PART II: PHILOSOPHICAL FRAMEWORK

2

Getting into Data: Philosophy and Tactics for the Analysis of Complex Data Structures

CHRISTOPHER CARR

In Chapter 1, I argued that a quantitative analysis can produce meaningful results only when two conditions are met. First, there must be logical concordance between the assumptions that a chosen technique makes about which aspects of a data set's structure reflect the phenomenon of interest, and those aspects that actually are relevant in this way. This concordance ensures the accurate *representation* of the data. Second, there must be logical concordance between the theoretical framework that guides or emerges from analysis, and the assumptions that underlie the chosen analytic technique. This is necessary for appropriate *meaning to be assigned* to the analytically derived representation of the data.

Discordance between technical assumption and a data set's structure can be of two kinds. First, the model that underlies the statistical technique to be employed may require data that pertains to a single process or parallel, coterminous processes that define a single population, whereas the data to be analyzed may actually reflect the effects of multiple processes that define multiple populations. This discordance usually arises when the variables and observations that are brought forward for analysis are defined too broadly and reflect a general problem area rather than a specific phenomenon of interest. The relevant and irrelevant variables, dimensions, and observations that are suggested by this discordance define the data's relevant and irrelevant *subset structures*. Second, the aspects of the data's structure that reflect the phenomenon (e.g., the scale of measurement of relevant relationships between

The ideas in this chapter stem from conversations I have had with W. Fredrick Limp, Michael Schiffer, Robert Whallon, David Braun, Dwight Read, and many graduate students at the University of Arkansas over the past five years. For their stimulation, for the way they have enriched my life, I am most grateful. I also wish to thank David Braun, Dwight Read, and Michael Schiffer for a number of excellent comments that helped me in the revision of this chapter.

variables, the monothetic or polythetic nature of relevant relationships among observations) may not be those which are assumed relevant by the analytic technique and to which the technique is sensitive. This discordance often arises when there is inconsistency between the expected nature of organization of the phenomenon of interest (used as a basis for choosing the technique) and the actual nature of its organization. It suggests the distinction between the data's relevant and irrelevant *relational structures*.

In this chapter, the cause of such discordances between data structure, technical assumption, and theoretical framework is examined in philosophical terms. The cause is made clear by comparing the logical processes that are involved in analyzing *complex* data sets, which requires mathematical pattern recognition procedures, to those that are involved in analyzing *simple* data sets, which requires only mental pattern recognition abilities.

The complexity of a data set is defined here to be a function of its size (number of variables and observations), the number and complexity of patterns within it, and the strength of patterning. A complex data set can be either of two kinds. The first is a *multivariate* data set in which the number of variables, observations, and patterns among them are too large for the patterns and meaning of variation among data items to be assessed mentally. A data set having a structure resolvable only by principal components analysis would be an example. The second kind of complex data set is a *univariate* response to multiple factors, which forms a time or space series that is too complicated for mental dissection. To investigate a data set having this structure, Fourier analysis, spatial filtering techniques, or time series analysis would be required (Carr, 1982b, in press). In contrast to these complex data sets, a *simple* data set is taken to be one having a very limited, mentally manageable number of variables and observations that exhibit uncomplicated patterning.

It is important to emphasize that the dichotomy drawn here between complex and simple data sets pertains to only those of their characteristics that determine whether mathematical procedures are necessary to *recognize patterns* within them, as just described. The distinction is not used to refer to other data characteristics, which determine the *ease with which appropriate meaning is assignable* to the patterns found within them. These additional characteristics include, for example, the number and complexity of assumptions involved in collecting, subsampling, and/or screening the data. In regard to these kinds of characteristics, all data sets are more or less complex (Schiffer, personal communication, 1983).

Once the cause of the discordances that can occur between data structure, technique, and theoretical framework during the analysis of complex data structures has been clarified, the occasions when such discordances arise will be discussed. Although such problems can potentially occur at all stages of analysis, they are particularly notable in the initial stages when little is known about the relevant aspects of a data set's structure. At this time, the researcher may find himself in a *methodological double bind*: he cannot choose an appropriate

technique of analysis and an appropriate subset of the data for analysis without some knowledge about the data set's relevant relational and subset structures; yet he cannot obtain this knowledge without applying some pattern searching technique to summarize the data's structure in a simpler form that is more comprehendable by the human mind. In short, the researcher has a problem of "getting into the data."

Finally, several solutions to the problem of getting into data and maintaining logical consistency in analysis thereafter are offered. The solutions—some standard, others not—are complementary. None is completely adequate. Each, however, focuses on precise *specification* of the phenomenon of interest and its nature, and explicit *justification* of the variables, observations, and techniques to be used in analysis relative to the phenomenon and its nature. As a consequence, they facilitate the formalization of bridging arguments, which characterize a mature discipline, and the development of theory; they represent a fundamental source of scientific advance.

A CAUSE OF LOGICAL INCONSISTENCIES BETWEEN DATA STRUCTURE, TECHNIQUE, AND THEORY

The fact that archaeology seeks to reconstruct and understand nonobservables (past behavior and ideas) through the discovery and investigation of patterning among observables (archaeological phenomena) places archaeological method within the realm of scientific method, involving the generating and testing of hypotheses or models (complexes of hypotheses). Hypotheses or models that are concerned with nonobservable activities and ideas are formulated through the discovery of patterns (generalizations) and are tested through the seeking of specific patterns (test implications) in archaeological observables.

The logic that is involved in formulating or testing hypotheses can vary in its consistency and potential for leading the researcher to accurate and meaningful conclusions. The degree of logical consistency that is realized and the accuracy of the conclusions that are reached depend minimally on two factors. The first is a *technical determinant*, which involves the complexity of the phenomenon, observables, and data set under investigation, and thus whether pattern-seeking mathematical techniques (e.g., factor analysis) are required during pattern-searching stages of research.

The second is a *theoretical determinant*, which involves the accuracy of the auxiliary assumptions that are used in assigning meaning to recognized patterns. For the sake of argument, the second, theoretical determinant will be held constant. It will be assumed that the auxiliary assumptions that are used in assigning meaning to recognized patterns are correct. Instead, attention will be focused on variation in the technical determinant and its effects on the consistency and accuracy of hypothesis formulation or testing.

Two kinds of situations, varying in the logic that they involve, can arise. In

the first, a researcher studies relatively simple phenomena of interest. Observables and data observations pattern themselves clearly such that the researcher can observe the patterns by him/herself, without the aid of mathematical pattern-searching algorithms. In this case, the logic involved in the formulation or testing of hypotheses can be carried out with consistency between data structure, generalization/test implication, and hypothesis and can lead to accurate and meaningful conclusions. This can be done in the manner envisioned by philosophers of science (Hempel, 1966; Hanson, 1972). In the second situation, the phenomena studied are relatively complex. Observables and data observations pattern themselves in ways that require the researcher to use pattern-finding mathematical techniques rather than his own senses and mental capabilities, alone, to search for pattern. In this case, logical inconsistencies between data structure, technique, test implication, and hypothesis can creep into the analysis. Inaccurate quantitative results and distorted understanding may be derived.

Let us see how the study of complex phenomena and patterning among observables with mathematical techniques presents a problem in logic, whereas the study of simpler phenomena and patterning among observables with one's own mental capabilities does not. We can do this by comparing the two approaches, varying only the technical determinant of consistency and accuracy.

In the analysis of either simply or complexly structured data, the researcher's task is twofold: to *find* patterns inherent in the data and *interpret* them with regard to the phenomena that produced them. Interpretation of an empirical pattern can be achieved in two ways. The first involves comparing the empirical pattern to those patterns implied by extant predictive laws or models and then matching it to one of the expected patterns (Toussaint, 1978, pp. 191-192). This allows the logical subsumption of the specific case under the accepted general law or model (i.e., explanation). Alternatively, interpretation of a pattern can be achieved through the formulation of a new hypothesis/model that implies an expectable pattern which matches the one found (theory/hypothesis building), followed by logical subsumption of the specific case under the new general principle. This amounts to explanation only if the new principle can be confirmed with other, independent data. The problem with the logic of analysis of complex data sets is most apparent in those analyses that require the second means of interpretation—hypothesis generation—but also is an aspect of analyses that involve the application of extant laws/models for interpretation. Let us first consider analysis involving hypothesis generation.

