UNIVERSITY OF CALIFORNIA, SAN DIEGO

Expertise and Mixture in Automatic Causal Discovery

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Philosophy

by

Joseph Daniel Ramsey

Committee in charge:

Clark Glymour, Chair
Patricia Churchland
Paul Churchland
Charles Elkan
Paul Kube
Gila Sher

2001

The dissertation of Joseph Daniel Ramsey is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

____

____
Chair

University of California, San Diego

2001

To my mother, Peggy

And two brothers, Steve and Mike.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

For this project I owe special thanks to two people in particular: Clark Glymour, advisor for the dissertation, and Ted Roush, source of much needed advise in rock spectroscopy and subject of our experiment in Chapter 4. Suffice it to say that without Clark's wise and energetic counsel, this project would never have begun, would never have taken shape, and would never have been finished. Similarly, without Ted's willingness to serve as experimental subject and without his boundless advise on matters scientific, this project would have had no point. These debts cannot easily be overstated.

Two departments of philosophy supported me for many years of study and work. To the Department of Philosophy at the University of California at San Diego I owe thanks for all that I've learned there, which is considerable. I owe thanks to nearly everyone in that department from the period of 1990 through 1998, but I owe special thanks to Henry Allison, Patricia Churchland, Paul Churchland, Gerald Doppelt, Clark Glymour, Patricia Kitcher, Philip Kitcher, and Gila Sher, each for a variety of reasons. I am also indebted to the department itself for supporting me financially through a difficult period of illness.

To the Department of Philosophy at Carnegie Mellon University, I owe thanks for the opportunity to partipate in a range of challenging research projects and to develop skills and intuitions in computer science. I am grateful to Wilfried Sieg, Clark Glymour, and Richard Scheines for arranging financial support during the period of 1998 to 2001 and to those faculty interested in causal search procedures (most especially Clark Glymour, Peter Spirtes, and Richard Scheines) for finding ways to include me in

their projects, from which I have gained immensely.

I wish to express thanks to the members of my committee for reading and commenting on this dissertation. Their feedback has been exremely valuable to me and will no doubt find its way into future work.

I am indebted to NASA Ames for the support they've lent to the research presented in this dissertation. Aside from financial support, through NASA Ames Cooperative Agreement NCC2-1026, we have received immense personal assistance not only from Ted Roush, mentioned above, but also Paul Gazis, whose algorithms and expertise I've taken advantage of.

Finally, I wish to thank my partner Eugene, whose love and support has been instrumental in all of my successes for many years now.

VITA

April 6, 1964	Born, Washington, Pennsylvania
1986	B.S. Indiana University of Pennsylvania
1987–1988	Overseas Student, Nanjing University, People's Republic of China
1990–1998	Teaching Assistant/Research Fellow Department of Philosophy University of California, San Diego
1995	M.S., University of California, San Diego
1998–1999	Research Programmer Department of Philosophy Carnegie Mellon University
1999–2001	Research Programmer/Director of Computing Department of Philosophy Carnegie Mellon University
2001	Ph.D., University of California, San Diego

PUBLICATIONS

Automated Remote Sensing with Near Infrared Reflectance Spectra: Carbonate Recognition.. (With P. Gazis, T. Roush, P. Spirtes, C. Glymour.) In preparation.

ABSTRACT OF THE DISSERTATION

Expertise and Mixture in Automatic Causal Discovery

by

Joseph Daniel Ramsey

Doctor of Philosophy in Philosophy

University of California San Diego, 2001

Clark Glymour, Chair

Critics of automatic causal discovery have claimed that Tetrad-style algorithms are inferior to domain experts at discovering causal structure from real scientific data and especially poor when applied to data that is highly mixed, either in a physical sense or in a mixtures of records sense. We compare a domain expert in geological spectroscopy head-to-head with a variety of machine algorithms on the task of predicting mineral class composition from visual to near infrared reflectance spectra. A simplified Tetrad algorithm ("modified PC") outperforms all other machine algorithms tested and performs comparably to the domain expert. We conclude (1) that Tetrad algorithms can perform as well as human experts on this task, and (2) that mixtures do not necessarily undermine the reliability of Tetrad algorithms on this task. This constitutes a counterexamples to the claim that machine algorithms are necessarily inferior to domain experts on tasks involving causal reasoning with real scientific data.

Chapter 1

Basic Concepts

1.1 Tetrad in "Normal Science" Contexts

According to a recent reviewer in *Cell* Magazine,[1] there is a change underway in how experiments are done in cell biology. On the way out is a reliance on the "one hypothesis/one experiment" approach. On the way in is a growing reliance on the "many hypotheses/many experiments" approach—in the following sense. Instead of designing specific experiments to test specific hypotheses, large programs of experiments are carried out to collect huge databases of information, that are then subsequently used to test numerous hypotheses, with followup experiments to fill in missing details. The order of experiment and hypothesis is effectively reversed; on the former approach, data are acquired after hypotheses are postulated, whereas on the latter approach, data are acquired before the relevant hypotheses are even thought of and are used to generate hypotheses. As a result, data mining techniques are increasingly being used to do basic

[1] See Vidal (2001).

science. The use of large databases and data mining techniques for this purpose is by no means limited to cell biology; rather, it's a practice that is growing in popularity across the sciences. As databases grow, our understanding of them shrinks, and as our understanding of data shrinks, data mining techniques become increasingly important as a way of telling us what our data means.

The method of collecting large databases that we barely understand, in the hopes that future colleagues will possess the concepts needed to understand them, is still new enough to be considered alternative, but it is no longer really new. The number of database-building experimental projects is already impressive across the sciences, and it is growing. There are stellar examples in cell biology. The project to sequence the yeast genome (*s. cerevisiae*) has been completed for some time now, and several databases are available with the information resulting from this effort. The project to sequence the human genome has recently been completed as well, after enormous effort by myriad scientists. Other genome sequencing projects for other organisms are on their way as well, in addition to data-gathering projects of other sorts, e.g., protein function for the yeast genome and for other organisms. We find similar efforts in astronomy, medicine, neuroscience, biology, planetology, geological physics, and many other fields. This is a way of working in science which is proving to be effective.

Causal information is one of the most important types of information scientists will need to mine from these large data sets. We need to know whether or not causal connections exist among variables, because this teaches us how to manipulate the world and therefore aids in our understanding of how the world works. With large databases, the usual recommendation of statisticians, that we do controlled experiments to discover causal connections between variables, is unhelpful. The content of these databases has already been measured—we don't have the opportunity to manipulate the relevant variables again for the measured data in an experimental context. Also, controlled experimentation over the variables in a database is limited by the number of variables we can consider at once. In a context where we only want to know whether one gene's expression affects one protein's level (say in a yeast cell), it's plausible that we could do a controlled experiment of our own (if we knew how) to see whether the causal relation obtains. However, if what we want to know is the pattern of causal connectivity over a collection of genes and proteins, we need to devise more sophisticated techniques for detecting these patterns, based on the kinds of data that are available (or that can plausibly be obtained). For this kind of inference about causal structure from data, some kind of automated algorithm is required.

Two types of well-tested causal inference algorithms in the literature are: (a) Bayesian algorithms, and (b) constraint-based algorithms. Bayesian algorithms (less well-known among philosophers) use Bayes' rule[2] to determine plausible causal connectivity from data. An excellent example of how this might be done is an algorithm by Cooper and Herskovits (Cooper and Herskovits 1991; Cooper and Herskovits 1992). Constraint-based algorithms, by contrast, test constraints that true patterns of causal connectivity (causal graphs) tend to obey with respect to observed data. One constraint-based approach (quite well known among philosophers) is the approach set forth in *Causation, Prediction, and Search* (Spirtes, Glymour, and Scheines 2000).[3] This approach

[2] $P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{m} P(B_i)P(A|B_i)}$.

[3] This is the second edition, which contains most of the text of the first edition (Spirtes, Glymour, and Scheines 1993) plus a large additional chapter (Chapter 12) detailing developments since the first

has been implemented in a series of computer programs entitled "Tetrad"; for shorthand, we will refer to the theory behind the programs, the algorithms, and the programs themselves[4] as "Tetrad" ("Tetrad theory," "Tetrad algorithms" or "Tetrad programs," respectively). The Tetrad algorithms (PC, FCI, etc.) perform searches over possible causal structures guided primarily by information about the conditional independencies of the variables being searched over; given data that is properly collected from a statistical point of view, the results of these algorithms can be remarkably insightful. There are, however, numerous causal inference problems from scientific contexts for which the Tetrad algorithms are known to fail or have difficulty. It is helpful, therefore, to develop a sense of just how well these algorithms can perform under difficult inference conditions from normal science.

1.2 Ideal Manipulation and the Causal Markov Condition

Tetrad theory, much to the consternation of critics, does not give any definition of causality, but it does offer a helpful intuition about what makes causal graphs useful for causal analysis—viz., a general account of manipulation.[5] When we can change the value of Y by fixing the value of X, and when we can do this reliably over many experiments, it is generally safe to conclude that X is a cause of Y. (Whether the cause is direct or indirect is a further question.) There is some risk in concluding this, since there may be conceptions of causal metaphysics on which we will be judged to have gotten the

edition.

[4] For an explanation of the Tetrad II program, see Scheines, Spirtes, Glymour, and Meek (1994).

[5] For a mathematical account of manipulation from the point of view of Tetrad theory, see the Manipulation Theorem (below, and Spirtes, Glymour, and Scheines (2000), Chapter 3).

wrong answer. Nevertheless, there exists a long-standing tradition in statistics, applied in a variety of sciences for many years, that identifies causal relationships in precisely this way—namely, the tradition of controlled experiments, argued for classically by Fisher (1951). It is a risk, though not a dangerous risk, to see manipulability as a hallmark of causality: Far more often than not this view will yield the correct answers about causal relationships, even when we disagree among ourselves about what the metaphysics of causality is. But Tetrad theory is a data mining theory for causal relations; it is not typically concerned with finding causal connections between pairs of variables but rather among networks of many variables.[6] How does the notion of manipulation for two variables extend to a system of many variables with a possibly complex network of interconnections?

Suppose we have a system of variables that we can manipulate ideally—that is, we can intervene from outside and alter the state of any subset of these values so that the system takes on whatever combination of values we please (within the legal ranges of values for each variable). If we could do this, then we would be able arbitrarily or randomly to set the values of our chosen variables to any experimental setup we choose, but otherwise retain any functional relationships that the variables may bear to one another.

Imagine all of the idealized randomized experiments that might then be done. With respect to the system of variables of the experiment, an intervention to manipulate

[6] For two variables alone, Tetrad turns out not to be very helpful, or at least not as helpful as controlled experiments. If we assume that there are no latent variables and run a Tetrad algorithm on data with variables X and Y, the most it can tell us is whether a causal connection exists between X and Y; it cannot tell us the nature or direction of that causal connection.

some variable X might, for some values of the intervention, alter the value of some other variable Y, with values for other variables in the system fixed. If we perform such an experiment for each pair of variables in the system, we can recover information about which variables are causes of which other variables in a way that is consistent with the intuition of manipulation that statisticians rely upon for binary systems of variables.

As an example, assume that we perform experiments like this for three variables, X, Y, and Z, and we discover that X causes Y, Y causes Z, and X causes Z. We can assemble this information into a directed acyclic graph (DAG)[7] as shown in Figure 1.1(a). We may wish, however, to distinguish between the following two possible reasons for X causing Z. It may be that X's influence on Z is entirely due to its influence on Y, in which case the alternative graph in Figure 1.1(b) would be more appropriate. The alternative graph still represents the fact that X causes Z; the influence of X on Z, however, is graphically shown to be indirect. On the other hand, it may be that the influence of Y on Z does not adequately characterize the variation of values of Z in the data and that values of X must also be considered, as an additional cause of Z. In this case, the first graph would be the more informative representation. While it is possible to assemble the information from an ideal randomized experiment into a causal graph in which all causes between variables V_1 and V_2 in the graph are represented as directed edges $V_1 \rightarrow V_2$, by distinguishing between direct causes and indirect causes we can produce a graph that is less cluttered and more informative.

Partly as a way of distinguishing direct from indirect causes and partly as a

[7] By way of definition, a directed graph is a graph containing only directed edges ($A \rightarrow B$); a cycle in a directed graph is a path $X_1 \rightarrow X_2 \rightarrow ... \rightarrow X_n$ such that $X_1 = X_n$. A directed acyclic graph (DAG) is a directed graph containing no cycles.

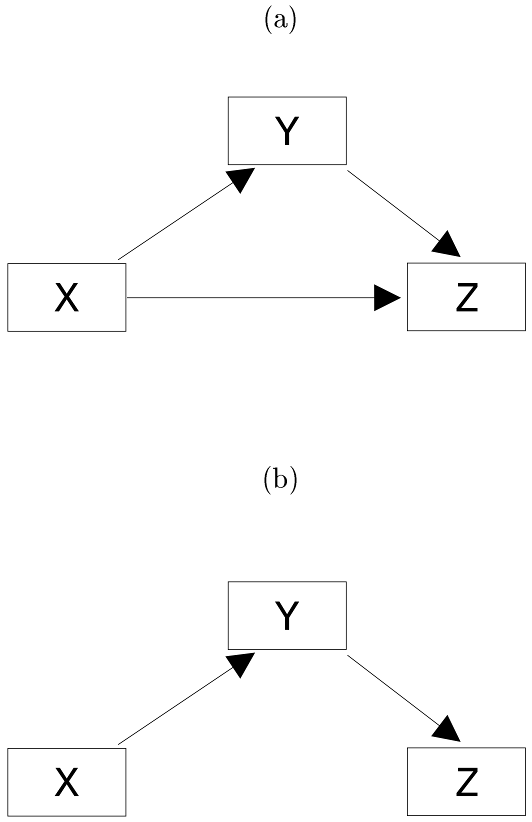


Figure 1.1: Two graphical representations of an idealized manipulation experiment where X causes Y directly, Y causes Z directly, and X causes Z indirectly, such that the influence of X on Z is entirely due to its influence on Y. In (a), both direct and indirect causal relationships are shown. In (b), only direct causal relationships are shown.

way of distinguishing causes from non-causes and of choosing orientations for causal edges, Tetrad theory takes for granted a general condition (and its converse) that relates graphical structure to statistical properties of underlying distributions. This condition is known as the "Causal Markov Condition," and its converse is known as the "Faithfulness Assumption." Both the Causal Markov Condition and the Faithfulness Assumption restrict the patterns of conditional independence relations that are assumed to hold of causal graphs with respect to the probability distributions they represent. Conditional independence is given a standard definition as follows: First, two random variables X and Y are independent (written hereafter as $X \perp\!\!\!\perp Y$) if the joint density of the product space $X \times Y$ is the product of the density of X and the density of Y for all measurable sets of values of X and Y. X and Y are independent conditional on a set of random variables $\mathbf{Z}$ (or "given" $\mathbf{Z}$; written $X \perp\!\!\!\perp Y \mid \mathbf{Z}$)) if the density of X and Y conditional on any measurable set $\mathbf{z}$ of values of $\mathbf{Z}$ equals the product of the density of X conditional on $\mathbf{z}$ and the density of Y conditional on $\mathbf{z}$, for all measurable sets of values of X and Y and for all measurable sets of values $\mathbf{z} \in \mathbf{Z}$, for which the density of $\mathbf{z}$ is not equal to zero. For sets of random variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ (written $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$)) if x is independent of y given $\mathbf{Z}$ for each $x \in \mathbf{X}$ and $y \in \mathbf{Y}$. Using this notion of conditional independence, the Causal Markov Condition may be stated as follows:

Definition 1 (Causal Markov Condition) *Let G be a causal graph with vertex set $\mathbf{V}$ and P be a probability distribution over the vertices in $\mathbf{V}$ generated by the causal structure represented by G. G and P satisfy the Causal Markov Condition if and only* *if for every W in* $\mathbf{V}$*,* W *is independent of* $\mathbf{V}\backslash(\mathbf{Descendants}(W) \cup \mathbf{Parents}(W))$ *given* $\mathbf{Parents}(W)$.[8]

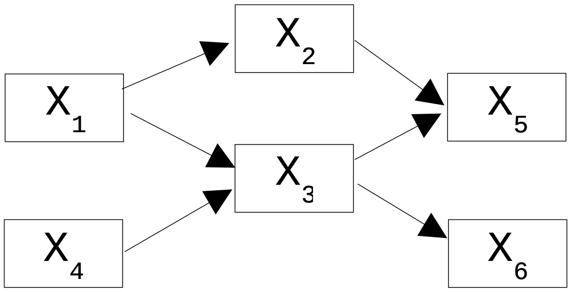
This condition relates the graphical structure of a causal graph to conditional independence facts about the probability distribution it represents.

