in a conceptual model to provide an integrating framework for identifying fundamental relationships. By using sampling theory, the final data structure organizes forest vegetation data gathering as a scientific process, rather than as specific business functions.

A data model is "a picture or description which depicts how data is to be arranged to serve a specific purpose" (Modell 1992). Without such a model, developers are prone to build data systems that incorporate existing relationships instead of more fundamental data relationships. The rapidity with which forest values are evolving and the complexity of forest resource data attest to the need for comprehensive data systems. The high cost of such systems motivates the use of modeling procedures and design concepts that promise longevity, flexibility, and stability. Although only one of several data management issues, the storage and availability of forest inventory and other vegetation measurements are crucial to the credibility of any management information system.

Social forces altering forest management activities are creating demand for resource information previously considered of little value. Small wood marketing requires information on ever smaller logs, while the resulting thinned stands prompt re-inventory of conditions largely ignored in the past. Forest practice laws requiring retention of residual live and dead trees create a need to "track" relicts and to develop inventories on dead woody materials. In addition, research continues to highlight dimensions of the forest not traditionally examined; recent examples include managing woody debris in forests (Graham *et al.* 1994), the need for risk assessment of refugia-like forests (Camp *et al.* 1997), and recognition of highly diverse forest structures important for endangered species (Everett *et al.* 1997). These events have led to expanded inventories that include new elements and characteristics such as snags, dead woody material, stumps, platforms and cavities, understory plants, canopy structure, and successional status.

Because of the extensiveness of forest vegetation data and the need for user support, larger integrated data structures are being promoted. Multiple small databases are simpler, but each system is supported by fewer people, while larger data systems allow redundancy in support personnel. A tradeoff is made for larger support capability at the expense of greater complexity. The difficulty of keeping separate systems functioning with changing events and technology is also seen as a benefit of a larger integrated system, e.g., year 2K transition, distributed processing, etc.

The purpose of this paper is to show how sampling theory was used to guide the development of a forest vegetation database. Sampling theory aided in the integration of disparate business operations into a single structure and provided principles for evaluating data model logic. We believe that sampling theory provides the foundation for more stable and longer lived data structures.

STEPS IN DATA MODELING

A basic issue in database development is how to deal with the sheer volume and complexity of data generated by our activities. This is not a trivial issue; how one chooses to organize data for storage has a profound bearing on the

Forest Biometrician, Washington Department of Natural Resources, Olympia, WA, USA; Computer Programmer, USDA Forest Service Forest Management Service Center, Ft. Collins, CO, USA; Systems Analyst, Washington Department of Natural Resources, Olympia, WA, USA; Operations Research Analyst, USDA Forest Service Forest Management Service Center, Ft. Collins, CO, USA.

subsequent ability to retrieve and use it. History, even recent history, is filled with examples of data structures that could not pass the test of daily use. How does one organize data so that they can be readily stored, retrieved, and, most important, used? How can different organization units share the same data? How can redundancy and its consequent implications for rising maintenance costs and introducing errors be reduced? Lastly, how can all this be done and still maintain a level of flexibility, since processes and needs change over time?

One approach is data modeling. In an earlier age, data models were simple and largely intuitive. However, as data and the demands of end users have increased, both the building process and the model itself have become increasingly complex. The payoff to thorough data modeling is a system that is comprehensive, flexible, and reliable.

A precursor to data modeling is business function modeling, which begins by reviewing business practices and products to determine business requirements for a data structure (Barker and Longman 1992, Baskerville and Moore 1988). This process progresses through a series of steps to identify the hierarchy of functions executed by an enterprise. High-level forestry business practices include inventory, timber cruising, land appraisal and exchange, experimentation, monitoring, regeneration certification, and timber sale compliance surveys. The process of examining different functional levels ultimately leads to identifying elementary business functions. Examples of elementary business functions (tasks that once started must be completed entirely to be useful) are inventory projections, measurement of vegetation data (sampling events), calculation of population parameters, and outputting of tree lists for further modeling. An important result of business function modeling is identification of function commonality. Commonality occurs when data can be shared between more than one business function, e.g., collected tree data are used to estimate parameters needed for timber sale, land exchange, habitat assessment, etc. Identifying relationships and dependencies between different functions is the basis for model building and is crucial to eliminating data redundancies. Examination of the relationships between elementary, common, and dependent business functions revealed sampling to be an integral part of all elementary functions and that all high-level functions relied on population parameter estimates.