For simply structured data sets with simple patterning among observations, the tasks of finding patterns and interpreting them through the generation of adequate explanatory hypotheses can be done with a *single* mental operation called *abduction* (Hanson, 1972). Abduction is the simultaneous discovery of a *pattern* and its *significance* in suggesting a hypothetical cause of the pattern, as one

searches data. Abductive reasoning is the simultaneous dawning that a pattern exists in one's data, and that "the pattern could be explained if hypothesis X were true."

Abduction is more than induction—the process by which generalizations (patterns) are formulated (Hempel, 1966). Abduction involves the realization of the higher-level *meaning* of a generalization (i.e., the development of an hypothesis) as well as the formulation of the generalization, itself. Abduction is also more than retroduction—the process of understanding that a pattern could be explained if a given, new hypothesis were true. (Here, I depart from Hanson's [1972] use of the term retroduction.) Abduction involves the *perception* of a new pattern—previously unperceived—as well as the retroduction of a new explanatory hypothesis. Abduction, then is the seeming "conceptual gestalt" by which, simultaneously, new hypotheses are born and patterning in data suddenly becomes clear and explicable.

For more complexly organized observables, the processes of discovering patterns among data items and inferring the phenomenon that produced those patterns is a more intricate, *multistep*, *serialized* task. In the most common approach to analysis, first, sophisticated, pattern-finding mathematical techniques are used to search the data for multidimensional patterns and to summarize those patterns in two or three dimensional representations that humans can visualize, or with a few statistics (but see the "entry model" approach to data analysis, described below). This step, which involves the generalization of patterns among observables, is equivalent to logical *induction*. Second, the patterns that are found and summarized are then interpreted in terms of the phenomena that produced them. This step, which involves the logic that a pattern would be explicable if a given, new hypothesis were true, is equivalent to logical *retroduction*.

The difference between the multistep, serialized logic of analysis of complex data structures and the single-step logic of analysis of simple data structures is critical. It involves a *separation* of the process of *finding patterns* from the process of *interpreting patterns*, by the mathematical technique used to search the data. *This separation has the fundamental consequence of leaving room for the development of logical inconsistencies between data structure, pattern-finding technique, and interpretive framework during the course of analysis. These inconsistencies, in turn, may result in the distorted definition of relevant patterning and the drawing of false conclusions.*

To elaborate on this point, in the simple abductive process, data are searched for patterns by a human mind that both *knows* what kinds of patterns are possibly meaningful and expectable from an interpretive standpoint, and *searches* the data for precisely those patterns in some appropriate way. At the same time as the data are searched, the interpretive framework, the set of possibly meaningful patterns (the aspects of the data's structure considered relevant), and the mode of search are questioned and reformulated continuously, in light of the patterns found in the data. Feedback is instantaneous and

continuous between data structure, mode of pattern searching, and interpretive framework, which brings and keeps all three in logically consistent relationships with each other. This feedback and logical consistency is possible only because the mental activities of searching for patterns and interpreting pattern occur simultaneously, in parallel, as part of the process of abduction.

In the case where mathematical techniques are used to search for patterns in complexly structured data items that pertain to complex phenomena, such feedback does not occur on a continuous basis. Data are not searched for patterns by a human mind with a *variable* search strategy, a variable theoretical framework, and a variable list of potentially relevant aspects of the data's structure—all of which may change as knowledge is gained about the data's structure during the search. Rather, data are searched by a mathematical technique with a *fixed* search strategy that is consistent with and implies a fixed interpretive framework and a fixed list of aspects of the data's structure that are considered relevant—none of which change when applied to the data. Critically, whether or not the technique produces mathematical results that accurately represent the relevant structure of the data depends on the degree of logical consistency between aspects of the data that are assumed relevant by the technique and those that actually are relevant. Interpretation then *follows* the search for patterning, *after* logical inconsistencies between relevant data structure, technique, and implied theoretical framework, as well as misrepresentation of the data, have had a chance to be incorporated in the analysis. Interpretations and conclusions of questionable accuracy and meaning may be derived. Thus, logical inconsistencies in analysis and erroneous conclusions can result from the *separation and serialization* of the processes of searching for pattern and interpreting pattern, and the inopportunity in a single search-pass over data for *feedback* between the currently known aspects of their structure and the nature of the search technique and interpretive framework.

The separation of the processes of searching for pattern and interpreting pattern by technique, and the undesirable effects of this separation on the logic of analysis, characterizes not only analyses that involve the formulation of new hypotheses for interpretation. It also characterizes analyses that involve the application of extant laws/models for interpretation. First, test implications are deduced from alternative models/laws that might have interpretive value for the specific case under investigation. Then the data are searched by mathematical techniques having assumptions that are concordant with the theoretical framework, rather than the relevant aspects of the data, in order to find patterns that match one or more of the test implication(s) and that allow the logical subsumption of the specific case under one or more of the explanatory models/laws. The patterns that are found within the data by the search technique may represent the relevant aspects of the data's structure with varying degrees of accuracy, depending on whether the technique's design and the assumptions it makes are logically consistent with the relevant aspects of the data. The matches obtained

between found pattern and deduced test implications, and the interpretations made, will correspondingly vary in accuracy and meaning. Once again, in a single search-pass over the data, there is no opportunity for feedback between currently known aspects of the structure of the data, on the one hand, and the search technique and interpretive model implied by it, on the other. Logical inconsistencies are allowed to develop between data and technique, and erroneous conclusions are allowed to be drawn.

If one considers that in a real analysis, an interpretive framework involves uncertain auxiliary assumptions in addition to the primary hypotheses/laws or models, the problem of how to analyze a complex data set with logical consistency and how to define relevant patterns within it that have potential for being assigned appropriate meaning becomes all the more apparent and troublesome. In a real analysis, choice of analytic technique and the patterning in the data that is revealed may be influenced by the auxiliary premises as well as the primary premise assumed true. A delay in the feedback between known aspects of data structure and the interpretive framework (including the auxiliary assumptions) may result in a poorer choice of techniques for searching the data, a greater potential for inconsistency in analysis, and the definition of léss relevant patterning. This circumstance will increase the possibility of inaccurately assigning meaning to analytic results, additional to the effects of any inaccuracies in the auxiliary assumptions, themselves.

In sum, maintaining logical consistency during the analysis of complex data sets often cannot be achieved for any single pass over the data. This problem does not result from the use of a pattern finding technique, per se. Both mental scanning of simple data sets and mechanical scanning of complex ones require the use of some search technique, yet the effectiveness of the latter may not match that of the former. Rather, the problem with the analysis of complex data, as typically approached, results from separating and serializing the processes of finding pattern and interpreting pattern, which does not allow continuous feedback between data structure, search technique, and theoretical framework. For circumstances involving hypothesis formulation as opposed to hypothesis testing, the problem posed by complex data sets, compared to simple ones, can be summarized in logical terms. *Analysis of complex data requires inductive and retroductive logic, whereas analysis of simple data can be achieved through abduction.*

THE PROBLEM OF GETTING INTO DATA

The serial process of finding pattern and interpreting pattern that is commonly involved in the analysis of complex data sets can result in the two potential kinds of discordance between data, technique, and theory that are described in this chapter's introduction. In brief, an analytic technique and the interpretive framework with respect to which it is chosen may assume that the data of interest pertain to a single process and population (relevant subset

structure) and have a certain organization (relevant relational structure) when in actuality the data may reflect multiple processes and populations and have a different relevant relational structure.

These kinds of discordance can occur at any stage in the analysis of complex data sets. They are particularly problematic, however, at the beginning of the analytic process. At this point, little may be known about the relevant aspects of the data's structure. As a consequence, the researcher is put in a bind. He cannot choose an appropriate analytic technique and an appropriate subset of the data for analysis without some knowledge about the data set's structure; yet he cannot obtain this knowledge without applying some pattern-searching technique to summarize the data's structure in a simpler form that is comprehendable by the human mind. If the researcher uses an inappropriate technique and subset of the data, the patterns that are found may not be an accurate representation of the data's relevant structure, nor meaningful. Furthermore, these distorted patterns—if used as the basis for making basic transformations of the data (screening it) in order to bring concordance in later analytic steps—can instead focus the analysis in a direction of greater discordance.

A very simple example of this bind during the initial stages of analysis of complex data is given by Christenson and Read (1977, p. 171). Concerned with the typology of a set of projectile points, they note that they could not do an R-mode factor analysis of the data to determine relevant dimensions of morphological variability without first eliminating extreme cases and defining a homogeneous population. (To not eliminate such cases would introduce distortions in the magnitudes of the correlation coefficients serving as a basis for the factor analysis.) At the same time, proper multivariate (as opposed to univariate) identification of the outliers required that the dimensions of variability present in the data be known.