To illustrate the way in which the Causal Markov Condition implies facts about conditional independence for graphs, consider the graph in Figure 1.2, over the variables $\{X_1, X_2, X_3, X_4, X_5, X_6\}$. Some of the conditional independence facts implied by the Causal Markov Condition for this graph are shown in Figure 1.2.

The converse of the Causal Markov Condition is the Faithfulness Assumption. Whereas the Causal Markov Condition asserts of a causal graph and an underlying distribution that a certain list of conditional independencies implied by the graph must hold in that underlying distribution, the Faithfulness condition asserts that any conditional independence fact that holds in the underlying distribution is implied by the Causal Markov Condition. When both the Causal Markov Condition and the Faithfulness Assumption are satisfied, it follows that all of the conditional independence facts implied by the Causal Markov Condition (and only these) hold in the underlying distribution.

Pearl (1988) defines a property of graphs, known as "d-separation," that allows the conditional independence facts implied by the Causal Markov Condition to be calculated quite easily. (D-separation also turns out to be a useful generalization of the Causal Markov Condition, as it allows Tetrad-style algorithms to be defined over directed cyclic graphs for which the Causal Markov Condition fails.) The d-separation

[8] See Spirtes, Glymour, and Scheines (2000), p. 29. For node X in graph G, $Y \in \mathbf{Descendants}(X)$ if and only if $X = Y$ or there is a directed path from X to Y. $Y \in \mathbf{Parents}(X)$ if and only if $Y \to X$.



$\{X_1\} \perp\!\!\!\perp \{X_4\} \mid \{\}$

$\{X_2\} \perp\!\!\!\perp \{X_3, X_4, X_6\} \mid \{X_1\}$

$\{X_3\} \perp\!\!\!\perp \{X_2\} \mid \{X_1, X_4\}$

$\{X_4\} \perp\!\!\!\perp \{X_1, X_2\} \mid \{\}$

$\{X_5\} \perp\!\!\!\perp \{X_1, X_4, X_6\} \mid \{X_2, X_3\}$

$\{X_6\} \perp\!\!\!\perp \{X_1, X_2, X_4, X_5\} \mid \{X_3\}$

Figure 1.2: An example of a causal graph to illustrate the use of the Causal Markov Condition. Some conditional independencies implied by this graph are listed.

property may be defined in terms of the notion of a collider, as follows:

Definition 2 (Collider) *Let G be a directed acyclic graph and let $U = \langle V_1, V_2, ..., V_n\rangle$ be an undirected path in G. Then V' is a collider on U just in case there are two distinct edges on U containing V' and both are into V'.*[9]

Definition 3 (D-Separation) *Let G be a directed acyclic graph, X and Y vertices in G with $X \neq Y$, and W a set of vertices in G not containing X or Y. Then X and Y are* ***d-separated*** *given W if and only if there exists no undirected path U between X and Y such that: (i) every collider on U has a descendant in W; and (ii) no other vertex on U is in W.*[10] *(X and Y are d-connected given W just in case they are not d-separated given W.)*

D-separation can be used to determine whether particular conditional independence facts are implied by the Causal Markov Condition. It has been shown that if vertices X and Y are d-separated given a set of vertices $\mathbf{W}$ in a causal graph G, then the Causal Markov Condition implies that X is independent of Y given $\mathbf{W}$, and vice-versa.

Notice that assuming the Causal Markov Condition holds of causal graphs allows one to distinguish between the two representations of the causal graph portrayed in Figure 1.1 (a) and (b). In that figure, on the assumption that X influences Z only by influencing Y, the Causal Markov Condition will hold of the graph in Figure 1.1(b) but not of the graph in Figure 1.1(a). The Causal Markov Condition itself is therefore a way to distinguish direct from indirect causal connections for a given set of variables.

[9] See Spirtes, Glymour, and Scheines (2000), p. 10.

[10] See Spirtes, Glymour, and Scheines (2000), p. 14.

When doing causal analysis over a set of variables, it's possible that all of the relevant causes of the variables being analyzed are present in the given set. If this is the case, then the set of variables is considered to be *causally sufficient*. Sets of variables which are not causally sufficient are missing an explicit representation of one or more *latent* (or *unmeasured*) variables—i.e., variables which are relevant to a causal account of the measured variables in the set but which have not themselves been measured. Among such latent variables might be latent *common causes*—i.e., latent variables which are causal parents of two or more variables in the measured set.[11] Algorithms for causal discovery are faster and simpler when causal sufficiency can be assumed. Under many conditions of causal insufficiency, however, reliable causal discovery is still possible. Of the standard Tetrad algorithms, the PC algorithm is designed to operate under the assumption of causal sufficiency, whereas the FCI algorithm is designed to operate under the assumption of causal insufficiency.

The results of causal search algorithms such as the PC or FCI algorithms are typically not single causal graphs but rather sets of causal graphs which share some, but not necessarily all, features. In particular, the output of the PC algorithm can be represented as a *pattern*, and the output of the FCI algorithm can be represented as a *partial ancestrial graph*, or *PAG*. The notion of a pattern may be defined as follows.

Definition 4 (Pattern) Π *is a pattern representing class* Δ *of causal graphs just in case* Π *is a graph containing directed and undirected edges only and for* $G \in \Delta$*:*

1. *G has the same adjacency relations as in Δ;*

[11] Latent variables are generally represented in causal graphs using ovals rather than rectangles.

2. *If $A \to B$ is in Δ, then $A \to B$ is in G for all nodes $A, B \in \Delta$; and*

3. *Unshielded colliders in G are unshielded colliders in Δ, where an "unshielded collider" is a node Y along a path $\langle X, Y, Z \rangle$ in a graph such that X and Z are not adjacent in the graph.*

The notion of a PAG is somewhat more involved; it is fully defined in Spirtes, Glymour, and Scheines (2000), p. 300.

If we accept the Causal Markov Condition for causally sufficient systems, then it is possible to give a general account of manipulation using causal graphs, as summarized in the following theorem:

Theorem 5 (Manipulation Theorem) *Given directed acyclic graph G_{Comb} over vertex set $\mathbf{V} \cup \mathbf{W}$ and distribution $P(\mathbf{V} \cup \mathbf{W})$ that satisfies the Causal Markov Condition for G_{Comb}, if changing the value of $\mathbf{W}$ from $\mathbf{w_1}$ to $\mathbf{w_2}$ is a manipulation of G_{Comb} with respect to $\mathbf{V}$, G_{Unman} is the unmanipulated graph, G_{Man} is the manipulated graph, and*

$$P_{Unman(\mathbf{W})}(\mathbf{V}) = \prod_{X \in \mathbf{V}} P_{Unman(\mathbf{W})}(X \mid \mathbf{Parents}(G_{Unman}, X))$$

for all values of $\mathbf{V}$ for which the conditional distributions are defined, then

$$\begin{aligned} P_{Man(\mathbf{W})}(\mathbf{V}) &= \textstyle\prod_{X \in \mathbf{Manipulated(W)}} P_{Man(\mathbf{W})}(X \mid \mathbf{Parents}(G_{Man}, X)) \\ &\quad \times \textstyle\prod_{X \in \mathbf{V} \setminus \mathbf{Manipulated(W)}} P_{Unman(\mathbf{W})}(X \mid \mathbf{Parents}(G_{Unman}, X)) \end{aligned}$$

for all values of $\mathbf{V}$ for which each of the conditional distributions is defined.[12]

[12] See Spirtes, Glymour, and Scheines (2000), p. 51.

1.3 Failures of the Causal Markov Condition

Tetrad theory makes two basic assumptions, embodied by the Causal Markov Condition (or, more generally, causal d-separation) and Faithfulness Assumption. First, it assumes that if variables are constrained to include all common causes directly affecting two or more variables, or that if the population is a collection of causally sufficient systems with the same graph, then the probability relations among variables in that system satisfy the d-separation criterion for the causal graph that would be constructed by the idealized randomized experiment described above. Second, it assumes that associations do not cancel one another perfectly, which would prevent d-separation from computing all causal relations among a set of variables. In addition, there may exist more specific assumptions for particular algorithms—for instance, we may assume that causal graphs must be acyclic, or that the probability distributions from which distributions frequencies are drawn are from particular families (e.g., normal, multinomial, etc.), or that dependencies between variables take on particular functional forms (e.g., linear, as with structural equation models).

Any of these assumptions could be in error; there is no “transcendental deduction” that the Tetrad procedures are valid. If these assumptions are in error, problems in the application of the algorithms will arise.

Problems with the Causal Markov Condition alone are interesting in their own right. It’s possible, for a given set of data, that the Causal Markov Condition will not be satisfied for any causally sufficient graph over the variables in that data set. Interesting problems arise where observed frequencies don’t satisfy the Causal Markov

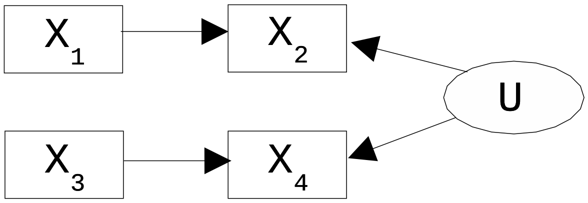


Figure 1.3: A simple graph with a latent variable to illustrate how a Tetrad search may find no causal graph satisfying the Causal Markov Condition under conditions of causal insufficiency.

Condition for any causally sufficient structure. Much algorithmic work in the last decade has been devoted to determining, on the assumption that data is ultimately derived from a distribution satisfying the first and second assumptions, how the Causal Markov Condition might not be satisfied for the observed frequencies, no matter what causal graph we choose to postulate for the data.

The first and most obvious way in which observed frequencies might not satisfy the Causal Markov Condition for any causal graph is if there are unrepresented latent variables in the graph. Consider the graph in Figure 1.3. In this graph, X_1, X_2, X_3 and X_4 are measured variables, and U is a latent variable. If we tried to construct a causal graph over just the measured variables, we would be thwarted, since an edge would need to be postulated between X_2 and X_4, in keeping with the Causal Markov Condition, but orienting the edge as either $X_2 \rightarrow X_4$ or $X_2 \leftarrow X_4$ would be inconsistent with the Causal Markov Condition. The underlying problem is that the true graphical structure contains a latent variable not included among the measured variables. Various techniques have been devised to detect the presence of latent variables under certain conditions—the FCI algorithm, for instance, is designed in part to do just this.

A second way in which observed frequencies might not satisfy the Causal Markov condition is that the sample data might itself be conditioned on a common effect. This is known as "sample selection bias," and it leads to spurious associations. This is another form of latent variable confounding.

A third way in which observed frequencies might not satisfy the Causal Markov Condition is if in estimating conditional independence we use the wrong probability distribution assumptions. For instance, vanishing conditional correlation is standardly used in structural equation models as an estimate of conditional independence. Conditional correlation is equivalent, however, to a form of simultaneous linear regression[13] and therefore assumes, for each variable X and each set of variables $\mathbf{Y}$ in a given causal graph, that there is an appropriate model regression of X onto $\mathbf{Y}$ that is linear. If this assumption fails, then tests for conditional independence using conditional correlation may fail as well. Tetrad algorithms rely on tests of conditional independence: When these tests produce false result, Tetrad algorithms may be able to find alternative graphs for data that satisfy the Causal Markov Condition, but it's at least as likely that they will be able to find no graph at all for the data that satisfies the Causal Markov Condition.

A fourth way in which observed frequencies might not satisfy the Causal Markov Condition is if the frequencies come from multiple systems with differing probability distributions. In this case, we can add a variable to represent which subpopulation the system being measured comes from, so the problem can be treated as a latent

[13] One typical way of calculating the correlation of X and Y conditional on $\mathbf{Z} = \{Z_1, Z_2, ..., Z_n\}$ is to linearly regress X onto $\mathbf{Z}$ and Y onto $\mathbf{Z}$ and then to correlate the two columns of residuals that result.

variable problem. There are well-known examples in the literature of mixed systems that yield unintuitive results not satisfying the Causal Markov Condition or the Faithfulness Assumption, such as the so-called "Simpson's Paradox." In this (hypothetical) example, the question is whether treatment leads to survival of a certain disease. Data from male and female subjects are compiled separately and then afterwards combined into one table (Table 1.1). For both males and females separately, it appears that treatment raises the chances of survival, but in the combined population treatment is independent of survival. On the principle that, if $X \rightarrow Y$ in each subsystem of a mixed system, it should the case that $X \rightarrow Y$ in the combined system, in Simpson's example the Faithfulness Assumption is violated in the combined system. Similarly, in the classic example due to Kendall,[14] a positive causal effect for males is exactly balanced by a negative causal effect for females, producing the perhaps unintuitive result that the causal effect for males and females cancels out in the mixture (Table 1.2). These are not, however, examples that undermine the usefulness of Tetrad theory. When we look at the possible ways of parameterizing a linear causal model and consider the subset of those parameterizations that represent probability distributions unfaithful to their true causal graphs, it turns out that this subset has Lebesgue measure zero.[15] More recently, results have been established that place hard limitations on the effectiveness of any causal search method from data, including causal searches by human experts with access to relevant background knowledge.[16]

[14] See Spirtes, Glymour, and Scheines (2000), p. 38-40.