The process of data modeling follows business function modeling. Data modeling designs a database using a series of related hierarchical models (Weldon 1997). The first modeling level is a conceptual model of the primary business activity to be supported; the business activity considered here is sampling. Based on the conceptual model, a logical model is developed that captures specific data items and relationships in a logical but application-independent way. Lastly, a physical model is constructed that implements the data and relationships of the logical model using a specific database management system (DBMS), such as Oracle. Limitations on the physical implementation may arise because the DBMS is unable to achieve all of the relationships identified in the logical model, or they may arise from basic business constraints such as limited computer capabilities or staff expertise.

Recognition that sampling data were being stored and manipulated confounded traditional database expectations—a paradigm shift was required. In a conventional database system, each "record" (observation, instance) is important in its own right, i.e., assumed a true population parameter known without error. A business payroll database lists every individual, their position and salary; no variance or sample error is considered or allowed. Each record represents only itself and every record is critical. But, when dealing with sample data, the interest is in parameters of sampling distributions that are related to a fundamental probability set (or population) and to sample size (O'Regan and Palley 1965). The identity of an individual element is less important than its contribution to a parameter estimate. For example, line-intersect sampling of a piece of woody material estimates volume per unit area, not volume of the observed piece; a regeneration survey returns a stocking sufficiency value and the identity of any particular seedling is insignificant; and the presence of a tree on a variable-radius plot estimates basal area per unit area and individual tree dimensions are less important. The expected value of a *population* is the information of primary interest, not the value of a single object. The population estimate is more important than the sampled element.

In a "sample" data system, information does not equal data. Data are facts used to *infer* information; information is knowledge about a population of interest. Sample data are the measurements about a population used to *derive* information. Rarely are sampled data the principal objects of concern; rather, they are a means to an end. This paradigm shift required that sampling and data modeling requirements be integrated. Important requirements in sampling consist of defining objectives, identifying populations, selecting a sample, making measurements, and estimating parameters. Important data modeling requirements are capturing fundamental versus existing relationships, eliminating redundancy and enhancing data integrity. Examples of fundamental relationships in forest measurements are recognizing that dbh is both a diameter *and* a height measurement, that crown ratio is just a ratio of height measurements, and that Girard form class is an arbitrary stem form measurement. Eliminating redundancy improves the quality of both data and processes, reducing problems with updates, and seeking a single "best" data storage location. Good

modeling designs integrity into the data model rather than imposing integrity through external processes. Relationships between basic entities can enforce integrity, eliminating multiple code checks and redundant attributes. For example, by including event entities in the model, knowledge and counts of occurrences (such as plot "taking") are explicit without the need for additional attribute counters.

CONCEPTUAL DATA MODEL DESCRIPTION

Cochran (1977) listed 11 principal steps in a sample survey. Steps relevant to data model design are the objectives of the survey, population to be sampled, data to be collected, degree of precision desired, methods of measurement, the frame, selection of the sample, and summary and analysis of the data. These steps were redefined into eight conceptual data model entities (fig. 1).

1. Sample design
2. Population rule
3. Element selection rule
4. Characteristic measurement rule
5. Conceptual population
6. Sample event
7. Sample element
8. Measured characteristic

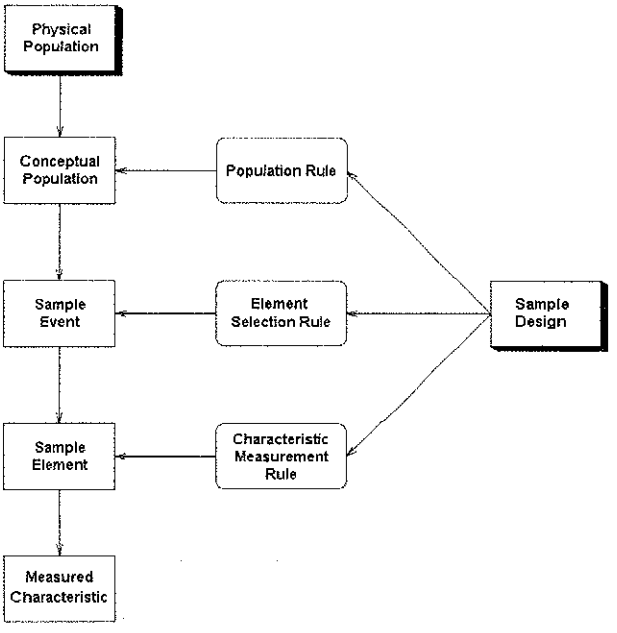

Figure 1.—*Conceptual model entities and relationships for a sampling survey database.*