Thus, complex data pose to the researcher a problem of how to enter or "get into" them without violating the relevant aspects of their structure. To circumvent this problem, at least four different strategies of analysis can be used. These are discussed in the remaining sections of this chapter and the following chapter by Read.

SOLUTIONS FOR MAINTAINING LOGICAL CONSISTENCY BETWEEN DATA STRUCTURE AND TECHNIQUE

To enter unknown, complex data yet maximize consistency between relevant aspects of its structure and technical assumption, four complementary strategies of analysis can be used:

1) deductive specification of potentially relevant variables and observations, and an appropriate technique;
2) "constrained" exploratory data analysis;

3) the "entry model" approach; and
4) stepwise, cyclical analytic designs.

Each of the strategies improves the researcher's chance of 1) selecting a subset of variables and observations that reflect the phenomenon of interest, and 2) choosing an analytic technique that is concordant with and sensitive to the relevant structure of the data. In this way, the strategies help to resolve the two potential kinds of discordances that can occur between data and technique, which were summarized in the beginning of this chapter, and to overcome the methodological double bind.

Deductive Specification of Potentially Relevant Variables and Observations, and an Appropriate Technique

One direct strategy for improving the degree of logical consistency within an analysis is to deductively specify that subset of the available data and that analytic technique which are *likely* to be relevant to the phenomenon of interest. This is done on the basis of extant theory about the nature of that kind of phenomenon in general. To the extent that the expected nature of the phenomenon of interest does not concord with its actual nature, irrelevant variables and items may be included in analysis and some meaningful ones may be deleted. In addition, the technique that is chosen for analysis may assume that certain kinds of relationships among variables or observations are relevant, when in fact other kinds reflect the phenomenon of interest more accurately (see Carr, chapter 1).

This strategy is usually employed at the beginning of a multistep analysis, when little is known about the structure of the specific data that are available for analysis. It can be followed by more inductive exploration of the chosen data using techniques that are justified on the basis of the initial insight that is obtained into the data's structure.

In archaeological studies, middle range theory (e.g., Binford, 1977a; Schiffer 1976; Raab & Goodyear, 1984) is frequently used to deduce the subset of data and/or the technique that is likely to be appropriate for analysis. Middle range theory is useful in this regard because it specifies the archaeological observables that are expected to manifest particular phenomena of interest, and/or their expected nature of organization. Let us consider some examples of this application of middle range theory.

Middle Range Theory Used to Specify Relevant Subsets of Variables and Observations

Artifact Style Analysis. One area of currently active research, in which middle range theory has been used to deduce potentially relevant variables and observations for analysis, is artifact style analysis for the purpose of testing or formulating propositions about prehistoric social organization. Wobst's (1977) information exchange theory of style specifies the characteristics of items that are likely to indicate group affiliation through their morphology, thereby sug-

gesting the *observations* that are probably relevant to the study of prehistoric social organization. Items having this potential are those that 1) probably were used in contexts ensuring their visibility to all members of the group and members of other nearby groups, as opposed to items used in the domestic sphere; 2) are long-lived, making their expression of group affiliation efficient over time; and 3) probably were not exchanged between groups.

Voss (1980a, p. 4), following Wobst, goes on to specify, for such items, the different kinds of stylistic *variables* that are often useful for determining group affiliation versus group interaction. Discrete characteristics that are highly visible and that thus can function effectively as symbols, such as discrete design elements and configurations (Fredrick, 1970; Stanislawski & Stanislawski, 1978), are more likely to be accurate measures of group affiliation. In contrast, continuous stylistic variables that encompass the "nuances" of style, such as the dimensions of design zones and counts of design element repetitions, are more likely to be accurate measures of group interaction. Finally, Braun and Plog have suggested that the stylistic characteristics of an artifact define a hierarchy. Attributes at different levels of the hierarchy represent different stages of the decision process that are involved in the manufacture of the artifact (Plog 1978, p. 161), but are also sensitive to different social factors and groups (Braun & Plog, 1982, p. 511), perhaps at different geographic scales (Braun, 1980, pp. 12-13).

In total, these middle range principles define a very powerful framework. From it, the kinds of observations and variables in an artifact style data set that are likely to be relevant to the study of prehistoric social organization can be deduced with a great degree of specificity. When applied within the bounds of their limitations (see Wiessner, 1983 for a discussion of limitations), these principles suggest the subset of variables and observations that probably pertain to a single social process, or a limited range of social processes, and that tend to be accommodated to statistical techniques based on models assuming some single process. Thus, in this case, deductive specification of variables and observations can improve the likelihood of concordance between data structure and technique.

Some studies of prehistoric social organization that have used these principles in this manner include those of Braun (1977, 1980), Plog (1976, 1978, 1980), Voss (1980a, 1980b), and Hinkle (1984). We may also note that Spaulding's emphasis on using nominal scale measures (or higher-scale measures reduced to a nominal scale) as the basis for artifact typologies, derives from conclusions of his that are concordant with the information exchange theory. Spaulding (1982, pp. 5-6, 10) argues that it is nominal scale (discrete) variables that indicate culturally imposed patterns of artifact manufacture and that may be used to define types having cultural significance (i.e., indicating the group affiliation of the artifact's makers).

In other fields of study, middle range theoretical arguments similarly allow

one to deduce variables and observations that are probably relevant to some phenomenon of interest.

Principles of lithic technology. In this volume, Hoffman (chapter 18) uses principles of lithic technology to select several variables for investigating morphological variation in a set of projectile points that is relevant to maintenance and reduction processes. These include measures of blade edge angle and blade size.

Mortuary analysis. Braun (1979, p. 69) argues that the archaeological variables that are relevant to the identification of ascriptive, hierarchical social distinctions include grave good classes that do not occur in village middens, that occur rarely overall within burials, that involve a relatively substantial labor input to produce, and that do not associate with age or sex. He also argues (p. 67) that it is qualitative rather than quantitative burial ritual attributes that symbolize formal authority and hereditary ranking. These arguments were then used by Braun to select a potentially relevant subset of variables from a burial set for factor analysis. O'Shea (1981, p. 42) has made similar arguments specifying the kinds of mortuary variables that are likely to distinguish horizontally or vertically differentiated social segments. These, in turn, were used to select potentially relevant variables for a factor analysis.

Middle Range Theory Used to Specify the Nature of the Phenomenon of Interest, Relevant Relational Data Structure, and Appropriate Technique

Over the past ten years, there has been a growing, general concern about the proper use and misuse of higher-level quantitative techniques (Thomas, 1971, 1978; Cowgill, 1977; Hole, 1980; Vierra & Carlson, 1981; Scheps, 1982; Moore & Keene, 1983). These concerns have been met by active research into the nature of organization of the archaeological record and the relevant structure of archaeological data in various contexts, the development of middle range theory about that organizational variation, and specification of the contexts in which applications of various quantitative techniques are appropriate.

Intrasite spatial analysis. Carr (1984) has reviewed most spatial quantitative methods that are currently used in intrasite studies in regard to their concordance with a model of intrasite artifact organization that commonly typifies archaeological sites. The model specifies that depositional sets may be polythetic and overlapping in organization. It also specifies that depositional areas may vary in their size, shape, orientation, spacing, artifact density and composition, border crispness, and in whether they overlap and are hierarchically arranged in space. Similarly, Whallon (1984) has modeled the kinds of variability encompassed by depositional areas, and has evaluated the use of factor analysis and other global methods in relation to it.

Seeing the discordance between most currently used techniques of intrasite spatial analysis and the structure of intrasite archaeological records, both

Whallon (1979, 1984) and Carr (1977, 1981, 1982b, 1984) have formulated new analytic methods that exhibit greater concordance with and sensitivity to the behaviorally relevant aspects of intrasite artifact distributional variability. In chapter 13 of this volume, Carr continues to enumerate additional mathematical models of intrasite artifact organization, some behavioral and natural contexts in which intrasite artifact organization can be expected to concord with those models, and some quantitative techniques (new and old) that are appropriate for use in those contexts. In this way, with a limited understanding of the behavioral and natural context of a site, it is possible to deduce the probable relevant organization of artifacts within it and the techniques most likely concordant with that organization.