[15] See Spirtes, Glymour, and Scheines (2000), pp. 41-42.

[16] See Spirtes, Glymour, and Scheines (2000), Chapter 12.

Table 1.1: Hypothetical data for Simpson's so-called "paradox," illustrating for mixed data how a positive causal effect in two subpopulations can disappear in the mixed population.

Males:

	Untreated	Treated
Alive	4	8
Dead	3	5

Females:

	Untreated	Treated
Alive	2	12
Dead	3	15

Combined:

	Untreated	Treated
Alive	6	20
Dead	6	20

Table 1.2: Hypothetical data for Kendall's example, illustrating how a positive and negative causal effect can cancel one another in a mixed population.

Males:

	Untreated	Treated
Recovery	100	80
Nonrecovery	80	40

Females:

	Untreated	Treated
Recovery	100	20
Nonrecovery	20	10

Combined:

	Untreated	Treated
Recovery	200	100
Nonrecovery	100	50

For these reasons and many others, causal searches over mixed systems present a challenging technical problem. It should be pointed out, though, that the problem is not universally insoluble, even analytically, since for some classes of problems analytic solutions exist. If we begin with a collection of linear causal models $\mathbf{M} = \{M_1, M_2, ..., M_n\}$ over the same variables $\mathbf{V}$, such that for no $X_1, X_2 \in \mathbf{V}$ is it the case that for $M_i \neq M_j \in \mathbf{M}$, $X_1 \rightarrow X_2$ in M_1 while $X_2 \leftarrow X_1$ in M_2, then it is reasonable to assert that the causal graph of the mixture is the union of the edges in the causal graphs of the subsystems M_i. In this case, searching over mixed data from the systems in $\mathbf{M}$ does reliably yield the causal graph of the mixture.

A fifth way in which observed frequencies might not satisfy the Causal Markov Condition is if we measure not the variables themselves but rather summations over instances of the variables. Consider a typical situation that arises in cell biology. One might wish to measure the expression of a certain gene for a particular cell type. It is usually difficult or even impossible to measure the expression of that gene in a particular cell. So instead of measuring gene expression in a particular cell, one measures the aggregate expression of that gene over a collection of cells, such as all the cells on a particular plate. The variable measured is not then an instance variable, but rather a summation variable. Causal models in this case might involve the causal effect of one summation variable on another summation variable. Again, if the causal models can safely be assumed to be linear, there are analytic reasons to believe that Tetrad searches should work correctly over such summation variables. If the linearity assumption turns out to be false, the situation is much more difficult to assess analytically.

In each of these cases it's plain that given a full reconstruction of the causal system generating the data, the Causal Markov Condition will hold in the reconstructed system. In the case of absent latent variables (i.e., causal insufficiency), once the latent variables are added back in, the Causal Markov Condition holds in the resulting system. In the case of sample selection bias, once we add a variable representing sampling, the Causal Markov Condition holds once again. In the third situation, where our model of the joint distribution of the system is incorrect, once we correct our method for calculating conditional independencies from the data, the Causal Markov Condition will again be satisfied. In the fourth situation, where we are taking measurements over mixtures of systems, the Causal Markov Condition might fail for a variety of reasons. If we distinguish the various subsystems from one another, however, the Causal Markov Condition should hold for each of the subsystems. The same should be true in the fifth situation, where we are taking measurements over summations of variables, even if separating out the individual variables may not be realistic (e.g., measuring the gene expression a single cell). In each of these cases, the Causal Markov Condition appears not to hold for the variables actually measured, though it does hold of the causal systems rightly considered.

The question, then, is whether there exist search procedures that, despite these failures of the Causal Markov condition, still provide some information about underlying causal structures of observed frequencies. This leads to a broader empirical question: Even when we're dealing with systems including complexities such as those outlined above, how much do these complexities matter for estimating the underlying causal structure?

There are two views on this issue in the literature. The first view is that such complexities almost always matter in real science, that humans have a unique, distinctive understanding of them, and that inquiry by humans not using search algorithm will almost always be more fruitful than inquiry by automatic search algorithms. This is a view urged, for instance, by Freedman and Humphreys,[17] who argue that in designing methods for causal analysis, we should avoid the "Automation Principle"—i.e., the belief that the only worthwhile knowledge is knowledge that can be taught to a computer. The implication of this criticism in the context of causal analysis is that humans must have abilities to discover causal relationships from data that causal discovery algorithms such as the Tetrad algorithms cannot hope to match. It is also a view urged, in less strident form, by Nancy Cartwright, who argues that while the use of computer algorithms to aid in causal discovery is not unreasonable, one should not expect a single type of algorithm to apply correctly across the various causal discovery contexts one normally encounters in science.[18] Different causal situations have different mathematical forms and require different statistical tests to establish causal relationships. This means that different causal discovery problems require different algorithms, and so methods need to be invented to deal with each type of causal discovery situation one encounters in turn. Human beings, not computers, are required to determine which types of algorithms will work in which contexts. On both Freeman and Humphreys' and Cartwright's view, the fact that the Causal Markov Condition can can fail so easily implies that human intervention is indispensible in the causal discovery process.

The second view of the implications of the Causal Markov complexities listed

[17] See Freedman and Humphreys (1999).

[18] See Cartwright (1999a), Cartwright (1999b).

above is that any knowledge relevant to search that humans actually possess can be entered into a computer as background knowledge and an automated search can then be conducted that is not only constrained by, but actually takes advantage of, that background knowledge. Some types of background knowledge can be exploited quite easily in a computational context, such as information about the time order of variables. Other types of background knowledge may require more effort to exploit. If we know, for instance, that partial conditional correlation is not a good estimate of conditional independence for a model (because, for example, relationships between variables cannot be modeled as linear), then before we can use Tetrad-style algorithms we must construct a better estimate of conditional independence. Nevertheless, adherents to this second view typically maintain that any measure of conditional independence that human experts can take advantage of can be programmed into a computer and used by a Tetrad-style algorithm and that the same comment could fairly be made of any other type of background knowledge relevant to causal search.

So the suggestion of this second view is that in "normal science" contexts, given the background knowledge humans have of a causal system, human experts do no better at deciphering causal structure than could computers that take advantage of the same background knowledge automatically, and computers decipher causal structure faster and more thoroughly. To test this hypothesis, one needs an established science with a long tradition of expertise, where a great deal of basic information remains to be discovered, where hypotheses can be independently tested, and where bona-fide experts in the field are willing to undergo direct comparisons of their own skills with the skills of computer programs at the task of causal analysis.

We will consider geological spectroscopy as such as field. The tradition of using spectroscopy to analyze the composition of substances goes back well into the nineteenth century, and the tradition of applying these techniques to mineralogy goes back at least to the 1940's. So there is certainly a long tradition of expertise in this field to draw on. Also, we have found a domain expert in the field, Ted Roush of NASA Ames, who has kindly allowed us to compare his expertise to that of a machine learning algorithm. We are grateful to Dr. Roush for giving us this opportunity. Finally, geological spectroscopy is a field in which sizeable amounts of data are available, in the form of spectra, for the mining of causal relations. There are numerous open problems involving causal analysis of rock and mineral spectra to which Tetrad-style algorithms can be applied. We will take advantage of this situation in upcoming chapters to construct a counterexample to certain claims about the limitations of automatic causal discovery, some of which have already been touched on briefly. In the next chapter, two problems in particular will be brought to focus—the role of expertise and the role of mixtures, with respect causal discovery using Tetrad algorithms.

Chapter 2

Issues and Criticisms

One might have thought that a clear representation of causal relations, together with a set of search strategies that dominates the most commonly used statistical search methods (i.e., varieties of regression), would be welcomed by causal reasoning methodologists. However, Bayes net strategies such as those briefly introduced in Chapter 1 have met with mixed reception. Some of the criticisms are principled. For instance, it has been objected that Bayes net search procedures do not allow for standard estimates of error, since automated search procedures do not output confidence intervals for Bayes nets. This is no accident; recent work by Robins *et al.*[1] has shown that no such error probabilities are possible, no matter what the search method may be—human or automated.[2] It is still true that search methods exist that converge to the partial ancestral graph of the true structure as the size of the sample increases without bound, as discussed in the previous chapter. But the Robins *et al.* results add as new information

[1] See Robins, Scheines, Spirtes, and Wasserman (1999).

[2] See Spirtes, Glymour, and Scheines (2000), Chapter 12.

that at every sample size alternative models exist, not represented by the PAG of the true model, whose marginal sampling distribution for that sample size over the observed variables is as close as you like to the marginal sampling distribution of the true Bayes net.

Other criticisms are less focused, but they are the subject of this essay. Freedman and Humphreys in several essays separately and jointly[3] have insisted that automated methods must be inferior to causal judgments of skilled scientific experts. They present no head to head comparisons, and their criticisms of correct predictions obtained from Bayes net search methods are based on misinterpretations of algorithmic output. For instance, Spirtes *et al.* use a small observational sample of measures from plugs of Spartina grass to predict (retrodict, actually) the outcome of a greenhouse experiment showing salinity having no influence on biomass when pH is held constant.[4] The Spirtes *et al.* prediction, incidentally, is contrary to the prediction of the biologist who conducted the experiment. Freedman and Humphreys object to the Spirtes *et al.* analysis that relations among other variables found in the output of the observational study are incorrect. But these connections involve features (binary effects of linear variables, complete graphical connections among multiple variables, etc.) that the manual for the Tetrad II program[5] states cannot be trusted to yield reliable information about causal connections. Nonetheless, while the claim by Freedman and Humphreys that human expertise always trumps Tetrad and related causal search methods—even

[3] See, e.g., Freedman and Humphreys (1999).

[4] See Spirtes, Glymour, and Scheines (2000), pp. 196-200.

[5] See Scheines, Spirtes, Glymour, and Meek (1994).

automated search in which explicit human knowledge of the domain has been encoded in the program—may be unsubstantiated, it is also essentially uncontradicted by any published data.

Nancy Cartwright has offered other criticisms. She claims, with a critical tone, that automated Bayes net methods cannot work universally,[6] which no one in the machine learning community has claimed of any algorithm, least of all the Bayes net learning algorithms. She claims, in the same tone, that the Causal Markov Condition is not a logical truth (i.e., that probability distributions and DAGs of causal graphs not satisfying the Causal Markov Condition are logically possible[7]), the contrary of which no one, to my knowledge, has ever suggested. She also claims that the correctness of the output of an automated search procedure is insufficient to guarantee the correctness of policy predictions obtained from the output structure using the Manipulation Theorem.[8] Her point is that any causal instruction, for example "Prevent people from smoking" can be implemented in many, many different ways, and the effects of the intervention will depend on the causal details of the situation—what she calls the "nomological machine." That seems entirely correct. She insists that what is required to obtain a correct policy prediction from a causal graph or causal Bayes net is the "stability" of the causal system,[9] knowledge of which can be had from knowledge of the "nomological machine." Her discussion of the Manipulation Theorem misrepresents it in an important way that

[6] See Cartwright (1999b), p. 104.

[7] See Cartwright (1999b), p. 107.

[8] See Cartwright (1999b), p. 104, p. 126.

[9] See Cartwright (1999b), p. 124.

need not concern us here,[10] but the upshot is that the "stability" she requires for correct prediction is exactly that the causal structure "downstream" from a manipulated variable not be altered by the manipulation. This is explicitly assumed in the definition of an "ideal manipulation" in the Manipulation Theorem, so the point of the objection becomes, once more, rather obscure.

A more substantive objection involves populations that are mixtures of units with different probabilities and causal structures. Cartwright argues that such systems will be mischaracterized by the automated Bayes search algorithms because even if two causal graphs over the same variables individually satisfy the Causal Markov Condition with respect to probabilities estimated from sample data, when the two samples are combined into one data set, the anticipated causal graph for the combined data will not satisfy the Causal Markov Condition. She illustrates her point using a hypothetical system of three variables X, Y, Z, over which two different mechanisms (indexed by the variable NM) operate separately.[11] When the first mechanism operates, X causes both Y and Z, as in Figure 2.1(a). When the second mechanism operates, there are no causal relations between X, Y, and Z at all, as illustrated in Figure 2.1(b). Cartwright hypothesizes data collected for these two mechanisms as shown in Table 2.1.[12] For the

[10] Cartwright (1999b) claims, on p. 127, that Spirtes, Glymour, and Scheines (1993) assume in the proof of the Manipulation Theorem that "$G_{Unman} = G_{Man}$ are subgraphs of G_{Comb}." She takes this to mean "that changes in the distribution of the externally manipulated 'switch variable' W are not associated with any changes in the underlying causal structure." However, this assumption is not made in the Manipulation Theorem; in fact, the point of the Manipulation Theorem is to show how manipulation of a causal structure via an external variable W *can* change the causal structure itself. For instance, if through W the value of a variable X in the causal structure is held fixed, then causal edges from the parents of X in the unmanipulated structure will be broken in the manipulated structure.

[11] See Cartwright (1999b), pp. 130–135.

[12] The numbers for these tables is taken from the charts and discussion on p. 132 of Cartwright (1999). They have been reformatted so that the comparison to Spirtes, Glymour, and Scheines (2000) will be more evident.