Building the conceptual model involved defining the entities, recognizing their relationships, and identifying instances of each entity type. Entity definitions were influenced by previous work of Palley and O'Regan (1961), O'Regan and Palley (1965), and Byrne and Stage (1988).

The sample design, population rule, element selection rule, and characteristic measurement rule entities (fig. 1) together form the protocol (metadata) for describing instances of sample events, sample elements, and measured characteristics. A sample design entity specifies the purpose and type of survey, the parameters to be estimated, and the number of conceptual population levels prescribed for the survey. Different types of surveys lead to different conceptual population hierarchies, e.g., a simple stand inventory has fewer population levels than a stratified multistage design. Conceptual populations originate from physical populations by application of population and element selection rules creating "sampled" populations for a specific design. Physical populations are the set of physical objects about which we desire information; it is the "target" population, such as all the trees in a stand or in a forest. Sometimes the target and sampled populations are the same, but often there are "gaps" between them. A physical population can be sampled with many different designs, each design creating a distinct set of conceptual populations. For example, a watershed may be intensively sampled using a "wall-to-wall" stand-level inventory design, or it may be sampled as part of a stratified extensive sample. Each design defines a different set of conceptual populations from the same physical population. A sample design protocol is static over time; changes to any of the protocol entities create a new design. Relationships between the design protocol entities and the instances of sample events, elements, and characteristics ensure integrity between the "why, where, and how" of a survey and the "what and when" of data collection.

Population rules establish the hierarchy of nonoverlapping conceptual population levels for a design (fig. 2). The rule describes the type and form of conceptual populations at each hierarchical level. For example, in stratified sampling, the population as a whole is the top-level population, strata compose the second level, and sample point locations form the lowest level. The top-level physical population might be a forested region, while the top-level conceptual population consists of all lands within mapped strata, where each stratum consisted of units greater than some minimum mapping-unit size. Likewise, an experiment might be composed of blocks, containing replicates of treatment plots. An experiment top-level conceptual population could comprise the totality of elements that might be treated in some fashion, as in a random effects model. The blocks may occupy physical areas, such as harvest units, while conceptually