The progress that has been made in these studies is based on nearly a decade of previous research that has focused on evaluating the response of various techniques to different spatial organizations of artifacts. These earlier studies, however, did not involve the construction of models of artifact organization that allow particular sites to be subsumed under them and that specify the techniques appropriate in those instances. For example, Schiffer (1975), through simulation, assessed the ability of factor analysis to reconstruct depositional sets of artifact types when the percentage of multipurpose (as opposed to single purpose) types becomes large and when correlation coefficients based on type counts within grid cells are used as the factored coefficients of similarity. Speth and Johnson (1976, pp. 50-53), using grid cell counts of artifact types, evaluated the response of intertype correlation coefficients to different artifact arrangements that result from different depositional processes.

Economic analysis. Another area of active modeling of the nature of organization of the archaeological record and behavior, and the techniques appropriate to their analysis, is economic analysis of settlement location choice and subsistence resource choice. Limp and Carr, Parker, Kvamme, and Keene (chapters 7, 8, 9, 10, respectively) each propose models of the nature of such decision processes and evaluate the concordance between various quantitative methods and those models. Among the technical assumptions considered in the evaluations are the level of information that is assumed accessible to the decision unit, the information processing capabilities that are implied of the decision unit, the assumed degree of continuity of settlement locations over space, and the implied degree of importance of social and ideological factors in subsistence and settlement decisions. These studies parallel previous evaluations of the concordance between technique and decision process in economic anthropology (e.g., Gladwin, 1975) and archaeology (Reidhead, 1979). It should be noted that in all these studies, general anthropological and economic theory, rather than middle range theory, is used to specify the appropriate analytic technique.

Mortuary analysis. Braun (1977), for example, has argued that ascribed and achieved status positions differ in the *predictability* (institutionalization) that is demanded of the behaviors associated with them and, hence, the constancy of

mortuary ritual treatments of individuals that occupy those positions. He also suggests (1979, p. 67) that ascribed status is symbolized at death by *multiple, redundant* forms of variation in burial ritual. On the basis of these postulates, Braun deduces that factor analysis can be a useful technique for identifying indications of ascribed status within a mortuary data set.

Artifact typology. Hodson (with Doran 1975; 1982, pp. 25-26) and Spaulding (1977, 1982, p. 18) have debated the appropriateness of object clustering techniques relative to attribute clustering techniques for creating artifact typologies. Cowgill (1982, pp. 45, 47-48, 50-53) argues that the two approaches can product equivalent or complementary results, and that the preferability of one approach over the other can vary. This depends on whether the data are measured on a nominal scale or continuous scale, and the data's particular structure (e.g., the distribution of marginal frequencies in the case of nominal data in contingency table format).

In sum, during the initial stages of analysis, the relevant structure of the data in hand is not often well-known. In this circumstance, deductive specification, from theory, of the subset of data and the technique that are likely to be relevant to the phenomenon of interest can be a powerful means for getting into the data and reducing the degree of discordance between data and technique. As theory develops—particularly middle range theory—and evaluation of the response of various techniques to different kinds of relevant archaeological data structures continues, we can expect this means for getting into data to become more helpful.

"Constrained" Exploratory Data Analysis

Deductive arguments can help one to narrow the range of variables and observations within a data set to those having greatest potential for reflecting the phenomenon of interest. They can also suggest a technique that is most apt to be concordant with the relevant aspects of the data's structure. However, deductive argumentation is seldom sufficient. Individual data sets need not—usually will not—conform to expectation in every way. If the theory employed to make deductions does not have strong predictive capabilities, the phenomenon of interest may manifest itself in unsuspected sets of variables and observations and forms of relationships among them (e.g., association rather than covariation). The same problem can arise if the predictive theory is incorrectly applied beyond the limits of its boundary conditions or if the auxiliary assumptions that are made when relating the theory to the data at hand are wrong. For example, consider the auxiliary assumptions that are made about sources of variation that are supposedly controlled during data collection. When unsuspected extraneous factors as well as those of interest affect the measurements brought forward for study, the data set's relevant structure may take an unexpected form. An expected linear relationship between two natural environmental

variables, for example, might instead take the form of a cyclical function with a linear trend, as a result of the compounding of diurnal variation with the variation of interest. Thus, totally deductive specification of the variables and observations to be analyzed and the techniques to be used need not ensure the complete relevance of the selected data to the phenomenon of interest, nor the concordance of the selected technique with the data's relevant structure.

To compensate for these problems—to get into the data more successfully—it is necessary to supplement the deductive strategy with an inductive one that examines the data on its own terms. *"Constrained" exploratory data analysis* (CEDA), having at least two variants that differ in the kinds of techniques they employ, is useful for this purpose.

CEDA vs. Exploratory Data Analysis

As defined here, CEDA includes all the analytic approaches for getting into data that comprise exploratory data analysis (Tukey, 1977; Hartwig & Dearing, 1979; Clark, 1982) but only a *portion* of the philosophy of exploratory data analysis that motivates their use. Like exploratory data analysis, CEDA is an inductive approach to recognizing patterns in a data set. Both have the goal of finding "any unanticipated structures or relationships that occur within a data set, regardless of expectation" (Tukey & Wilk, 1970, p. 371). Both involve searching for any patterning in the data in order to reach a better understanding of the nature and causes of its total structure. Unlike in exploratory data analysis, however, in CEDA, this understanding of the data's total structure is sought in order to *isolate* the *relevant* aspects of it—*those that reflect some one explicitly specified phenomenon of interest as defined deductively* by the larger theoretical framework or paradigm of the researcher. In contrast, in exploratory data analysis, understanding of the total data structure is sought explicitly in order to generate new ideas, problem areas, and hypotheses (Tukey, 1979, p. 122; 1980, pp. 23-24) within a primarily inductive framework. Discovery of *many* relevant data structures pertinent to many phenomena, rather than the *single* structure pertinent to the single phenomenon of interest, is the goal of exploratory data analysis. Because CEDA is undertaken within a larger deductive framework and is more focused in its aim, whereas exploratory data analysis occurs within an inductive, less focused context, the designation *constrained* exploratory data analysis is used.

Whereas exploratory data analysis was developed by Tukey in reaction to the strongly deductive, "confirmatory" mode that dominates theoretical statistics (Tukey, 1979), CEDA is meant to articulate with it. Analysis is begun in a deductive manner with the specification of variables and observations that are probably relevant or irrelevant to the phenomenon of interest. Data items that are thought to be irrelevant are dropped from analysis. The search for relevant and irrelevant data items is continued in an inductive manner with CEDA

procedures. The subset of the data that results from *both* the deductive and CEDA steps can then be used in either hypothesis testing or hypothesis formulation. In either case, both data screening steps are motivated by the researcher's larger theoretical framework, which specifies the phenomenon of interest. *CEDA, then, is an inductive middle-step within a stepwise analytic design that has an overall deductive orientation and that is begun with deduction.* In contrast, exploratory data analysis is an inductive approach for initiating analytic process. (See Carr, chapter 13 for a discussion of the strengths and weaknesses of exploratory data analysis in this capacity.)

To examine the total structure of a data set and determine those aspects of it that are probably relevant to the phenomenon of interest, CEDA uses the same methods as exploratory data analysis, plus some additional ones. First, to view the multiple structures within a data set, CEDA involves the re-expression of the data on various scales of measurement (e.g., nominal, ratio, logarithmic, square root) and the examination of the re-expressed data with different techniques that are concordant with those scales of measurement (Tukey, 1980, p. 24; Hartwig & Dearing, 1979, p. 10). Also, techniques assuming multiple mathematical models are used to investigate the data from multiple perspectives. An effort is made to find any patterns in the data, regardless of whether they reflect the phenomenon of interest, and to consider how potentially relevant structure in the data then might be isolated from irrelevant patterning (e.g., removal of outliers, selection of variables, expression of the data on a particular scale, use of a technique that is sensitive to the scale most likely appropriate to the relevant structure). Second, CEDA, like exploratory data analysis, stresses the importance of graphic representations of the data (e.g., histograms, crossplots, the box and whisker, maps) in aiding the search for patterns (Tukey, 1970, p. 372; Hartwig & Dearing, 1979, p. 9).