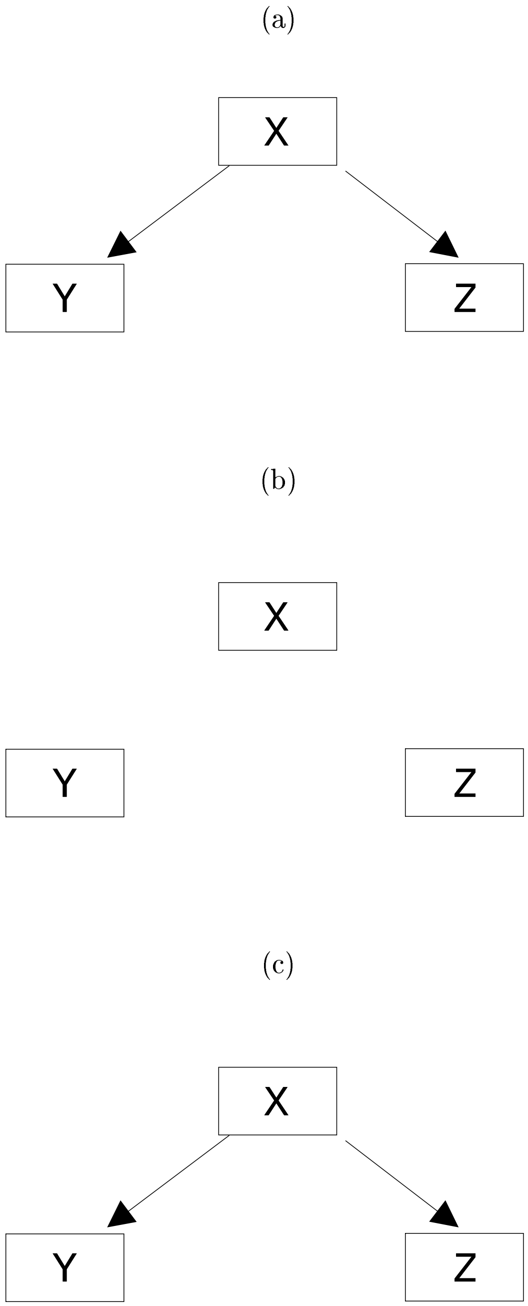


Figure 2.1: Causal graphs for the unmixed and mixed mechanisms Cartwright posits to illustrate how the Causal Markov Condition fails for mixed distributions. The first graph (a) represents the first mixed structure; the second graph (b) represents the second mixed structure; the third graph (c) represents the mixed structure. (An adaptation of diagrams by Cartwright in Cartwright (1999b), p. 131.)

Table 2.1: Hypothetical data for the mixed mechanisms Cartwright uses to illustrate how the Causal Markov Condition may fail in mixed distributions. The first table (a) is hypothetical data for the first mechanism; the second table (b) is hypothetical data for the second mechanism; the third table (c) is the data which results from mixing the records of (a) and (b). The graphs for these machanisms are shown in Figure 2.1 (a-c). (Adaptation of tables by Cartwright in Cartwright (1999b), p. 132.)

(a)

$NM = 1$	$Y = 1$		$Y = 2$	
	$Z = 1$	$Z = 2$	$Z = 1$	$Z = 2$
$X = 1$	200	0	0	0
$X = 2$	0	0	0	200

(b)

$NM = 2$	$Y = 1$		$Y = 2$	
	$Z = 1$	$Z = 2$	$Z = 1$	$Z = 2$
$X = 1$	50	50	50	50
$X = 2$	50	50	50	50

(c)

$(Combined)$	$Y = 1$		$Y = 2$	
	$Z = 1$	$Z = 2$	$Z = 1$	$Z = 2$
$X = 1$	250	50	50	50
$X = 2$	50	50	50	250

Causal Markov Condition to hold of the first mechanism, Y should be independent of Z given X. This is certainly true, given the data in Table 2.1(a). For the Causal Markov Condition to hold of the second mechanism, all possible conditional independence relations should hold in the data. Again, this is true of the data in Table 2.1(b). However, if we combine the data from Table 2.1(a-b) the result is shown in Table 2.1(c). The graph for the mixture should be the union of the graphs for the unmixed data—i.e., the graph in Figure 2.1(c). This implies that Y should be independent of Z given X, and this is clearly not true. For example, in order for this independence relation to hold, it would have to be the case that Y is independent of Z conditional on $X = 1$. The data for X and Z conditional on $X = 1$ are as follows:

	$Z = 1$	$Z = 2$
$Y = 1$	250	50
$Y = 2$	50	50

From this it is quite evident that the conditional independence relation does not hold. So Cartwright's contention is valid, that for these hypothetical[13] numbers, the Causal Markov Condition holds for the causal graphs of the two mechanisms separately (with respect to probabilities estimated from their respective data) but not for the combined causal graph (with respect to probabilities estimated from the combined data set).

Note that the output of the FCI and similar search procedures applied to the probabilities for the combined population would produce a complete PAG, because in the combined population no independence or conditional independence relations hold at

[13] It's important to keep track of which examples of failures of the Causal Markov Condition and/or Faithfulness Assumption are hypothetical and which are actually measured, because these examples are easy to construct, and yet the likelihood of such data being measured in the real world is theoretically insignificant. (See Spirtes, Glymour, and Scheines (2000), Chapter 12.)

all. Given the prior information that X occurs before Y and Z (and is therefore not influenced by Y or Z), the output of the FCI algorithm would be the graph shown in Figure 2.2(a). This PAG represents many alternative causal structures, among which is the graph shown in Figure 2.2(b), where U may be either an unobserved common cause or a parameter representing the fact that there are two subpopulations with different causal structures and/or probabilities. Note that these features, and examples similar to Cartwright's such as the Kendall example and Simpson's paradox,[14] are discussed fully in Spirtes, Glymour, and Scheines (1993) and that there is nothing new in Cartwright's discussion except a mischaracterization of their significance. When mixtures exist of the sort she considers, the output of the FCI algorithm will not be incorrect—it will be uninformative.[15]

There are other ways not discussed by Cartwright in which mixtures might be problematic for automatic causal search, e.g., the aggregation of variables, discussed in the previous chapter. Here, just as with the Cartwright's mixture example, the conditional independence relations among measured variables do not give us information about the conditional independence relations in the unerlying causal structure. Suppose, for example, that in each unit X causes Y, Y causes Z, and there is no other causal connection between X and Z. Suppose further that for each unit in a population, if for any unit one could do repeated measures of X, Y, and Z for that unit, the set of measurement triples would, in the limit, be distributed so that $X \perp\!\!\!\perp Z \mid Y$, for each value of X, Y, and Z. In short, our supposition is that the Causal Markov Condition holds

[14] See p. 17.

[15] Uninformative since it doesn't provide information directly about the proper interpretation of the double-headed arrow ($Y \leftrightarrow Z$).

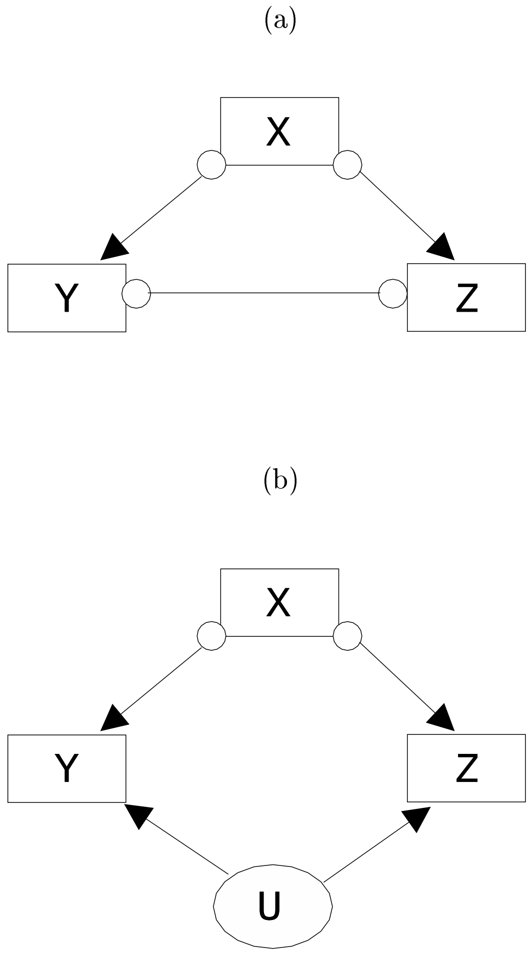


Figure 2.2: Results of FCI algorithm on Cartwright's combined data. The first graph (a) shows the results of the algorithm; the second graph (b) shows one graph in the equivalence class of (a).

for repeated measurements on each unit. Suppose, however, that our measurements are not actually of X, Y, and Z for individual units, but rather of the sum of X values, the sum of Y values, and the sum of Z values across all units in the population. Then, except in special cases, the sum of X and the sum of Z will not be independent of the sum of Y. (One such special case is when all dependencies are linear.[16]) Measurements of gene expression in most microarray data are, for example, aggregated in this fashion.

Discarding the straw men in Cartwright's discussion and the more or less deliberate misrepresentations in Freedman and Humphrey's discussion, one substantive, interesting issue remains: For complex systems, in which the data are likely to be produced by a mixture of causal processes or by aggregation of variables over many different units, and where there is a long tradition of human expert causal interpretation of data from such systems, do automated search procedures do as well or better than human experts? Of course, the answer may very well be different for different domains, and no single study can settle the relevant scientific policy questions: Should we, in a new domain or an old one, trust automated procedures; should automated procedures be our first recourse for likely hypotheses; should we trust them as much as or more than expert judgments based on the same data? Humphreys and Freedman in effect say, "never"—to all three questions. Cartwright's view on the same questions is less clear cut but at least clearly skeptical.

There are a great many comparisons of automated procedures with applied human experts where the goal is simply to recognize a feature or predict a feature of

[16] In that case aggregation can actually help in search, since the sums of independent random variables will, by the central limit theorem, approach normally distributed variables, so that $\sum X$, $\sum Y$, $\sum Z$ will look like linearly related, normally distributed variables.

behavior, without regard to causal constraints. Thus for fifty years it has been known that in predicting recidivism, for example, simple regression with adequate data does as well or better than parole boards, psychiatrists, or other experts. There are, however, few examples of head-to-head contests of human experts and expert systems where the issue is causal, scientific judgment. The *New England Journal of Medicine* published a comparison of the diagnostic accuracy of residents and experts in internal medicine, versus the *Internist* program developed at the University of Pittsburgh.[17] The program outperformed residents but not expert internists. The example is not, however, satisfactory for beginning to address the substantive issues about the effectiness of automated causal inference in complex mixed systems, for several reasons. The *Internist* algorithm was intensely specialized, aimed at internal medicine and nothing else, while the Bayes net search algorithms, like regression, are generic, and nothing about them or their ancestry tells us where they will work and where they will not. Further, the Internist algorithm was essentially an automated version of a person. It was based on rules and procedures of diagnosis extracted from a single expert physician at the University of Pittsburgh Medical Center.

The following chapters consider the substantive issue of how reliable a simple machine learning algorithm can be, compared to human expertise, in an area of physics where there is a long, well established tradition of expert practice, where the goal is recognition of components that play a causal role in generating the data, and the causal processes that produce the data are complex and may very well involve mixing, aggregation, and significant non-linearities. The domain is the identification of mineral

[17] See Miller (1982).

composition from visible/near infrared reflectance spectra, a technique that has been used in geophysics for 70 years. We will construct a simple Tetrad-style algorithm to solve this problem and treat it as a simple classifier alongside many other types of classifiers; this will allow us to take advantage, in a classification problem, of the sensitivity of Tetrad-style algorithms to causal structure. Discovering, as we shall, that given comparable data, simple, generic machine learning techniques perform this task as well or better than human experts will not, of course, establish that such techniques work everywhere, nor will it remove the demonstrations that the automated search procedures lose information in mixed, aggregated nonlinear systems. But it will at least raise the suspicion that in such contexts human experts lose information too, and perhaps more of it, than do their automated counterparts.

Chapter 3

The Spectrographic Investigation of Rocks and Minerals

3.1 Introduction

The identification of surface composition from reflectance spectra has traditionally relied on two methods. The older of the two is a direct examination of spectra by experts, seeking lines or bands characteristic of particular substances, sometimes taking account of overall luminosity of the spectrum, and sometimes, with computational aids, taking account of the shapes of bands. The standard alternative is simultaneous linear regression of an unknown spectrum against a library of known spectra for candidate materials. A number of spectral libraries have been compiled which can be used for this purpose. Some neural net procedures—notably Kohonen maps—have also been used to analyze spectral data, typically not for identifying surface composition directly, but rather for finding bounded regions of similar composition in an array of point spectra from a "visual" field.[1] Other automated techniques have been used explicitly to identify surface composition of minerals and rocks, including a Bayesian technique described below. Despite, however, its numerous applications for planetary and terrestrial exploration and for various military purposes, we have found no published systematic (or even unsystematic) study comparing automated examination of reflectance spectra to human expert examination of reflectance spectra.

So far as planetary exploration is concerned, reflectance spectroscopy techniques have already shown themselves to be useful. Visual to near infrared (VNIR) reflectance spectroscopy (from approximately 0.4 μm to approximately 2.5 μm) in particular has offered geologists an important potential source of petrological information for the exploration of planets, satellites, and other solar system objects. Lightweight, low-power commercial instrumentation is available, detailed physical models have been developed (e.g. Hapke (1993)), and data from VNIR instruments is routinely used by geological spectroscopists in practical mineral classification.[2] Were such instruments coupled with intelligent software for mineral classification from spectra, the resulting system could be used either for remote sensing or for surface based studies, reducing requirements for data storage and information transmission, and aiding autonomous, rational, scientifically-informed decisions by robot explorers about further directions for exploration and data acquisition.

The interest in planetary exploration motivates an examination of the problem of determining whether rock or soil samples contain carbonates and, in particular,

[1] Careful work on this subject has been carried out by E. Merenyi—e.g., Merenyi (2000).

[2] See, for example, Chapters 3, 14, 16, 20, and 21 of Pieters and Englert (1993) and references therein.