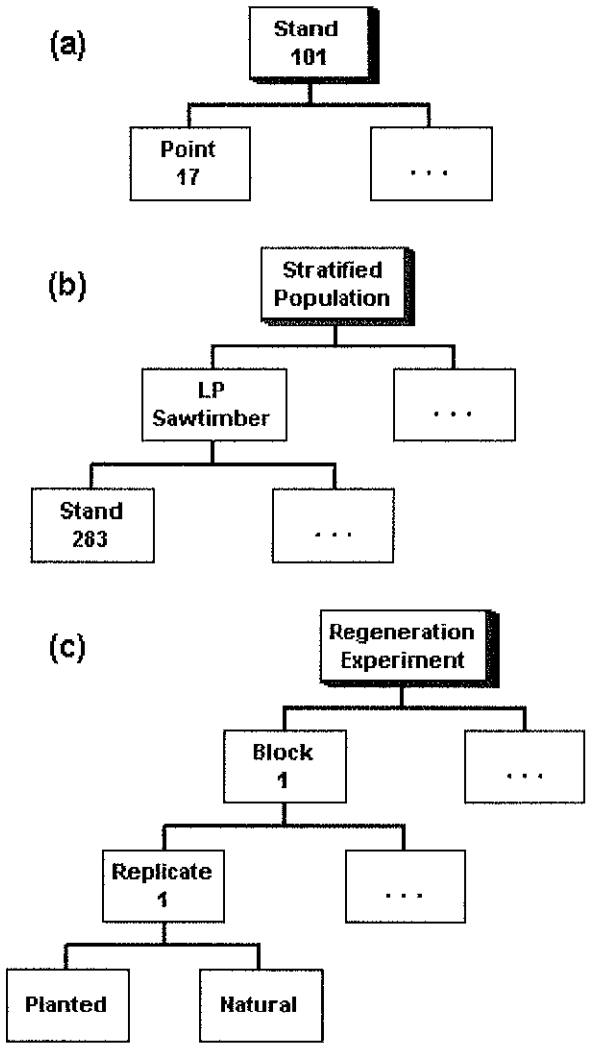

Figure 2.—*Schematic examples of conceptual population hierarchies for: (a) a single stand inventory, (b) a stratified stand inventory, and (c) a randomized complete block experiment.*

they represent a random sample of a harvest method. Similarly, a timber sale might be described by two population levels, a single harvest unit and a set of sample points; or it may require three levels—a top-level population for the entire sale, multiple second-level timber-harvest units, and sets of third-level sample point populations within each harvest unit. Besides defining the sampling frame, population rules also specify methods for determining probabilities (weights) at each hierarchical level. Conceptual populations may be spatial objects comparable to physical populations (polygons, lines, or points), they may encompass multiple physical locations, or they may represent theoretical populations. Further, for multiple resource inventories, a conceptual population may include several different types of elements; in essence, the population becomes a superset of different types of elements, each of which may be separately sampled from the population using different element selection rules.

Element selection rules define for each population level the types of elements to be sampled, the methods by which elements are selected for sampling, and criteria for creating subpopulations of elements. A population level may be sampled with many different selection rules, but a selection rule is unique for each population level; the same rule cannot be applied to the same set of elements at different population levels. This ensures that double counting does not occur. Further, within a sample design, element selection rules must be mutually exclusive so that the same individual element cannot be sampled with more than one rule. However, the same element may occur in two populations at the same hierarchical level, e.g., clustered variable-radius plots in which the same tree is observed at two or more points. The element selection rule also specifies the method for identifying which instance of an element is selected for sampling (e.g., fixed-area plots, horizontal point sampling, line intercept, 3P, census, etc.). Finally, the selection rule establishes criteria for creating subpopulations of elements (e.g., large trees versus small trees), each subpopulation of elements associated with a different sample selection rule. Examples of element selection rules are illustrated for a hypothetical "new forestry" timber sale cruise. Three conceptual population levels are defined: the harvest unit, several strips, and a systematic grid of points. For the top-level harvest unit population, select all large (>32-inch dbh), live cedar trees using 3P (rule 1) and select all large snags (>20-inch dbh) (rule 2). Within each strip (population level two), select all sawtimber-size live trees, except large cedars (rule 3). At each systematic point, select pieces of dead down wood with a 50-foot line intercept transect (rule 4), select understory vegetation groups (forbs, grasses, and shrubs) using a 1-m square grid (rule 5), and select all small snags (> 10 feet tall and < 20-inch dbh) using a variable-radius plot (rule 6). Over time the method appropriate for a specific element may change as the element moves from one subpopulation of elements to another. This causes no problem because sample event integrity is maintained (see below). The element selection rule is similar in concept to that described by Byrne and Stage (1988), but unlike Byrne and Stage, an element cannot be sampled with more than one selection rule in a given sample design. A change in the element selection rule results in a change in sample design and creation of different conceptual populations. However, this does not preclude evaluating sample results from multiple conceptual populations created from the same physical population. Such evaluations are effected by the relationship between conceptual and physical populations rather than through "linking variables" applied to individual measurements, as in the Byrne and Stage design.

A characteristic measurement rule controls which characteristics of elements are measured and the measurement protocols. Measurement controls include the order that characteristics are measured, criteria for deciding if a characteristic is to be measured, and the probability and minimum frequency of measurement. The order that characteristics are measured, along with measurement criteria, can be used to control which characteristics of an element are to be measured. For example, tree age might be observed if measurements of species, tree size, and damage were all within specified bounds, i.e., a site tree. Subsampling of characteristics is provided by specifying both a probability and a minimum frequency of measurement; e.g., subsampling of tree heights might specify measuring 25 percent of heights with a minimum of four heights per species. Measurement protocols include units of measure, acceptable measurement procedures and devices, measurement resolution, and legitimate values. Including units of measure within rules, rather than as attributes of actual measurements, reduces redundancy and enhances data integrity. Specification of procedure, device, and resolution attributes help document data quality. Legitimate value attributes provide filters to omit recording of extreme characteristic values, e.g., excessive ages or minor damages. Storing legitimate values (particularly for categorical or class values) in the design protocol allows new designs to evolve while protecting data integrity. For example, if class values change over time (e.g., tree damage codes or structural stage classes), originally recorded values are retained but made equivalent to current values through translation tables, a process known as image journaling.