CEDA vs. Data Screening

CEDA and exploratory data analysis involve many of the same techniques and operations traditionally used to *screen* data in preparation for the application of higher-level statistical techniques. These include histograms; crossplots; simple univariate descriptive statistics; bivariate techniques of association, rank correlation, and correlation analysis; elimination of outlying observations; segregation of modalities for separate analysis, should the data be composed of observations within several suspected populations; and transformation of the form of the frequency distributions of individual variables or the functional relationships between variable pairs. All of these techniques and operations can be used in CEDA to obtain a basic understanding of the data to be analyzed. However, CEDA departs from traditional data screening in that these methods are not applied in order to transform the structure of the data into a form that is concordant with some particular analytic technique to be used. Data are not screened to *fit to technique*. Rather, the data set is examined to find and isolate the

potentially relevant aspects of its structure, *in regard to which an appropriate technique of analysis is chosen or developed.*

Two Forms of Constrained Exploratory Data Analysis

CEDA encompasses two variants, which differ to some extent in the nature of the techniques they encompass. The first variant emphasizes the use of techniques that make *minimal assumptions* about the data's structure when displaying it for pattern-searching. The second variant emphasizes the use of techniques that are capable of handling *a heavy load of irrelevant variables or observations, or variation in general*—a common characteristic of data sets that have been screened only by deductive selection.

A good example of a technique that makes minimal assumptions and that might be used with a CEDA framework, but which to date has been used in only an exploratory data analysis framework, is Whallon's (1984) *unconstrained clustering* method of intrasite spatial analysis. As Whallon (1984, p. 275) notes, unconstrained clustering "is hardly more than an elaborate approach to a descriptive summary or display of the data or a series of such summaries and displays." (For a more detailed discussion of the method in relation to exploratory data analysis see Carr, chapter 13.) Other examples of techniques that make minimal assumptions include other graphic displays—such as the stem-and-leaf display, the box-and-whisker, and scatterplots—and certain "resistant" descriptive statistics— such as the trimmed mean, the Winzorized mean, and the median absolute deviation (Tukey, 1977; Hartwig & Dearing, 1979, pp. 16-26). These various techniques can be used to determine the variables, observations or the relationships among them that are probably relevant to the phenomenon of interest.

One example of a technique that is capable of handling a heavy load of irrelevant variables, but that does not make minimal assumptions, is *R-mode factor analysis*. Although it can be used for multiple purposes, R-mode factor analysis is ideally suited for defining clusters of variables, making it "easier to decide upon their relevance to a problem" (Christenson & Read, 1977, p. 174) with a CEDA framework.

Christenson and Read (1977, pp. 167, 170-174) have used factor analysis along with a multivariate identification-of-outliers program explicitly this way in preparation for developing a projectile point typology with cluster analysis. A factor analysis of projectile point data was used to identify two dimensions of morphological variability (groups of variables) that seemed relevant to the researchers' typological goals, and other dimensions that seemed irrelevant. The two relevant dimensions were then selected as the "variables" to be used in creating the point typology. The following chapter by Read continues discussion of this approach. It illustrates how factor analysis and scatterplots of factor scores can be used in an alternating, iterative manner to refine the selection of variables (factors) and set of observations that are chosen to represent the one or

more phenomena of interest that are potentially reflected in a data set's structure.

A second example of a technique that is capable of handling a heavy load of irrelevant variation but that does not make minimal assumptions is *spectral analysis*. This technique allows the researcher to identify multiple forms of variability of different scales that are compounded within the track of a *single* response variable over time or space. The results of a spectral analysis can be used by a researcher to design "filters" that allow the extraction and isolation of these individual forms of variation from the compounded response variable, which in turn are defined as new variables. Those of the new variables that are considered to reflect the phenomenon of interest can then be subjected to further analysis, free of the confusing effects of the other, irrelevant sources of variation.

Carr (1982b) has used spectral analysis in this manner to identify and analyze several sources of variability within an intrasite resistivity survey data set. He has also suggested its use for identifying the different kinds of depositional processes that are responsible for artifact density variation within a composite artifact distribution (a palimpsest) that has been formed by the partial spatial overlap of multiple depositional processes (e.g., different kinds of activities of different scales). Artifact density variation that is thought pertinent to each of the depositional processes of interest (relevant data structure) can then be extracted from the palimpsest for individual study using filtering techniques (Carr, 1982a; 1984; 1986 this volume, chapter 13).

In conclusion, analysis of complex data sets often requires inductive as well as deductive specification of the variables, observations, and relationships among them that are likely relevant to the phenomenon of interest, and a concordant analytic technique. In this regard, in the case of complex data, phases of scientific investigation that are concerned with hypothesis testing and that are supposedly "deductive" are seldom *completely* deductive. Theory may be used to deduce a model or hypothesis, but the formulation of a test implication—which states an expectable relationship among observables in *terms of the variables, cases, and technique that is selected for analysis*—is a process that often requires both deductive and inductive logic. The expectable relationship follows from the theoretical framework, but its expression depends on the data and technique to be used, which often must be selected in part by induction.

Entry Models

A third strategy for getting into an unknown data set while maximizing consistency between its relevant structure and technical assumption involves the construction and use of what may be termed *entry models* and *parallel data sets*. This strategy involves both inductive and deductive logic, and requires the use and development of middle range theory. It gives the researcher insight into the organizational nature of the phenomenon of interest and the relevant relational

structure of the data set (e.g., nominal vs. ratio scale organization), thereby allowing the researcher to choose an analytic technique which is more concordant with that form of organization. It does not necessarily involve the selection of relevant observations and variables, though it may. The strategy is summarized in Figure 1.

An entry model has three essential components: 1) The most critical is a *general mathematical model* or description of the *form of organization* of the archaeological observables that represent the phenomenon of interest. An example would be a model that specifies the form of spatial organization of coarranged artifact types within a site as monothetic or polythetic. To this organizational model are linked the other two components of the entry model.

2) The second component is an *enumeration of the kinds of processes* that could lead to the archaeological observables being organized in the way that the mathematical model specifies. These classes of processes will always include cultural and natural formation processes of the kind documented by middle range archaeological theory (e.g., curation, lithic reduction and maintenance processes, means by which rank is symbolized in mortuary remains). However, they may also include processes to which general anthropological theory pertains (e.g., the pattern and tempo of fission-fusion of hunter-gatherer bands). Continuing our intrasite spatial example, above, a list of processes that can cause coarranged artifact types to be organized monothetically or polythetically might include differential artifact preservation, artifact curation, artifact recycling, misclassification of artifacts, or the occurrence of alternative tool types within a tool kit (see Carr, chapter 13). Although specification of such linkages between form of archaeological organization and process may be difficult, it is currently the subject of active research on middle range archaeological theory.

3) The last component is an inventory of the quantitative techniques that are concordant with the mathematical model of organization of the archaeological observables. For example, some kinds of "polythetic association" methods of spatial analysis (Carr, chapter 13) would be concordant with a polythetic coarrangement of artifact types.

An entry model usually is one of a series of such models. Each entry model specifies a different mathematical model of the organizational form of the archaeological observables that represent the phenomenon of interest. The differences between the mathematical models of organization in the various entry models reflect the different effects of different classes of formation processes or higher-level processes, which are enumerated by the entry models. The entry models will also specify different quantitative techniques that are concordant with their different mathematical models of organization. For example, Carr (chapter 13) defines six alternative entry models. They involve different mathematical models of possible organizations of "tool kits" that have been deposited in the archaeological domain (archaeological observables) and that represent activities (the phenomenon of interest). The different mathe-

matical organizations of the "tool kits" reflect the different effects that various kinds of formation processes, which are enumerated by the entry models, would have. The entry models also specify different sets of quantitative techniques that are concordant with the different organizations of "tool kits" and that can be used to search for them in archaeological data.

Entry models are useful when two circumstances occur. 1) The data set that documents the phenomenon of interest and that is slated for analysis is very complex. As a result, the researcher is unable initially to specify—by simple inspection of the data—the aspects of its structure that are likely to be relevant, the probable organization of the archaeological observables that represent the phenomenon of interest and that are expressed in the data, and an analytic technique that probably is appropriate. 2) There exists a simpler, *parallel data set* that gives the researcher insight into the processes that are responsible for the archaeological observables, their consequent organization, and the relevant and irrelevant structural aspects of the complex data set. In these two circumstances, it is possible for the researcher to learn something about the complex data set's relevant relational structure (e.g., the relative frequency of monothetic or polythetic relationships of association among coarranged artifact types) and to specify an analytic technique that is likely to be appropriate by examining the *parallel* data set and using the entry model. This is done in lieu of *directly* but possibly discordantly examining the *complex* data with a higher-level pattern-searching technique. In this way, the researcher is removed from the bind of not being able to choose an appropriate analytic technique without knowledge of the complex data set's structure, yet not being able to obtain this knowledge without applying some pattern-searching technique to the complex data.