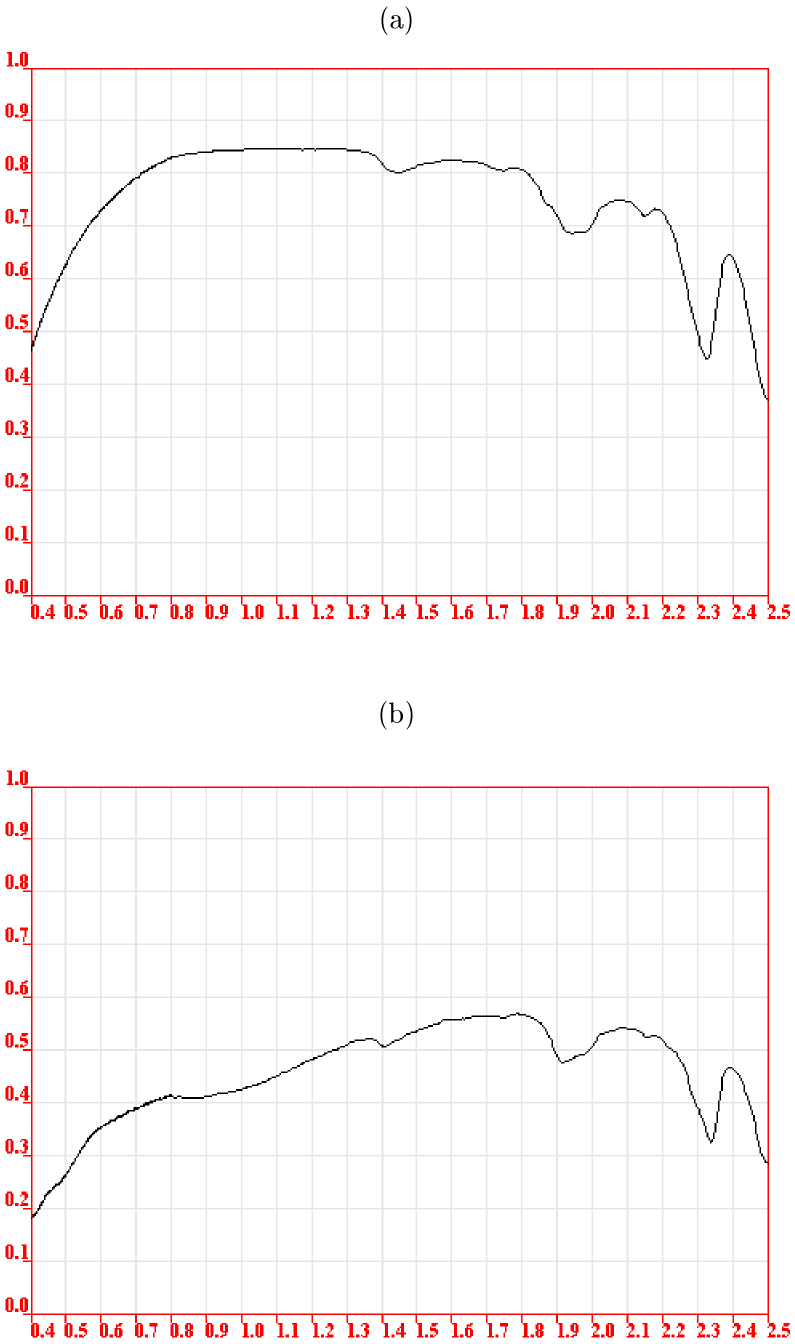
whether such samples contain either of the most frequently occurring forms of carbonate materials—calcites and dolomites. Carbonate identification is particularly interesting for extraterrestrial exploration, as carbonates are typically formed by processes—such as deposition from water—which could indicate the presence of a life-supporting environment at some point in the past. Focusing, therefore, on the task of carbonate identification and reflecting an interest in comparing human expertise to machine algorithms for identifying mineral composition from reflectance spectra, we compare the reliabilities of an expert human spectroscopist, an expert system that models human expert procedures, and a variety of automated techniques, including linear regression, each with various resampling and cross-validation techniques, on the task of mineral (mostly carbonate) identification from visual to near infrared reflectance spectra.

In our tests, an adaptation of the PC algorithm (Spirtes, Glymour, and Scheines 1993; Spirtes, Glymour, and Scheines 2000)) implemented in the TETRAD II program (Scheines, Spirtes, and Meek 1994) for constructing causal Bayes nets from data, combined with appropriate data selection and data preprocessing, performed more reliably than any other automated procedure we tested. We will refer to this procedure as the "modified PC algorithm." In some tests this procedure was more informative than a human expert spectroscopist and almost as reliable, and in other tests the procedure performed almost as well as human experts with access to both physical samples and measured spectra of physical samples. These claims are made more precise in the Chapters 4 and 5, where we additionally compare various pre-processing and data selection procedures. In this chapter, we give an introduction to the history and tradition of expertise of the problem itself of identifying the composition of rocks from their VNIR spectra.

3.2 Statement of the Problem

The task of the experiments described in the following two chapters is to determine, based on the spectrum of a rock, what its mineral class components are likely to be. Rocks are composed of amalgamated granules of pure minerals. These pure minerals themselves, though, often appear in the world in rock-sized chunks, so it is possible to study the pure minerals outside of the rock mixtures that they form. One way to examine them is to use a spectrometer—i.e., an instrument that measures light at each of a range of chosen frequencies. A reflectance spectrometer measures light reflected from objects. If we use a reflectance spectrometer to measure the light reflected from rocks and their component minerals, we find interesting relationships. Some of these relationships are hard to see, while others are quite evident. For example, Figure 3.1(a) shows a reflectance spectrum in the range $[0.4\mu m, 2.5\mu m]$ of a particular mineral (calcite), in the class of carbonates. Figure 3.1(b), on the other hand, shows a reflectance spectrum for a limestone rock that contains carbonates of this same type.

In the range [2.0 μm, 2.5 μm], these two spectra are remarkably similar. Based on this similarity, one might guess, without knowing so in advance, that the limestone contains some kind of carbonate—in particular, some kind of calcite. One would be right. This is exactly the kind of analysis which needs to be given, either automatically or by a human expert, in order to solve the problem of mineral class identification from spectra. Presented with a spectrum of a given rock, such as limestone, one needs to



(a)

(b)

Figure 3.1: Two examples of carbonate spectra. The first (a) is a calcite (a mineral); the second (b) is a limestone (a rock).

be able to say what mineral classes the spectrum reveals. In this case, the spectrum of the rock (i.e., the limestone) shows evidence for the existence of some kind of carbonate (perhaps the calcite shown, but perhaps some other carbonate).[3]

The spectra graphed above are in the visual to near infrared range; they include the visual range ([$0.4\mu m$, $0.7\mu m$]) plus most of the near infrared range ([$0.7\mu m$, $3.0\mu m$]); these spectra actually extend only to 2.5 μm. The range [0.4 μm, 2.5 μm] includes considerable information for many types of rocks and avoids a region of strong atmospheric noise beginning at 2.5 μm. It is thus a good range in which to test various procedures on the task of recovering information about mineral classes from spectra.

3.3 Expert Understanding of Mineral Identification from Spectra

It's interesting to note that the tradition of expertise in mineral reflectance spectroscopy extends back well over half a century and the tradition of expertise for spectroscopy generally extends even further back than a century. Work on spectroscopy in general began in the late nineteenth century, with early work concentrating on transmission spectroscopy, i.e., measuring how light is transmitted through substances rather than reflected from their surfaces. Work on reflectance spectroscopy proper began in the 1920's, and the practice of applying reflectance spectroscopy to minerals, using infrared

[3] These examples are chosen to illustrate how spectra in the same category can have similar features. Of course, spectra in the same category can also have quite different features too, and spectra in different categories are sometimes comparable in appearance. This is one of the reasons this inference problem is so difficult, not just for machines, but for human experts as well.

spectra, began in the 1940's.[4] (One might consider Anderson (1950) to be an early example of spectral identification of mineralogical substances, even though the objects he was studying—varieties of glasses—aren't, strictly speaking, minerals.) Since the 1950's, expert knowledge about the identification of rock and mineral composition from infrared reflectance spectra has developed more or less continuously, with major contributions being made in the 1970's. So from the point of view of human expertise, there is a well-grounded tradition that attempts to solve instances of the problem of mineral class identification from spectra using expert judgment.

Experts distinguish the background of a spectrum from its features, give explanations for the existence of particular features of rock or mineral spectra in terms of the quantum mechanical features of their chemistry and crystallograpy, and provide explanations for how the spectra of mineral mixtures (rocks) relate to the spectra of those minerals separately. In broad outline, this story has been established for some time, although in some details (especially in the details of how mineral spectra mix) experts still disagree about the details.

Experts tell us that a reflectance spectrum generally consist of a background upon which are imposed various features. Figure 3.1(a) provides an illustration. The background is a curve which changes relatively slowly; features look like dips or spikes below this curve.[5] Features can be wide or narrow, deep or shallow, skewed or symmetrical. They can be well-separated from one another or superimposed. Although one can get some information about a mineral from the shape of its background curve—one can

[4] See Wendlandt and Hecht (1966).

[5] In our analysis, we use a hull differencing algorithm to calculate this slowly changing background. See Section 5.3.4 for details.

tell whether it is dark or bright, for instance—the positions and shapes of its features provide valuable clues as to the chemistry of the rock or mineral. If the chemistry of the rock or mineral includes certain configurations of iron, for instance, one can generally get a hint that this is the case by looking for a feature with a certain shape at about 0.9 μm to 1.0 μm.

The features and background shape that one observes for a rock or mineral spectrum are, according to established expert knowledge, the result of three general sorts of processes: reflection, transmission, and emission.[6] Reflection occurs when light meets the surface of an object and bounces back; transmission occurs when light continues through the surface. Emission occurs when an object spontaneously generates black-body EM radiation. In the range $[0.4\mu m, 2.5\mu m]$, emission is not generally a major concern; its effects are much more noticeable in the mid-infrared range ($[3.0\mu m, 30.0 mum]$). Reflection and transmission in the visual to near infrared range are, however, in the expert's view, diagnostic in this range, since they yield information which helps to distinguish classes of minerals and even particular minerals from one another. Consider reflection. Even perfect reflection from surfaces of minerals is not uniform, since the angles and wavelengths at which reflection occurs are distinctive features of particular minerals. Reflection by itself, therefore, can yield complicated, diagnostic features in mineral spectra. The effects generated by transmission are however even more complex (and therefore more diagnostic); these account for most of the features observed in the reflectance spectra of rocks or minerals. When light is transmitted through a mineral or scattered through irregularities or particles on the surface of the mineral, the pat-

[6] The discussion here follows Clark (1999).

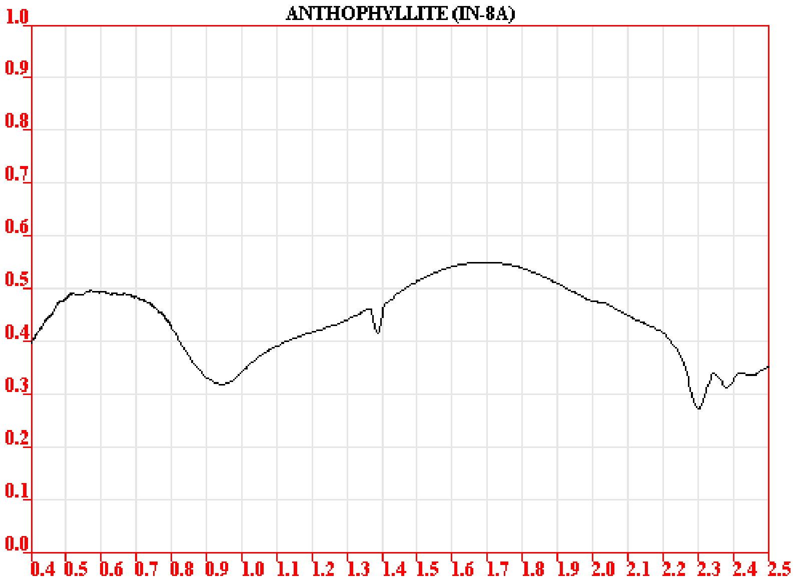


Figure 3.2: An inosilicate spectrum, illustrating a Fe^{2+} feature at about $0.9\mu m$.

tern of absorption can be affected by crystal field effects, charge transfer absorptions, conduction bands, color centers, and vibrational modes.

Crystal field effects are due to unfilled electron shells of transition elements (nickel, cobalt, iron, etc.). When these atoms are isolated, their two *d*-orbitals have identical energies, but when they are placed into a crystal field, these orbital energies split. In this split configuration, electrons can be moved from a lower energy to a higher energy if they absorb photons whose energy accounts for the difference. The exact amount of energy needed depends on the type of atom, its electronic configuration, and a variety of other facts. Figure 3.2 shows an example of a crystal field effect for Fe^{2+} at about $0.9\ \mu m$.[7]

Charge transfer absorptions are due to electrons being transfered from one

[7] This wavelength, $1.0\ \mu m$, lies within the $[0.4\ \mu m,\ 2.5\ \mu m]$ range of the data to be examined in the next chapter.

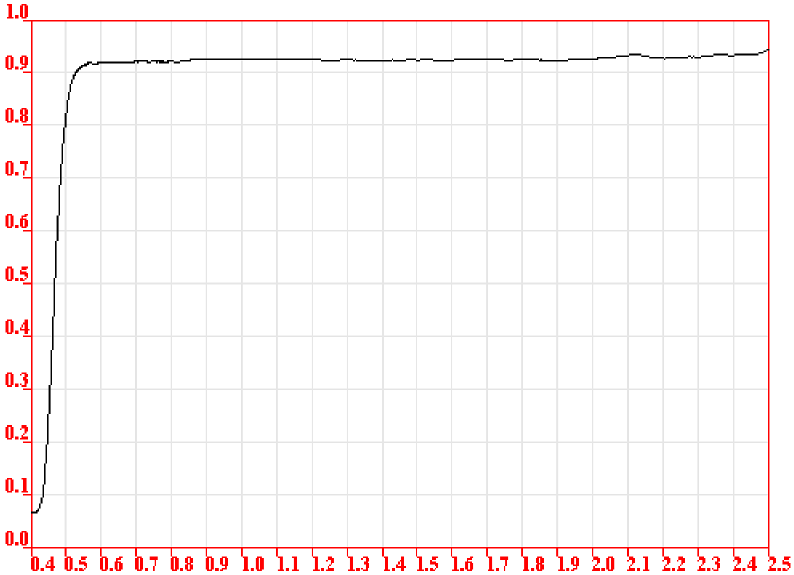


Figure 3.3: A spectrum of sulphur, an elemental mineral, showing the “step function” shape.

element to another—e.g., from one ion to another—or from one valence state to another for the same metal. Here again, the energy needed to make this electronic transfer needs to be supplied from somewhere, and photons with exactly the right energy are such a source. These effects are typically centered in the ultraviolet range, with some influence in the visual range. These kinds of absorptions explain, e.g., why iron oxides are red.

Conduction bands are heightened energy levels in certain minerals in which electrons may move freely about the mineral without being tied to particular atoms. In some minerals, such as sulfur, the band gap between the valence band (the lower energy band in which electrons are tied to particular atoms) and the conduction band (the higher energy band where electrons can move about freely) corresponds to the energy of a photon whose wavelength falls in the visible to near-infrared range. In minerals like this the reflectance spectrum looks more or less like a step function (cf. Figure 3.3).

Color centers sometimes result when a crystal that has imperfections in it is irradiated. Defects in the crystal produce discrete jumps in energy levels within the crystal, and electrons can be drawn into these discrete levels if they absorb photons with the right energy. This effect is responsible for variations in color of certain minerals due to impurities—e.g., the yellow, purple, and blue versions of flourite.

Vibrational modes are the various ways in which atoms and molecules vibrate. Typically, each type of molecule can vibrate in a number of a number of different ways at once, depending on how many atoms and bonds there are in the molecule. If a molecule has N atoms, there are $3N - 6$ basic modes of vibration called *fundamentals*. Just as with a cello string, in addition to the fundamental vibrations, there are also *overtones* (i.e., multiples of fundamental vibrations) and *combinations* (when different modes of vibration interact). A mineral will absorb photons at a particular wavelength if those photons provide its components the right amount of energy to vibrate in particular ways. Carbonates provide good examples. Carbonates are minerals that contain the group CO_3; this group has certain identifiable vibrational characteristics that show up in the visual to near-infrared range as overtones and combinations. It has a symmetric stretch (ν_1)at $9.407\mu m$, an out-of-plane bend (ν_2) at $11.4\mu m$, an asymmetric stretch (ν_3) at $7.067\mu m$, and an in-plane bend (ν_4) at $14.7\mu m$. All of these are outside of the range $[0.4\mu m, 2.5\mu m]$, but two combinations and overtones are prominent within this range—viz., (a) $\nu_1 + 2\nu_3$ at $2.50\mu m$ to $2.55\mu m$ and (b) $3\nu_3$ at $2.30\mu m$ to $2.35\mu m$.[8] In many carbonate reflectance spectra in the visual to near infrared range, these lines are clearly visible.