Application of the population rules to physical populations creates specific instances of conceptual populations. These instances are recursive within the defined population hierarchy, forming a parent-child relationship. Information for higher level populations can be inferred from information at lower levels, while lower levels "inherit" characteristics from higher levels. Usually an instance of a conceptual population is defined on a physical population identifiable by a closed polygon, line, or point. However, as previously mentioned, some populations may not be directly related to physically identifiable populations, such as in an experiment. Two important population attributes are the area (when it exists) and the sampling weight. Sampling weight is especially important because different populations may be sampled with different probabilities, as in stratified random sampling. Keeping sampling weights with the population instance instead of with the population rule allows greater flexibility for applying the same design protocol to different physical populations. Assemblages of populations within levels and between levels with their appropriate weights permit estimating parameters for either *a priori* or *posteriori* populations.

A sample event entity is the application of an element selection rule to a conceptual population level at a specific time. The sample event is a pivotal entity for data modeling. It uniquely relates measurements of elements to a population for a point in time. The sample event "intersects" an element selection rule entity and a conceptual population entity bringing into being a cluster of elements (trees, dead wood, shrubs, etc.) with known probability of selection. A collection of sample events for a single type of element at a single point in time appears equivalent to the definition given to a "conceptual population" by Palley and O'Regan (1961) and O'Regan and Palley (1965). We use the term more broadly; a conceptual population can be a collection of sample events for a single element, or a collection of sample events for multiple element types, or even for a single sample event. The extent of elements encompassed by a conceptual population depends on the population level and the sample events associated with it. Different sampling events arise from application of different sampling rules to the same conceptual population, even when applied at the same time. However, sampling the same potential set of elements using a different element selection rule would constitute an entirely different sample design protocol, leading to a new set of conceptual populations and a new set of sample events. This condition, or constraint, maintains integrity between conceptual population weights and sample events within the same sample design protocol. For example, sampling trees in a stand using both variable-radius and fixed-radius element selection rules would require two separate design protocols resulting in two separate conceptual populations and two separate sample events, all for the same physical population. An important attribute of a sample event is the probability associated with the selected element(s). Although the element sample rule defines the basic sampling method, the precise probability of an element may not be known until or even after the time of element selection, e.g., 3P sampling and cluster sampling across type boundaries. The count of sample events from a single time for a given population level is equal to the sample size from that population; an additional sample size attribute is unnecessary. Assembling sample events for a given time is the first step in deriving parameter estimates for a population level.

Sample elements are the basic units of a population on which characteristic measurements are made. Element types defined to date are: (1) an individual tree; (2) a group of trees; (3) a piece of woody material; (4) an individual forb, fern, grass, or shrub; (5) a group of forbs, ferns, grasses, or shrubs; and (6) the surface of the land. Although elements may occasionally mimic sample units, such as in 3P sampling or single stage cluster sampling (Shiver and Borders 1996), they are always related to some population level through the sample event entity.

Instances of elements become part of a conceptual population sample by the application of an element selection rule. For example, a tree is included in a conceptual population if its distance from a point is within some proportion of its squared-diameter when applying a horizontal point selection rule; likewise, some of the land surface will be included in a conceptual population if it falls with the area of a fixed-radius plot. Each element is considered to have an area of influence; the conceptual population establishes a location from which an element's influence area is appraised using an element selection rule (Stage and Rennie 1994). An element can be selected using only one sample selection rule at a single time, but the same element can occur in several different conceptual populations, e.g., the same tree selected on two different variable-radius plots. Again, the marginal importance of a specific element is noted; the requisite information is provided by parameter estimates.