The logical process involved in the use of a parallel data set and an entry model is shown in Figure 1 and can be described as follows:

1) Archaeological observables that reflect the phenomenon of interest are described in two separate data sets: a complex one that is the ultimate target of analysis and a simple, parallel one that is sensitive to the processes of formation of the archaeological observables in the complex one. An example of a complex data set would be a matrix of point locations of artifacts of many classes within a hunter-gatherer site. Examples of a simple data set that parallel this complex one and gives insight into the complex data's organizational nature would be (a) one that contains information on intrasite spatial variation in soil acidity (reflecting the potential for differential preservation of bone artifacts over space), (b) one that documents the orientation and dip of artifacts (indicating the possibility of disturbance in the spatial distribution of artifacts by fluvial activity), or (c) one that describes the grain of the surrounding environment (suggesting the likelihood of tethered mobility patterns, repeated reuse of the site, and the palimpsest nature of the artifact distribution).

2) The processes that are responsible for the archaeological observables to be

studied in the complex data set are reconstructed on the basis of information in the parallel data set. This reconstruction can be accomplished by logical deduction, in which case the specific patterns within the parallel data set are subsumed under general, accepted models of the observable consequences of formation processes. It can also be achieved by abduction from patterns within the parallel data set, followed by testing of one's conclusions with other, independent data that also comprise the parallel data set. For example, in deductive mode, the disturbance of an intrasite artifact distribution by fluvial processes might be determined by noting patterns in the orientation, dip, and size sorting of artifacts and then subsuming such patterns under established models of fluvial displacement of artifacts (Behrensmeyer & Hill, 1980; Shackley, 1978). This step conforms to Schiffer's (1983) call for "up front" identification of the processes that are responsible for archaeological observables, prior to behavioral interpretation.

3) The specific processes that are found to be responsible for the archaeological observables within the parallel data set and that are also pertinent to the organization of the complex data set are then matched with processes that are enumerated in a more general way in one or more of the entry models.

4) On the basis of (a) the association of the archaeological observables in both the parallel and complex data sets with a particular entry model via common processes and (b) the model's specification of both the effects of formation processes on the organization of archaeological observables in the complex data set and the analytic techniques that are concordant with that organization, two

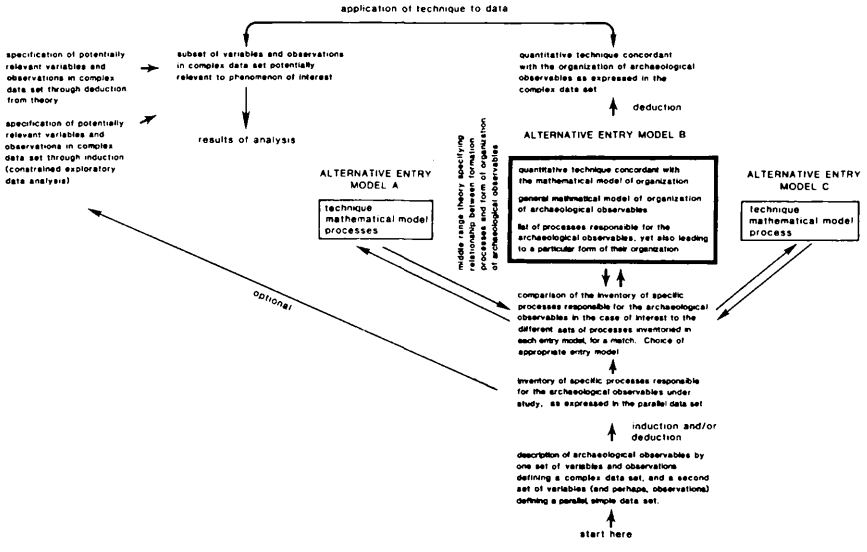

Fig. 2.1. The use of entry models and parallel data sets to get into a complex data set.

things are concluded. These are 1) the probable general nature of organization of the archaeological observables in the complex data, which is specified in mathematical terms, and by implication, some aspects of the complex data's relevant relational structure, and 2) the techniques of analysis that are most likely appropriate for investigating the archaeological observables in the complex data set. Note that information in the *parallel* data set is used to determine the appropriate entry model, whereas information on the relevant structure of the *complex* data set is derived from the entry model. Also note that the process of associating the archaeological observables in the parallel and complex data sets with a given entry model is equivalent to logically subsuming them under the entry model.

An example of the steps that have just been outlined is given in chapter 13 by Carr. Here, the target, complex data set is composed of the spatial distributions of many artifact types within a site. Formation processes that would have affected the nature of organization of spatially coarranged artifact classes within the site are identified with various kinds of aspatial data, which constitute a parallel data set. These processes are matched to those that are enumerated in two of several alternative entry models, which also include mathematical models of the organizational form of coarranged artifact types. From this match are concluded the two most probable forms of spatial organization of coarranged artifact types within the site, which is specified in mathematical terms (relevant relational structure), and the two techniques that are most apt to be appropriate for analyzing the artifact type distributions.

5) The quantitative technique that is determined to be most probably concordant with the relevant relational structure of the complex data is applied to that data, or some subset of its variables and observations that is thought potentially relevant to the phenomenon of interest. The potentially relevant subset might have been specified by deduction from theory or by inductive examination of the parallel data set. For example, again consider the application of the entry model strategy used by Carr in chapter 13. The artifact type, flint pebbles, might have been removed from the complex data set of artifact type distributions, and from the search for tool kits in that set, on the basis of an aspatial piece of information in the parallel data set: the fact that many of the smaller pebbles were probably of natural, fluvial origin, and thus, irrelevant to behavioral reconstruction.

In sum, the entry model strategy can be a powerful approach for getting into a complex data set while minimizing the violation of relevant aspects of its structure by an applied technique. The strategy allows the researcher to determine the probable general nature of the data set's relevant structure, and thus, the technique(s) that are most likely appropriate for its analysis, *without directly analyzing it with some possibly discordant method.* This is accomplished through the examination of a parallel data set for the processes of formation of the archaeological observables in both the parallel and the complex data sets, rather than

through a direct, inductive examination of the many facets of the complex data set's structure using multiple techniques. In essence, the entry model strategy allows one to investigate a complex data set by "slipping in a side door," which is provided by parallel data on formation processes, rather than by affronting it.

Although the entry model strategy can be more powerful and bring greater concordance between data and technique than the deductive or CEDA strategies, the entry model approach has a disadvantage. It requires a good foundation of middle range theory on the processes that are responsible for the archaeological and behavioral variability of interest, and also processes that are *not* of interest. As this foundation broadens, it will become more practicable (see Schiffer, 1983).

Stepwise, Cyclical Analytic Designs

A final means for improving consistency between data structure, technique, and theoretical framework, during analysis, is the well known stepwise, cyclical process of scientific investigation, itself (Fig. 2). This process requires repeated analysis of a data set—including modification of the data, the analytic technique(s), and/or the interpretive framework that guides analysis, with each pass over the data—such that all three approach greater concordance with each other. Modifications of these three entities with each cycle are made in light of 1) discrepancies between expectable results and those obtained (external inconsistencies), 2) discrepancies between the interpretive implications of different

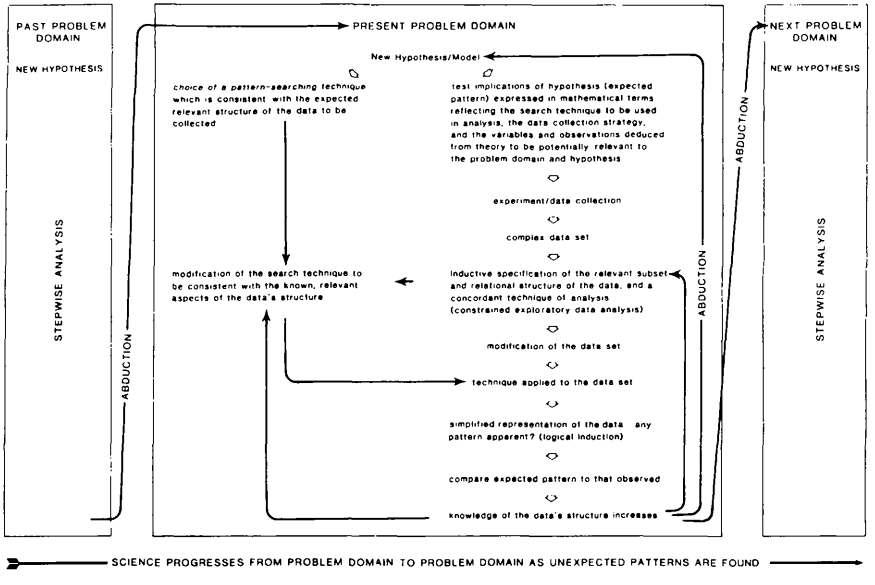

Fig. 2.2. The sequential, stepwise, cyclical process of scientific investigation.

subsets of the results that are obtained (internal inconsistencies), 3) whether the discrepancies increase or decrease from trial run to trial run, and 4) knowledge that is gained about the data's relevant structure. In short, the stepwise approach to data analysis is a logical process that *simulates*, in serial format, the continuous, simultaneous feedback between data, technique, and theoretical framework that characterizes the mental process of abduction.