[8] This illustrates how features can shift wavelengths.

Each of these processes—crystal field transfers, charge transfers, conduction band transfers, color center transfers, and vibrational mode absorptions—are processes of absorption, not reflection, but they have vivid effects in reflectance spectra, largely due to the process of scattering, in which light is bounced around from one mineral granule to another before finally being sent back out as a reflection from the surface of a rock or mineral. The ideal mineral would have a completely flat surface with no irregularities, but real mineral surfaces are not completely flat. Also, minerals may be ground up into powders, so that their surfaces are uneven for a different reason. Rocks, which are made up of mineral grains, also have uneven or grainy surfaces. Photons encountering an uneven surface of any of these types will follow a path that is essentially a complex random walk[9] before emerging again from the surface in the form of a reflection. Thus the reflections observed in reflectance spectra are in actual fact a record of a combination of absorption and reflection processes.

For rocks, the situation is especially interesting, as while as photons do random walks about the surface of a rock, they pass through (and are affected by) mineral granules of different types. Mineralogists believe that this scattering process for rocks is what largely accounts for the way in which the absorption effects for different minerals combine to generate reflectance spectra for rocks that are combinations of those minerals. "Scattering," as Clark puts it, "is the process that makes reflectance spectroscopy possible":

> In transmission, light passes through a slab of material. There is little or no scattering (none in the ideal case; but there are always internal reflections from the surfaces of the medium). Analysis is relatively simple. Reflectance

[9] See Clark and Roush (1984).

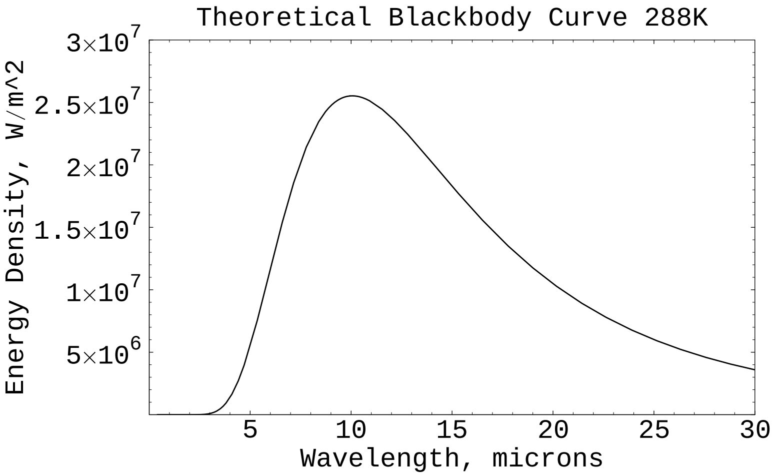


Figure 3.4: Theoretical blackbody radiation in the range $[0\mu m, 30\mu m]$ at 15 degrees Celsius (288 Kelvin).

> of a particulate surface, however, is much more complex and the optical path of photons is a random walk. At each grain the photons encounter, a certain percentage are absorbed. If the grain is bright, like a quartz grain at visible wavelengths, most photons are scattered and the random walk process can go on for hundreds of encounters. If the grains are dark, like magnetite, the majority of photons will be absorbed at each encounter and essentially all photons will be absorbed in only a few encounters....
>
> The random walk process of photons scattering in a particular surface also enhances weak features not normally seen in transmittance, further increasing reflectance spectroscopy as a diagnostic tool. Consider two absorption bands of different strengths, such as a fundamental and an overtone. The stonger absorption will penetrate less deeply into the surface, encountering fewer grains because the photons are absorbed. At the wavelengths of the weaker absorption, few photons are absorbed with each encounter with a grain, so the random walk process goes further, increasing the average photon path length. The greater path length will result in more absorption, thus strengthening the weak absorption in a reflectance spectrum.[10]

Finally, even though emission is not a process of great importance in the visual to near-infrared range, it may be helpful to say something about it. Every object emits blackbody radiation that depends on its temperature. Figure 3.4 shows how this blackbody radiation is related to wavelength at 288 K (15 C); the curve peaks at about $10\mu m$,

[10] See Rencz (1999), p. 35

which is well into the mid-infrared range, and drops essentially to zero at about $3\mu m$.[11] In the range $[0.4\mu m, 2.5\mu m]$ at this temperature (and even at much higher temperatures) blackbody radiation plays no significant role.

3.4 Calculation of Intensity

It is interesting to note how reflection spectra are actually measured. The device used to do the measurement is called a *reflectance spectrometer*. This is a device which shines white light onto a sample, then measures the amount of light reflected at each of a pre-selected set of wavelengths. Of course there is no device that emits perfectly white light, i.e., light that has exactly the same intensity at all desired wavelengths. To compensate for this imperfection, one typical method used is to measure the spectrum first of a *white sample*—a piece of material that reflects as perfectly as possible over the range of wavelengths being tested. Afterwards, the spectrum of the target object is measured. That way, the reflectance value of the target object can be divided through by the reflectance value of the white sample for each wavelength of interest. The quotient that results is called the *intensity value*. All of the data considered in the upcoming chapters is either already provided as a set of intensity values or else has been converted to intensity values for use in the experiments.

[11] The formula for theoretical blackbody radiation as a function of wavelength and temperature is $E(\lambda, T) = \frac{2\pi hc^2}{\lambda^5(e^{hc/\lambda kT}-1)}$, where λ is the wavelength (in meters), T is temperature (in Kelvin), h is Planck's constant, c is the speed of light, k is Boltzmann's constant, and $E(\lambda, T)$ is energy density per unit time per unit wavelength.

3.5 Preview

In the next two chapters, experiments will be reviewed which compare the performance of an expert to the performances of several machine algorithms on the task of identifying mineral class composition from spectra. Chapter 4 will review an experimental assessment of expert skill level. Chapter 5 will review experiments assessing the performance of several machine learning algorithms; the performance of these algorithms will also be compared to expert skill level. Chapter 6 will respond to arguments from Chapter 2 in light of this experimental evidence.

Chapter 4

Experiment: Expert Assessment of Mineral Class Composition

4.1 Introduction

We begin a review of experimental work by examining an experiment performed in 1998 to assess the skill level of a human expert spectroscopist at judging the mineral composition of rocks, based on graphs of their visual to near infrared reflectance spectra. The motivation for the experiment was to establish a standard for the design of automated systems intended to supplement or replace human experts in the analysis of rock spectra at remote locations. At the time of the experiment, there was no relevant literature to help establish such a standard, and even at present the only experimental evidence available is the evidence contained in this chapter. We therefore rely on this evidence in upcoming chapters to compare directly human performance and machine performance on this task.

In this experiment, unlabeled spectral graphs of rocks are presented in random order to an expert, who classifies them according to his own judgment of their mineral class composition. The options for mineral classes are set by the experiment. A gold standard is established separately using sample petrology, against which responses are compared.

4.2 Data Sources

Two sources of spectral data are used in the experiment: (a) the JPL Spectral Library and (2) the JHU Spectral Library. The spectra themselves come from the JHU library; they are formatted, however, using the parameters of the JPL library. This formatting allows results from the expert experiment to be compared directly to results from the machine algorithms experiments of the next chapter. Brief descriptions of each library follow; more elaborate descriptions may be found in Appendices A and B.

4.2.1 The JPL Spectral Library (Minerals)

The JPL Spectral Library consists of spectra for 160 different minerals, with wavelengths covering the range $[0.4\mu m, 2.5\mu m]$ (826 wavelengths altogether). The mineral samples for this library were ground into powders of three different grain sizes: small ($< 45\mu m$), medium ($45 - 125\mu m$), and large ($125 - 500\mu m$). Each mineral was measured at one, two, or three of these grain sizes, and out of the 160 minerals, 135 of them were measured at the largest grain size. These 135 large grain mineral spectra will be referred to as the "JPL Large Grain Mineral Library."

The minerals in the JPL Spectral Library are divided into 17 mineral classes, as shown in Table A.1.[1] These mineral classes are a subset of the mineral classes presented in the standard reference, *Dana's Mineralogy*.[2] (Alternative classifications of minerals exist in the literature, but for the most part the 17 JPL classes can be mapped, more or less unproblematically, onto these alternative classifications.)

4.2.2 The JHU Spectral Library (Rocks)

The JHU Spectral library consists of several hundred spectra with various wavelength ranges, 192 of which are spectra for rocks in the wavelength range of about $[0.4\mu m, 14.0\mu m]$.[3] These 192 spectra represent 96 different rocks, each of which is presented in two different forms. The library is also divided into rocks of three very general sorts—igneous, metamorphic, and sedimentary. Igneous rocks are presented in solid form as well as in fine-grained powdered $(0 - 75\mu m)$ form. Metamorphic and sedimentary rocks are presented in fine-grained and coarse-grained powdered $(500 - 1500\mu m)$ form. The number of rocks in each of the six resulting categories is shown in Table B.1.

[1] Minerals naturally belong to only one mineral class, while rocks may contain a number of minerals, and hence belong (in a compositional sense) to more than one mineral class.

[2] See Gaines, Skinner, Foord, Mason, and Rosenzweig (1997).

[3] The exact set of wavelengths used by rock spectra in the JPL Library tends to be somewhat variable in the upper reaches of the $[0.4\mu m, 14.0\mu m]$ range, but through interpolation there is no difficulty producing spectra from these in the range $[0.4\mu m, 2.5\mu m]$, i.e., the same wavelengths as the JPL mineral spectra.

4.3 Data Preparation

4.3.1 Preparation of the JHU Large Grain Rock Library

The 192 JHU rock spectra were interpolated to the set of wavelengths used in the JPL Mineral Library using the simple linear interpolation algorithm defined in Section 5.3.4. This results in a set of spectra which cover the range $[0.4\mu m, 2.5\mu m]$, so that the subject in this experiment is present with spectra in the same range as the machine algorithms reviewed in Chapter 5. This set of 192 spectra interpolated to the JPL wavelengths will be referred to as the "JHU Rock Library."

4.3.2 Preparation of JHU Gold Standard

In order to determine the skill level of an expert attempting to classify the 192 spectra of the JHU Rock Library, an objective method for determining the true classification of the JHU spectra is required—viz., an objective estimate of the 192 JHU spectra in terms of the mineral classes used in the experiment (the 17 JPL mineral classes as shown in Table A.1).

The procedure we used was as follows. In the file descriptors of each data file an expert petrology is provided, containing information which can be used to perform such a classification. Several weeks before acting as subject in the expert experiment, Ted Roush, Senior Geologist at NASA Ames, examined each of these file descriptors in turn (96 in total) and classified each rock in the library using the 17 JPL categories. For each rock and for each category, Roush specified whether the category was present, absent, or possibly present in the composition of that rock. In our data analysis, we considered two interpretations of “possibly present”: (a) interpreting “possibly present” as “present,” for the sake of creating a gold standard and (b) interpreting “possibly present” as “absent” for the sake of creating a gold standard. The inclusive gold standard resulting from interpreation (a) will be referred to as G_1 and the exclusive gold standard resulting from (b) will be referred to as G_2.

As an example of how this procedure works for establishing a gold standard, consider the petrology for the sample in the file named “andesi1f”:

> The sample is about 4 x 3 cm, brown on the weathered surface and dark gray on fresh surfaces. It is porphyritic with the phenocrysts making up about 25-30% of the rock. The groundmass is gray and microcrystalline. The phenocrysts are approx. $< 1mm$ and consist of plagioclase laths, pyroxene and opaques, in that order of abundance, with pyroxenes nearly as abundant as the feldspars.

Interpreting the petrology, Roush determined that the sample contains components belonging to the following three JPL classes:

Inosilicates, Oxides, Tectosilicates.

Roush found no evidence in the petrology that it belonged to any of the other 17 classes. The gold standard for the sample “andesi1f” therefore consists of the assertion that it contains minerals of these three classes and none of the others. In this case, G_1 and G_2 agree.[4]

[4] The gold standard classifications for the Carbonate class are shown in Table B.2.

4.4 Experiment

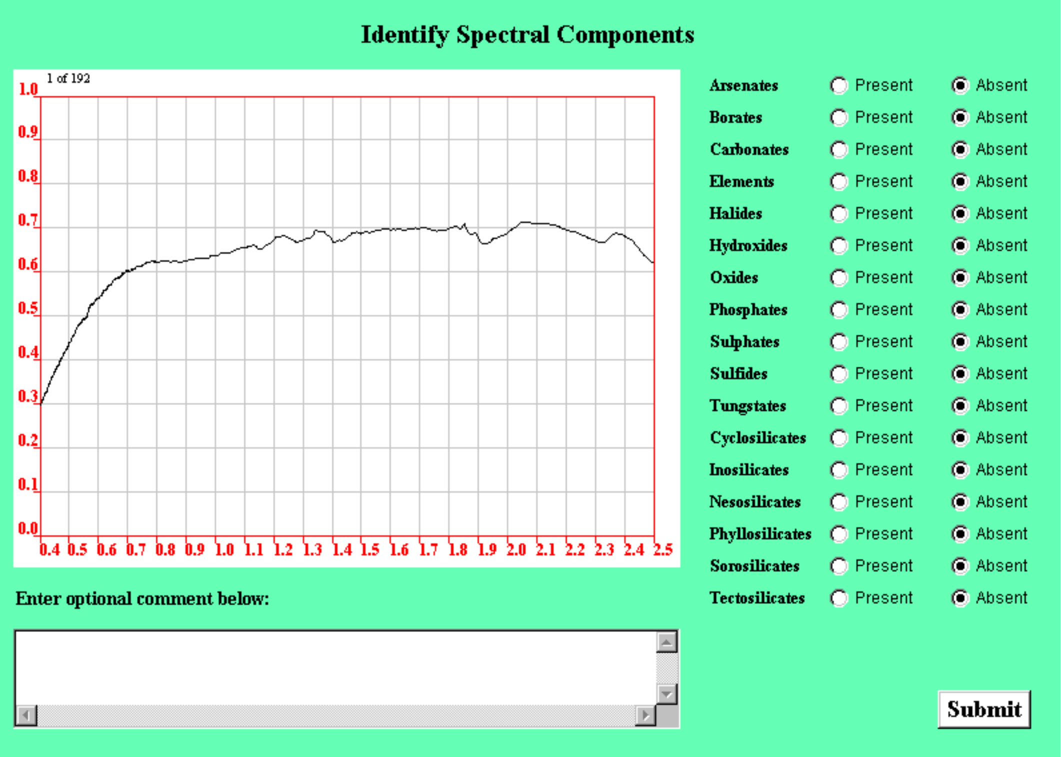
4.4.1 Methods

For the experiment itself, unlabeled spectra from the JHU Rock Library were presented in random order to a subject (an expert in the field of geological spectroscopy), who predicted, for each spectrum, which types of minerals were present in the rock measured by that spectrum. The permissible mineral classes to be used for this purpose were the 17 JPL mineral classes.