A characteristic is a quality or feature of an element that can be measured or assessed. Examples of quantitative characteristics include tree height, canopy cover of a group of trees, the aspect of the land surface, the cover of grasses, and the height of a shrub. Qualitative characteristics include tree crown class, tree group structural stage, decay class of a tree or piece of woody material, and land surface habitat type. Characteristics are the typical attributes found in forest databases, and their measured or assessed values are the data. Criteria for deciding which characteristics to measure and the probabilities and protocol for measurement are regulated by the measurement rule. A characteristic may be measured several times on the same element, e.g., multiple values of tree dbh using calipers or multiple damage agent classes for a group of trees. The number of times that characteristic values can be recorded for an element is determined by the final database structure.

INFORMATION GENERATION

Generating information is the *process* of assembling sample data and calculating population parameter estimates. Parameter calculation depends on the sample design including the probabilities associated with a measured element and its associated population level, as in stratified sampling. The process assembles element characteristic values from a set of sample events for a conceptual population and "expands" these values into a parameter estimate for the appropriate population. This assembly process may be iterative, estimating parameters at several levels of conceptual populations.

Parameter estimates need not be included in the database, but may be available on-demand as "views" or generated reports. Sometimes, parameter estimates will be generated and then used to "populate" attributes in related databases, such as a stocking table in a silvicultural database, a volume on a timber sale offering web page, or a listing of structural stages in a habitat monitoring system.

CONCEPTUAL MODEL APPLICATION

The concepts described have been used to build two different logical models (FSVEG of the USDA Forest Service and FRIS II of the Washington Department of Natural Resources) and one physical model (FSVEG). Although the two logical models differ in appearance, both implement a data-storage logic based on sampling fundamentals. However, neither model completely addresses all of the data management issues surrounding forest vegetation sampling due to the complexities of defining and translating sampling theory into physical entities and relationships. We continue to rely on separate supporting reports and maps to explain the full breadth of our sampling surveys.

ACKNOWLEDGMENT

Dr. James W. Flewelling, Biometrics Consultant, Kent, WA, USA, reviewed this manuscript.

LITERATURE CITED

Barker, R.; Longman, C. 1992. CASE*METHOD function and process modeling. New York: Addison-Wesley Publishing Co. 386 p.

Baskerville, G.; Moore, T. 1988. Forest information systems that really work. Forestry Chronicle. 64(4): 136-140.

Byrne, J.C.; Stage, A.R. 1988. A data structure for describing sampling designs to aid compilation of stand attributes. Gen. Tech. Rep. INT-247. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 20 p.

Camp, A.; Oliver, C.; Hessburg, P.; Everett, R. 1997. Predicting late-successional fire refugia pre-dating European settlement in the Wenatchee Mountains. Forest Ecology and Management. 94: 63-77.

Cochran, W.G. 1977. Sampling techniques, 3rd ed. New York: John Wiley & Sons. 428 p.

Everett, R.; Schellhass, D.; Spurbeck, D.; Ohlson, P.; Keenum, D.; Anderson, T. 1997. Structure of northern spotted owl nest stands and their historical conditions on the eastern slope of the Pacific Northwest Cascades, USDA. Forest Ecology and Management. 94: 1-14.

Graham, R.T.; Harvey, A.E.; Jurgensen, M.F.; Jain, T.B.; Tonn, J.R.; Page-Dumroese, D.S. 1994. Managing coarse woody debris in forests of the Rocky Mountains. Res. Pap. INT-RP-477. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 12 p.

Modell, M.E. 1992. Data analysis, data modeling, and classification. New York: McGraw-Hill, Inc.

O'Regan, W.G.; Palley, M.N. 1965. A computer technique for the study of forest sampling methods. Forest Science. 11(1): 99-114.

Palley, M.N; O'Regan, W.G. 1965. A computer technique for the study of forest sampling methods. I. Point sampling compared with line sampling. Forest Science. 7(3): 282-294.

Shiver, B.D.; Borders, B.E. 1996. Sampling techniques for forest resource inventory. New York: John Wiley & Sons, Inc. 356 p.

Stage, A.R.; Rennie, J.C. 1994. Fixed-radius plots or variable-radius plots? Designing effective inventory. Journal of Forestry. 92(12): 20-24.

Weldon, J.L. 1997. A career in data modeling. Byte 22(6): 103-106.