THE LOGIC OF ANALYSIS VS. THE LOGIC OF PUBLISHED ANALYSIS

Successful analysis of complex data sets usually requires and involves deductive, inductive, *and* stepwise strategies for getting into the data and developing concordance between data, technique, and theory. It can also involve the use of parallel data sets. Nevertheless, if we were to look at a typical journal article that reports research on a phenomenon that is described by complex data, we would probably find, instead, a more deductive tenor. Justifications of the variables, observations, and technique(s) that are used would more likely refer to theoretical considerations than to patterns noted in the data or to insightful discrepancies between expectation and result in trial runs. We might also find simply a report of the problem of interest and associated hypotheses, data collection design, data, results of analysis, and conclusions, without explicit justification of the data items that were analyzed as opposed to those that were collected, or of the technique that was employed. In either case, the reader is left with the impression of a primarily deductive analytic process that has one or a few steps, as opposed to the more complex, multistep, multistrategy process that typifies the analysis of complex data as just described (see Binford & Sabloff, 1982 for a discussion of this effect in the New Archaeology of the 1960s-1970s).

This difference, which often occurs between the logic implied by the format of published scientific investigations and the logical processes by which such investigations are achieved, results from at least two factors. First, as a result of publication expense and limitations on space, it is often impossible to include in a report of investigation the multiple, sometimes complex reasons for deleting certain observations or variables, or for selecting one technique over another. Second, broad, general, deductive justifications of variables, observations, and technique can often be stated more succinctly than inductive justifications, which may have multiple idiosyncratic or contextual facets to them that can be conveyed only at length. (For examples of the latter, see Whallon, 1984; Read, chapter 3; Carr, chapter 13; Braun, chapter 16.) Thus, if any justifications of analytic strategy are included in a research report, it is the deductive ones, which can be expressed most briefly, that tend to be reported.

It is important to realize the difference between the deductive-tending logic of published analyses of complex data and the logic by which those analyses are accomplished. To confound published logic with the logic of analysis and to proceed with the analysis of complex data in a largely deductive, single-step

manner can have at least two negative consequences. First, by limiting one's strategies for getting into data to a deductive approach—by not also using an inductive constrained exploratory approach and iterative processing—the researcher greatly decreases his capabilities for identifying the relevant aspects of his data and for selecting data and technique so as to maximize concordance between them.

Second, by not examining data in an inductive constrained exploratory manner, one decreases the opportunity for discovering unexpected data patterns that suggest new problem areas or alternative interpretive frameworks. The importance of constrained exploratory data analysis in the discovery of new problem domains and in escaping the blinding "tyranny" of an accepted theoretical framework and paradigm has been stressed by many authors (e.g., Hanson, 1972; Tukey, 1980; Clarke, 1972:8; Binford & Sabloff, 1982).

It also is important—in a groping, growing discipline like archaeology—that the analytic logic by which data and technique have been selected and justified be reported as much as possible, instead of reporting only succinctly statable deductive arguments or none at all. Many of the arguments that are used by a researcher to select certain observations, variables, or techniques imply (if they are not stated as such) formal *bridging arguments* (Hempel, 1966, pp. 72-75) that link the nonobservables of a theoretical framework to observables. Importantly, they involve the assumed nature of the phenomenon of interest. In an established discipline, where these bridges are well known and part of accepted theory and methodology, it is superfluous and too expensive to repeatedly report their use. However, in a quickly growing discipline where such bridges are not yet formalized and accepted, it is critical that they be stated explicitly and reported openly for criticism. *It is partly through explicit justification of the observations, variables, and techniques that are used in analysis and criticism of such justifications that the strong bridges between theory, method, and data, which typify a well established discipline, are built and communicated to the discipline at large.* Also, because such justifications pertain to the relevant structure of a data set, and thereby to the nature of the phenomenon of interest, *their refinement and formalization as bridging principles leads to or goes hand in hand with the refinement of theory and scientific advance* (e.g., Read, 1974).

Finally, in circumstances where bridging arguments are not well formalized or generally accepted, explicit statements of justification of the variables, observations, and techniques that are used in an analysis must be reported, if the analysis is to be assessable for its validity.

CONCLUSION

As a result of human limits to pattern recognition, the analysis of complex data sets cannot proceed with the same logic as the analysis of simple data sets. A sequential approach is required. This allows data structure and pattern-

searching technique to become discordant with each other and causes a problem, for the researcher, of how to get into data without violating its relevant structure. Several solutions to this problem are discussed in this chapter. All can be summarized in a word: JUSTIFICATION. The analysis of complex data requires the researcher to make a conscientious attempt to justify explicitly—through deductive or inductive argument—the relevance of the variables and observations that are brought forward for analysis to the phenomenon of interest. It also requires the researcher to justify the chosen analytic technique in relation to what is known about the data set's relevant relational structure, which is a reflection of the nature of the phenomenon of interest. Only when data and technique are concordant with each other and the phenomenon of interest can the results of an analysis accurately represent the phenomenon of interest, in turn laying the foundation for the assignment of appropriate meaning to the results.

"Fine tuning" of an analytic design, which involves the explicit justification of data and technique, however, has value beyond the scope of any single analysis. It is a critical aspect of the process of scientific advance. It is one means by which logical inconsistencies and false premises in current theory are uncovered and inadequacies in traditionally used analytic techniques are unveiled, and, hence, is a driving force behind the formulation of new theories and techniques.

REFERENCES

Behrensmeyer, A.K., & Hill, A.P. (Eds.). (1980). *Fossils in the making: Vertebrate taphonomy and paleoecology.* Chicago: University of Chicago Press.

Binford, L.R. (1977). *For theory building in archaeology.* New York: Academic Press, Inc.

Binford, L.R., & Sabloff, J.A. (1982). Paradigms, systematics, and archaeology. *Journal of Anthropological Research 38*(2), 137-153.

Braun, D.P. (1977). *Middle Woodland-early Late Woodland social change in the prehistoric midwestern U.S.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Braun, D.P. (1979). Illinois Hopewell burial practices and social organization: A reexamination of the Klunk-Gibson mound group. In D. Brose & N. Greber (Eds.), *Hopewell archaeology: The Chillicothe conference* (pp. 66-79). Kent, OH: Kent State University Press.

Braun, D.P. (1980, April). *Neolithic regional cooperation: A midwestern example.* Paper presented at the annual meetings of the Society for American Archaeology, Philadelphia, PA.

Braun, D.P., & Plog, S. (1982). Evolution of "tribal" social networks: Theory and prehistoric North American Evidence. *American Antiquity, 47*(3), 504-525.

Carr, C. (1977). *The internal structure of a Middle Woodland site and the nature of the archaeological record.* Unpublished preliminary examination paper, University of Michigan, East Lansing.

Carr, C. (1981, April). *The polythetic organization of archaeological tool kits and an algorithm for defining them.* Paper presented at the annual meetings of the Society for American Archaeology, San Diego, CA.

Carr, C. (1982a, April). *Dissecting intrasite artifact distributions as palimpsests.* Paper presented at the annual meetings of the Society for American Archaeology, Minneapolis, MN.

Carr, C. (1982b). *Handbook on soil resistivity surveying: Interpretation of data from earthen archaeological sites.* Evanston, IL: Center for American Archaeology.

Carr, C. (1983, June). *A design for intrasite research.* Paper presented at the National Park Service Research Seminar in Archaeology, Fort Collins, CO.

Carr, C. (1984). The nature of organization of intra-site archaeological records and spatial analytic approaches to their investigation. In M.B. Schiffer (Ed.), *Advances in archaeological method and theory* (Vol. 7) (pp. 103-222). New York: Academic Press.