By way of implementation, a Java applet was designed with a simple servlet backend, that allowed the subject to work at the experiment at his own pace. The applet presented an interface to the subject which permitted him to carefully examine each spectrum in turn, predict the mineral classes of its components, and submit his answer. The servlet backend collected the data submitted and appended it appropriately to a data file. The subject was permitted as much time as he liked to work on the experiment, was allowed to stop and start as often as he liked, was allowed to work at different locations if he chose, and was allowed to consult any reference works he deemed helpful. When all the data was collected from the experiment, it was compared to the gold standard described above; the results are shown in tables at the end of this chapter. Altogether, the subject spent approximately 12 hours on the task over a period of seven days.[5]

Figure 4.1 shows an image of the applet from the subject's point of view. The upper left hand portion of the applet exhibited an unlabeled spectral graph. (A counter

[5] The subject began on Friday, 12/11/1998, and finished on Friday, 12/18/1998.



Identify Spectral Components

Arsenates	○ Present	● Absent
Borates	○ Present	● Absent
Carbonates	○ Present	● Absent
Elements	○ Present	● Absent
Halides	○ Present	● Absent
Hydroxides	○ Present	● Absent
Oxides	○ Present	● Absent
Phosphates	○ Present	● Absent
Sulphates	○ Present	● Absent
Sulfides	○ Present	● Absent
Tungstates	○ Present	● Absent
Cyclosilicates	○ Present	● Absent
Inosilicates	○ Present	● Absent
Nesosilicates	○ Present	● Absent
Phyllosilicates	○ Present	● Absent
Sorosilicates	○ Present	● Absent
Tectosilicates	○ Present	● Absent

Enter optional comment below:

Figure 4.1: A screen shot of the applet used to collect expert mineral class identification data.

was included in the image so that the subject would know how far he'd progressed.) On the right were Present/Absent radio buttons for each of the 17 JPL mineral classes. On the bottom of the screen was a text field in which the subject could make whatever notes he wished. When the subject had inspected a particular graph to his satisfaction, made his classification, and recorded any desired notes, he pressed the SUBMIT button, after which all of the checkboxes were reset, the area for notes cleared, and a new unlabeled spectral graph displayed.

Each time the SUBMIT button was pressed, a record was sent to the backend servlet for the applet containing the time of submission, the name of the sample, the classification, and the notes from the textfield. These records were then appended to a file. In order to ensure that all of the records were submitted by the same person during the course of the experiment, the subject was required to sign in with a username and password each time he returned to the applet.

Using this mechanism, the 192 spectral graphs of rocks from the JHU Rock Library were presented in random order and without any labeling to the subject, who classified each of them but one ("qrtzit6f") in the prescribed manner. 191 records were therefore collected. The final collected set of data was then compared to the gold standards G_1 and G_2.

4.4.2 Results

For each of the 17 JPL mineral classes C, and for each of the two interpretations of the gold standards G_1 and G_2, two quantities were calculated:

1. The frequency with which the subject estimated rocks to belong to C given that

they belong to C according to the gold standard ("accuracy"), and

2. The frequency with which rocks belong to C according to the gold standard given that the subject estimated them to belong to C ("coverage").

Accuracy results for G_1 are shown in Table 4.1; for G_2 accuracy results are shown in Table 4.3. Coverage results for G_1 are shown in Table 4.2; for G_2 are shown in Table 4.4. Setting aside JPL mineral classes which are poorly represented among the JHU rock spectra (according to each gold standard, respectively), what the data show is that the subject is a particularly accurate classifier for certain mineral classes—viz., tectosilicates, carbonates, nesosilicates, and phyllosilicates—with slightly higher accuracies reported in connection with the inclusive gold standard. However, when one looks at the coverage for these same mineral categories, one finds that, except for phyllosilicates and inosilicates, the results fall off dramatically. Consider the carbonate class: When the subject classifies a spectrum as representing a rock with a carbonate component, it is fairly certain that the rock actually does have a carbonate component, according to the gold standard. However, out of all of the rocks that have carbonate components according to the gold standard, only some of them are correctly identified by the subject as having carbonate components. The subject is therefore accurate but conservative.

Table 4.1: Probability that the subject of the expert experiment identifies a rock as belonging to a mineral category given that the inclusive gold standard G_1 identifies it as belonging to that category.

p := # cases for which expert classification from spectra = present or uncertain.
q := # cases for which expert classification from petrology = present.

Category	$(p\&q)/q$	$p\&q$	q
arsenates	*	0	0
borates	*	0	0
carbonates	0.96	24	25
cyclosilicates	0.00	0	1
elements	0.28	9	32
halides	0.00	0	4
hydroxides	0.00	0	20
inosilicates	0.59	52	88
nesosilicates	0.77	10	13
oxides	0.61	37	61
phosphates	*	0	0
phyllosilicates	0.76	83	109
sorosilicates	*	0	0
sulfides	0.24	8	33
sulphates	0.00	0	1
tectosilicates	1.00	29	29
tungstates	*	0	0

Table 4.2: Probability that inclusive gold standard G_1 identifies a rock as belonging to a category given that the subject of the expert experiment identifies it as belonging to that category.

p := # cases for which expert classification from spectra = present.
q := # cases for which expert classification from petrology = present or uncertain.

Category	$(p\&q)/q$	$p\&q$	q
arsenates	*	0	0
borates	0.00	0	2
carbonates	0.26	24	92
cyclosilicates	0.00	0	14
elements	0.41	9	22
halides	0.00	0	2
hydroxides	0.00	0	4
inosilicates	0.62	52	84
nesosilicates	0.19	10	54
oxides	0.37	37	100
phosphates	0.00	0	22
phyllosilicates	0.68	83	122
sorosilicates	0.00	0	10
sulfides	0.33	8	24
sulphates	0.00	0	2
tectosilicates	0.17	29	166
tungstates	0.00	0	2

Table 4.3: Probability that the subject of the expert experiment identifies a rock as belonging to a mineral category given that the exclusive gold standard G_2 identifies it as belonging to that category.

p := # cases for which expert classification from petrology = present
q := # cases for which Expert classification from spectra = present.

Category	$(p\&q)/q$	$p\&q$	q
arsenates	*	0	0
borates	*	0	0
carbonates	0.96	24	25
cyclosilicates	0.00	0	1
elements	0.12	4	32
halides	0.00	0	4
hydroxides	0.00	0	20
inosilicates	0.59	52	88
nesosilicates	0.77	10	13
oxides	0.49	30	61
phosphates	*	0	0
phyllosilicates	0.72	79	109
sorosilicates	*	0	0
sulfides	0.12	4	33
sulphates	0.00	0	1
tectosilicates	1.00	29	29
tungstates	*	0	0

Table 4.4: Probability that the exclusive gold standard G_2 identifies a rock as belonging to a category given that the subject of the expert experiment identifies it as belonging to that category.

p := # cases for which expert classification from spectra = present
q := # cases for which expert classification from petrology = present

Category	$(p\&q)/q$	$p\&q$	q
arsenates	*	0	0
borates	0.00	0	0
carbonates	0.39	24	62
cyclosilicates	0.00	0	10
elements	0.50	4	8
halides	*	0	0
hydroxides	0.00	0	2
inosilicates	0.63	52	82
nesosilicates	0.19	10	52
oxides	0.38	30	78
phosphates	0.00	0	20
phyllosilicates	0.69	73	114
sorosilicates	0.00	0	8
sulfides	0.22	4	18
sulphates	*	0	0
tectosilicates	0.18	29	162
tungstates	*	0	0

Chapter 5

Algorithmic Mineral Class Identification Experiments

5.1 Introduction

In this chapter, a number of algorithms are compared to one another and to the human expert experiment of Chapter 4 on the task of judging mineral class composition from visual to near infrared spectra of rocks. The algorithms fall into two categories: those which judge mineral class composition by comparing the spectra of a rock to a background library of pure mineral spectra and those which do not. For algorithms of the first sort, we used the large grain spectra from the Jet Propulsion Laboratory (JPL) Spectral Library as our background library of pure minerals. For all of the experiments in this chapter, we used target spectra from two different sets: (a) the rock spectra from the Johns Hopkins University (JHU) Spectral Library; and (b) a set of field data collected at Silver Lake near Baker, California. The algorithms tested include simultaneous linear regression, stepwise linear regression, the modified PC algorithm (a Bayes net model search), logistic regression, classification and regression trees (CART), naive Bayes classifiers, probabilistic decision trees, and several varieties of neural nets.

This chapter will explain how the data used in these experiments were produced, how the algorithms compared in the experiments were designed, how the experiments themselves were carried out, and what the results of the experiments were. Where appropriate, comparisons will be made to the human expert data presented in the previous chapter.

The conclusion drawn from all of these experiments taken together is that the modified PC algorithm performed at least as well as any of the other algorithms compared and that it performed comparably to a human expert at the same task. This is interesting because the modified PC algorithm was designed using principles of automatic causal discovery described in Spirtes *et al.* (1993). This invites a causal interpretation as to why the algorithm is so successful. Such an interpretation will be given in Chapter 6.

5.2 Data Sources

Three sources of data are used: (a) the Jet Propulsion Laboratories Spectral Library, (b) the Johns Hopkins University Spectral Library, and (c) a set of field samples measured near Silver Lake, California. Descriptions for (a) and (b) are given in Chapter 4; a technical description for the Silver Lake field samples follows.

5.2.1 The Silver Lake Field Samples

In the Winter of 1999, NASA scientists conducted field tests of a robot and various instruments in and about Silver Lake, California, a dry lake bed in the Mojave desert (Stoker et al. 1999). Spectral data for field rocks was collected by a spectrometer mounted on a field robot, controlled remotely. The instruments included a near-infrared spectrometer (Johnson et al. 1999), described below. Spectra were taken, usually *in situ*, of rocks and soils; the spectra were identified as carbonates or non-carbonates both by the field geologists, from physical observations of the specimens and their spectra, and by a group of geologists located remotely at NASA Ames, who used both the spectra and the descriptions of the field experts (Gazis and Roush 1999). Paul Gazis at NASA Ames provided software to correct instrumental artifacts and to filter out spectra that, typically because of atmospheric effects, were too noisy to process. After preprocessing, 21 spectra remained; 13 samples were identified as carbonates and 8 samples identified as non-carbonates by the field geologists. Subsequently, eight of the 21 samples were analyzed by standard petrographic techniques. All eight analyses agreed with the judgments of the field geologists and the remote geologists. The expert judgments are shown in Table C.1.

Prior to receipt of the set of 21 field spectra measured at Silver Lake, we received a single field sample, measured by Roush and Glymour on an earlier excursion at the same site, which was used for a separate study (see Section 5.5.2). Details of the instrumentation, data collection, and artifacts are given in Appendix C.

Table 5.1: Schema for the tab-delimited file format used for all algorithmic experiments. Included are variables for wavelength and (prepared) spectral intensities for individual spectra at those wavelengths.

$[wave]$	$file_1$	$file_2$	$file_3$	...	$file_n$
[0.400]	$i_1(0.400)$	$i_2(0.400)$	$i_3(0.400)$	...	$i_n(0.400)$
[0.401]	$i_1(0.401)$	$i_2(0.401)$	$i_3(0.401)$	...	$i_n(0.401)$
[0.402]	$i_1(0.402)$	$i_2(0.402)$	$i_3(0.402)$	...	$i_n(0.402)$
...	...	...	...	...	...
[2.500]	$i_1(2.500)$	$i_2(2.500)$	$i_3(2.500)$	...	$i_n(2.500)$

5.3 Data Preparation

Data from all sources were prepared as tab-delimited text files in the format of Table 5.1. The set of wavelengths used in these files was typically a subset of the set of wavelengths used in the JPL Spectral Library. (Usually, all of the wavelengths of the JPL Spectral Library were used, though sometimes intervals of these wavelengths were removed due to atmospheric noise.) For data sets other than the JPL Spectral Library, spectra were interpolated to this set of wavelengths using a simple linear interpolator. After interpolation, two other preparations were often used. Hull differences of data were calculated, and wavelength intervals of spectral data were removed. Details of preparation for separate libraries follow.

5.3.1 JPL Large Grain Mineral Library

File headers aside, data from the JPL Spectral Library is organized into one, two, or three columns of percent spectral intensities, with wavelengths in a separate file (‘beck.dat’). (No interpolation of this library was necessary.) The wave function for the JPL mineral library includes the wavelengths in the interval $[0.400\mu m, 0.800\mu m]$ with a regular spacing of $0.001\mu m$ and the wavelengths in the interval $[0.800\mu m, 2.500\mu m]$ with a regular spacing of $0.004\mu m$, for a total of 826 channels. To produce a tab delimited file of the format in Table 5.1, individual data files were examined to determine whether they contain a large grain spectrum. All large grain spectra from these files were then combined into a single data file in the format of Table 5.1. Spectra were optionally hull-differenced, and in cases where wavelength intervals needed to be removed from the target, the same wavelength intervals were removed from the JPL Large Grain Mineral Library file. The result of the procedure was a library ocontaining 135 spectra, sometimes hull differenced, sometimes with water lines removed.[1]

The 17 mineral classes used throughout all of the experiments described in this chapter are the mineral categories from the JPL Large Grain Spectral Library, given in Table A.1.

5.3.2 JHU Rock Library

File headers aside, data in the JHU Spectral Library are formatted as one file per spectrum, with two columns in each file—one for wavelengths and the other for spectral intensities. The range of the data is $[0.4\mu m, 14.0\mu m]$. At the lower ranges, the wavelengths of the JHU rock spectra are evenly spaced at regular intervals of $0.001\mu m$; at higher wavelengths, the exact wavelengths measured are not as evenly spaced and show slight discrepancies from one file to the next. To produce tab delimited files in the

[1] The range of lines removed due to water lines was $[1.198\mu m, 2.004\mu m]$.

Table 5.2: Number of samples in the JHU Rock Library judged by a human expert to contain minerals of each given JPL mineral class.