Carr, C. (in press). Dissecting intrasite artifact palimpsests using Fourier methods. In S. Kent (Ed.), *Method and theory for activity area research. An ethnoarchaeological approach* (chap. 5). New York: Columbia University Press.

Christenson, A.L., & Read, D.W. (1977). Numerical taxonomy, R-mode factor analysis, and archaeological classification. *American Antiquity 42*(2), 163-179.

Clark, G.A. (1982). Quantifying archaeological research. In M.B. Schiffer (Ed.), *Advances in Archaeological Method and Theory* (Vol. 5) (pp. 217-273). New York: Academic Press, Inc.

Clarke, D.L. (1972). Models and paradigms in contemporary archaeology. In D.L. Clarke (Ed.), *Models in archaeology* (pp. 1-60). London: Methuen Publications.

Cowgill, G.C. (1977). The trouble with significance tests and what we can do about it. *American Antiquity 33*, 367-375.

Cowgill, G.C. (1982). Clusters of objects and associations between variables: Two approaches to archaeological classification. In R. Whallon & J.A. Brown (Eds.), *Essays on archaeological typology* (pp. 30-55). Evanston, IL: Center for American Archaeology Press.

Doran, J.E., & Hodson, F.R. (1975). *Mathematics and computers in archaeology.* Cambridge, MA: Harvard University Press.

Doran, J.E., & Hodson, F.R. (1982). Some aspects of archaeological classification. In R. Whallon & J.A. Brown (Eds.), *Essays on archaeological typology* (pp. 21-29). Evanston, IL: Center for American Archaeology Press.

Friedrich, M.H. (1970). Design structure and social interaction: Archaeological implications of an ethnographic analysis. *American Antiquity 35*, 332-343.

Gladwin, H. (1975). Looking for an aggregate additive model in data from a hierarchical decision process. In S. Plattner (Ed.), *Formal methods in economic anthropology* (Special publication) (pp. 159-196). Washington, DC: American Anthropological Association.

Hanson, N.R. (1972). *Patterns of discovery.* Cambridge: Cambridge University Press.

Hartwig, F., & Dearing, B.E. (1979). *Exploratory data analysis.* Beverly Hills: Sage Publications.

Hemple, C.G. (1966). *The philosophy of natural science.* Englewood Cliffs, NJ: Prentice Hall.

Hinkle, K.A. (1983). *Ohio Hopewell textiles: A medium for stylistic and social information exchange.* Unpublished master's thesis, University of Arkansas, Fayetteville.

Hole, B.L. (1980). Sampling in archaeology: A critique. *Annual Review of Anthropology 9*, 217-234.

Moore, J.A., & Keene, A.S. (1983). Archaeology and the law of the hammer. In S.A. Moore & A.S. Keene (Eds.), *Archaeological hammers and theories* (pp. 3-13). New York: Academic Press, Inc.

O'Shea, J. (1981). Social configurations and the archaeological study of mortuary practices: A case study. In R. Chapman, I. Kinnes, & K. Randsborg (Eds.), *The archaeology of death* (pp. 39-52). Cambridge: Cambridge University Press.

Plog, S. (1976). Measurement of prehistoric interaction between communities. In K.V. Flannery (Ed.), *The early Mesoamerican village* (pp. 255-272). New York: Academic Press, Inc.

Plog, S. (1978). Social interaction and stylistic similarity: A reanalysis. In M.B. Schiffer (Ed.), *Advances in archaeological method and theory* (Vol. 1) (pp. 144-182). New York: Academic Press, Inc.

Plog. S. (1980). *Stylistic variation in prehistoric ceramics.* Cambridge: Cambridge University Press.

Raab, M.L., & Goodyear, A.C. (1984). Middle-range theory in archaeology: A critical review of origins and applications. *American Antiquity 49*(2), 255-268.

Read, D.W. (1974). Some comments on the use of mathematical models in anthropology. *American Antiquity 39*(1), 3-15.

Reidhead, V. (1979). Linear programming models in archaeology. *Annual Review of Anthropology 8*, 543-578.

Scheps, S. (1982). Statistical blight. *American Antiquity 47*(4), 836-850.

Schiffer, M.B. (1975). Factors and "tool kits": Evaluating multivariate analysis in archaeology. *Plains Anthropologist 20*, 61-70.

Schiffer, M.B. (1976). *Behavioral archaeology*. New York: Academic Press, Inc.

Schiffer, M.G. (1983). Toward the identification of formation processes. *American Antiquity 48*(4), 675-706.

Shackley, M.L. (1978). The behavior of artifacts as sedimentary particles in a fluvial environment. *Archaeometry 20*, 55-61.

Spaulding, A.C. (1977). On growth and form in archaeology: Multivariate analysis. *Journal of Anthropological Research 33*, 1-15.

Spaulding, A.C. (1982). Structure in archaeological data: Nominal variables. In R. Whallon & J.A. Brown (Eds.), *Essays on archaeological typology* (pp. 1-20). Evanston, IL: Center for American Archaeology Press.

Speth, J.D., & Johnson, G.A. (1976). Problems in the use of correlation for the investigation of tool kits and activity areas. In C. Cleland (Ed.), *Cultural change and continuity* (pp. 35-75). New York: Academic Press, Inc.

Stanislawski, M.B., & Stanislawski, B.B. (1978). Hopi and Hopi-Tewa ceramic tradition networks. In I. Hodder (Ed.), *The spatial organization of culture* (pp. 61-76). Pittsburgh: University of Pittsburgh Press.

Thomas, D.H. (1971). On use of cumulative curves and numerical taxonomy. *American Antiquity 36*, 206-209.

Thomas, D.H. (1978). The awful truth about statistics in archaeology. *American Antiquity 43*, 231-244.

Toussaint, G.T. (1978). The use of context in pattern recognition. *Pattern Recognition, 10*, 189-204.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, J.W. (1979). Comment to "Nonparametric statistical data modeling." *Journal of the American Statistical Association 74*, 121-122.

Tukey, J.W. (1980). We need both exploratory and confirmatory. *American Statistician 34*(1), 23-25.

Tukey, J.W., & Wilk, M.B. (1970). Data analysis and statistics: Techniques and approaches. In E.R. Tufte (Ed.), *The quantitative analysis of social problems* (pp. 370-390). Reading, MA: Addison-Wesley.

Vierra, R.K., & Carlson, D.L. (1981). Factor analysis, random data, and patterned results. *American Antiquity 46*(2), 272-283.

Voss, J.A. (1980a). *Tribal emergence during the Neolithic of northwestern Europe*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Voss, J.A. (1980b, April). *The measurement and evaluation of change in the regional social networks of egalitarian societies: An example from the Neolithic of northwestern Europe*. Paper presented at the annual meetings of the Society for American Archaeology, Philadelphia.

Whallon, R. (1979, April). *Unconstrained clustering in the analysis of spatial distributions on occupation floors*. Paper presented at the annual meetings of the Society for American Archaeology, Vancouver.

Whallon, R. (1984). Unconstrained clustering for the analysis of spatial distributions in archaeology. In H.J. Hietala (Ed.), *Intrasite spatial analysis* (pp. 242-277). Cambridge: Cambridge University Press.

Wiessner, P. (1983). Style and social information in Kalahari San projectile points. *American Antiquity 48*(2), 253-275.

Wobst, M.H. (1977). Stylistic behavior and information exchange. In C. Cleland (Ed.), *For the director: Research essays in honor of James B. Griffin* (Anthropological papers 61) (pp. 317-342). Ann Arbor: University of Michigan, Museum of Anthropology.

For Concordance in Archaeological Analysis

BRIDGING DATA STRUCTURE, QUANTITATIVE TECHNIQUE, AND THEORY

Christopher Carr
GENERAL EDITOR

WESTPORT PUBLISHERS, INC.
in cooperation with the Institute for Quantitative Archaeology,
University of Arkansas

LIBRARY OF CONGRESS CATALOGING IN PUBLICATION DATA

Main entry under title:

For concordance in archaeological analysis.

Includes bibliographies and index.

1. Archaeology—Statistical methods—Addresses, essays, lectures.
I. Carr, Christopher, 1952- . II. University of Arkansas, Fayetteville. Institute of Quantitative Archaeology.
CC81.F66 1985 930.1'028'5 85-8861
ISBN 0-933701-00-4

Printed in the United States of America