Mineral Class	**Count**
Arsenates	0
Borates	2
Carbonates	92
Cyclosilicates	14
Elements	21
Halides	2
Hydroxides	4
Inosilicates	84
Nesosilicates	54
Oxides	100
Phosphates	22
Phyllosilicates	121
Sorosilicates	10
Sulfides	24
Sulphates	2
Tectosilicates	165
Tungstates	2

format of Table 5.1, the spectral data in each file was interpolated to the JPL Spectral Library wave function, optionally hull differenced, and assembled into six separate file, one each for the subcategories of JHU spectra shown in Table B.1. For some experiments, these six files are combined into one file.

In order to determine whether an algorithm has correctly classified a particular spectrum from the JHU Rock Library using the mineral categories of the JPL Mineral Library, a gold standard classification of the JHU rocks using the JPL mineral categories must be constructed. This gold standard, supplied to us by Ted Roush of NASA Ames, is described in Chapter 4. Table 5.2 shows the number of samples which were judged by this gold standard to belong to each JPL class.

5.3.3 Silver Lake Field Data

Data from the Silver Lake Field Study were formatted as 21 separate files which (headers aside) contained two columns each—wavelength and spectral intensity. The wave function for these samples covers the range $[0.350\mu m, 2.500\mu m]$, with wavelengths evenly spaced at intervals of $0.001\mu m$. To produce tab-delimited files of the form shown in Table 5.1, the spectral intensity data in these files was interpolated and arranged in that format.

Preprocessing of spectra was as follows. First, the spectra were interpolated to the JPL wave function, using the algorithm described in the following subsection. Second, since the spectra were measured under field conditions, noise caused by water absorption lines in the atmosphere at about $1.9\mu m$ was pronounced. A wavelength interval around this line was removed from the spectra (and from the corresponding background library files which were used). Third, hull differences (or first differences) of the spectra were optionally calculated. (By way of comparison, the effects of different preprocessing procedures are illustrated by analyses of the spectrum of a rock taken near Silver Lake, CA; the results are shown in Table 5.4.)

5.3.4 Auxiliary Algorithms

Interpolation

It was necessary to interpolate the spectra of the target library (i.e., the 192 JHU rock spectra) so that they used the same set of wavelengths (i.e., the same wave function) as the background mineral library for algorithms which required the use of both. The wave function for the JPL mineral library included the wavelengths in the interval $[0.400\mu m, 0.800\mu m]$ with a regular spacing of $0.001\mu m$ and the wavelengths in the interval $[0.800\mu m, 2.500\mu m]$ with a regular spacing of $0.004\mu m$, for a total of 826 channels. The wave functions for the JHU rocks cover the interval $[0.400\mu m, 14.000\mu m]$, with some variation from spectrum to spectrum at the higher end of this range. At the lower ranges, the wavelengths of the JHU rock spectra are evenly spaced at regular intervals of $0.001\mu m$; at higher wavelengths, the exact wavelengths measured are not as evenly spaced and show slight discrepancies from one file to another. These discrepancies can be compensated for by a linear interpolation, which at the same time allows the wavelengths of the JHU spectra to be converted into the wavelengths for the JPL spectra.[2]

Algorithm 6 (Simple Linear Interpolation Algorithm) *Given finite sets of wavelengths W_1, W_2 (where the range of W_2 is included in the range of W_1) and a spectral measurement $m_1 : W_1 \to [0.0, 1.0]$. Construct a second, interpolated measure $m_2 : W_1 \to [0.0, 1.0]$ as follows. For each $w_0 \in W_2$, find the largest $w_1 \leq w_0$ in W_1 and the smallest $w_2 \geq w_0$ in W_1 (w_1 may equal w_2). Form the (perhaps degenerate) line segment connection $(w_1, m_1(w_1))$ and $(w_2, m_1(w_2))$. (If $w_1 = w_0 = w_2$, this will just be a point.) Find the value along this line segment at w_0. Set $m_2(w_0)$ to this value. Return m_2.*

[2] For most channels, no interpolation is necessary, since the JHU channels are included in the set of JPL channels.

Hull Differencing

A standard strategy in econometrics for transforming a time series into a better approximation of i.i.d.[3] data is to form a series of the differences of data points and their neighbors. An analogous strategy can be applied to spectra. There are algorithms that fit a curve, called a hull, around any given spectrum. Depending on a parameter in the algorithm (window size), the hull may be chosen to fit the spectrum loosely or tightly, with few inflection points or many. Given a hull, a spectral series of hull differences can be calculated by taking the difference at each frequency of the hull value and the raw spectral value.

There are two advantages to using the hull differences (with a constant parameter for all spectra) of library and field spectra as data. The spectra of many different minerals may show the same overall shape; taking hull differences discounts this correlation. The hull differences are furthermore a better approximation to i.i.d data because the autocorrelation of neighboring frequencies is reduced.

The hull differencing algorithm used for these experiments was provided for us by Paul Gazis of NASA Ames:

Algorithm 7 (Hull Difference Algorithm) *Given a set of wavelengths* $W \in [w_0, w_1]$ *(including the endpoints), a set of measurements* $m : W \rightarrow [0.0, 1.0]$*, and a maximum window size* $r \in (0, w_1 - w_0]$,

1. Set $w' = w_1$.

[3] "Identically and independently distributed." A distribution is i.i.d. if the distribution that points are drawn from are not are not statistically different from one another (except for a shift in mean)—e.g., they are all normal with statistically indistinguishable standard deviations.

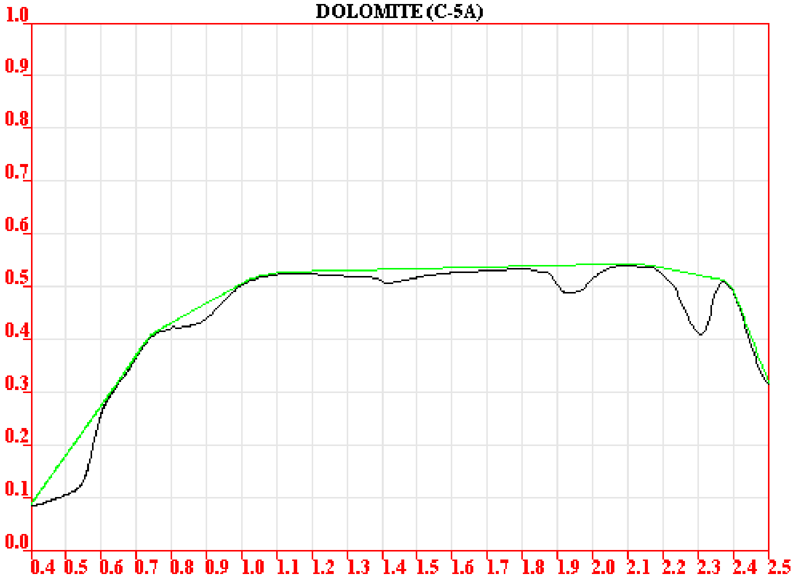


Figure 5.1: A dolomite spectrum in the range $[0.4\mu m, 2.5\mu m]$, showing the construction of background hull.

2. *Search to the left of* w' *for the first wavelength* $w'' \in W \cap [w' - r, w_1]$ *such that all points on the graph of the spectrum in the range* $[w' - r, w_1]$ *lie below the line through* $(w'', m(w''))$ *and* $(w', m(w'))$.

3. *Add the line segment connecting* $(w'', m(w''))$ *and* $(w', m(w'))$ *to the hull.*

4. *Repeat steps (2) and (3) until* $w' = w_0$.

5. *Return the hull that was constructed.*

An example of a hull difference calculated using this algorithm is shown in Figure 5.1.

First Differencing

An alternative preparation considered for data (which didn't work out very well) was first differencing. Again, there was concern that spectral data in general might not be i.i.d.. One way to compensate for data which is not i.i.d. is to apply a first differencing procedure to it—that is, to replace the data itself with the difference between each data point and its predecessor—viz.,

Algorithm 8 (First Differencing Algorithm) *Given a spectral measurement* $m : W \rightarrow [0.0, 1.0]$*, where* W *is a sequence of wavelengths* $w_0, w_1, ..., w_n$*, form a new function* $m' : W' \rightarrow [0.0, 1.0]$*, where* W' *results by removing the highest wavelength from* W*. Set* $m'(w_i) = m(\dot{w}_{i+1}) - m(w_i)$ *for* $i = 0, 1, ..., n-1$*. Return* w'*.*

5.4 Algorithms

5.4.1 Simultaneous Regression

Simultaneous linear regression fits a plane to a set of data using a least squares fitting algorithm. That is, given a set S of n data points $\langle i_1, i_{2,}..., i_m, d\rangle \in \prod_{i=1}^{m} I_i \times D$ where variables $I_1, I_2, ..., I_m$ are continuous independent variables and D is a continuous dependent variable, simultaneous linear regression finds the m-plane in $\prod_{i=1}^{m} I_i \times D$ (a linear function $r : \prod_{i=1}^{m} I_i \rightarrow D$) which minimizes the sum of the squares of the distances in the D dimension from each point in S to the plane. As it turns out, a fast algorithm for this least squares search is available, since it can be formulated as a linear transformation of the data points themselves. Once the regression plane for a set of data has been found, it can be used for prediction. Given a point $\langle a_1, a_{2,}..., a_m\rangle \in \prod_{i=1}^{m} I_i$, the predicted value is $r(\langle a_1, a_{2,}..., a_m\rangle)$.

Simultaneous linear regression was carried out in our experiments in different ways. In many cases, we used the REGRESS command in Minitab v. 10. In some cases, we performed the regression transformation ourselves using our own code. In other cases, we included regression among the algorithms to search over in the Model 1 package. We expect that these algorithms did not differ from one another in performance in any significant way.

Algorithm 9 (Simultaneous Linear Regression) *Let n be the number of sample points for $K-1$ variables, $\mathbf{Y}$ be an $n \times 1$ vector of estimated Y values, $\mathbf{X}$ an $n \times K$ matrix of data (X) values (with initial column of 1's), and e an $n \times 1$ vector of sample residuals. Calculate $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, which will be the $K \times 1$ vector of estimated coefficients. Calculate the p-value of each estimated coefficient. Determine which coefficients are within the specified significance. Return the corresponding minerals.*[4]

We applied simultaneous linear regression to the prediction of mineral class composition of rock spectra as follows: Taking wavelengths as units and rocks and minerals as variables, we constructed a background library by calculating a hull difference for each of the JPL mineral spectra. We next constructed a set of target spectra by interpolating each of the JHU rock spectra to the wavelength set of the JPL spectra and then calculating the hull difference for each of these interpolated spectra using the same method as for the JPL spectra. We then regressed each JHU rock against the JPL minerals and determined which JPL minerals had p-values less than 0.05 for each rock. We then determined which JPL mineral categories were represented among this set of JPL minerals with p-values less than 0.05 and returned this list of mineral categories. For instance, if the set $\{c03a, c03d, ts01a\}$ of JPL minerals had p-values less than 0.05,

[4] See Hamilton (1992).

then the following set of mineral categories was returned: $\{carbonates, tectosilicates\}$. (This same reporting scheme was used for several of the algorithms we tested.)

Simultaneous linear regression was also used by the Model 1 program in a different way. In the case of Model 1, wavelengths were taken to be variables and the various rock and mineral spectra were taken to be units. Additional variables were added for prediction—viz., variables reflecting the mineral classification of rock and mineral spectra according to the 17 JPL mineral categories. A simultaneous linear regression model was then calculated, regressing particular mineral content variables for the JPL large grain mineral data onto the 826 wavelength variables for that library. This model was then used to predict mineral class content for the JHU rock spectral data. The predicted results were then compared to the gold standard compiled by examining petrology in file descriptors for the JHU rock library. This same general schema was used for all of the Model 1 test procedures: Tests were carried out for all of the 17 JPL mineral categories, though only summary results for carbonates are reported below.

5.4.2 Stepwise Linear Regression

In stepwise linear regression, a set of predictors is given, and an automatic search is performed over these predictors for a locally optimal subset. The strategy is a "forwards and backwards" strategy: From an initial subset of predictors, predictors are either added to the specified set or removed from the current model based on an F-statistic. (A minimum and maximum F-statistic value are specified.) These steps are repeated until no variables can be added or removed. The specific stepwise regression algorithm used is the one from Minitab v. 10, which goes as follows:[5]

Algorithm 10 (Stepwise Algorithm) *Given set C of background variables a target variable Y, an initial set of predictor variables M, and two values F_0 and F_1 for the F-statistic (with $F_0 < F_1$).*

1. *Calculate the set $S_0 = \{P \in M : F(P) < F_0\}$. If $S_0 \neq \emptyset$, remove from M the predictor with the lowest F-statistic.*
2. *Calculate the set $S_1 = \{P' \in C - M : F(P') > F_1\}$. If $S_1 \neq \emptyset$, add to M the predictor with the highest F-statistic.*
3. *Repeat steps (a) and (b) until no more predictors can be added or removed.*

Applied to the prediction of mineral class composition, we used JHU rock spectra as targets and JPL large grain mineral spectra as background library, properly interpolated and hull-differenced, in the same fashion as for the simultaneous linear regression algorithm above. We used the same reporting scheme as well, reporting as classification for each JHU target rock the list of minerals represented in the JPL mineral spectra returned by the Stepwise algorithm.

A stepwise linear regression method was also used in the Model 1 analysis, though, as with simultaneous linear regression, variables and units were interchanged and prediction variables representing JPL mineral class composition for each rock or mineral were added to the model.

[5] In preliminary tests, we tried forwards stepwise regression and backwards stepwise regression in Minitab as well as the forwards-and-backwards combination procedure described below and found that the combination procedure worked the best.

5.4.3 Modified PC

The modified PC algorithm is an algorithm of our own design, motivated by the automatic causal discovery theory in Spirtes *et al.* (1993). A causal interpretation of the algorithm will be given in the next chapter; here the goal will be merely to motivate it as a response to regression.

There are three difficulties, one structural and two statistical, which make regression an inferior procedure in many applications—and in particular, in one of the applications of interest, where JHU rock samples are compared to JPL large grain minerals using regression:

1. Consider any two regression variables, X_1, X_2 among a set C of candidate causes of an outcome variable Y. Suppose X_1 and X_2 are correlated due to factors that influence both X_1 and X_2 values but are not themselves in C. Suppose, finally, that another factor U, not included in C, influences both X_1 and X_2. Then, even if X_1 has no influence on Y, even if there is no correlated error between X_1 and Y, and even if all common influences on X_1 and Y are included among the variables in C, for sufficiently large sample sizes the partial regression coefficient for X_1 will (almost certainly) have a significant value. The phenomenon is sometimes called conditional correlated error. In the present application, it can result in the identification of minerals that are not, in fact, components of the source.

2. Simultaneous linear regression computes the partial regression coefficient of a variable X_1 effectively by conditioning (assuming a Normal distribution) on all other regressors in the regressor set C—in our application, conditioning on all of the

other 134 minerals in the JPL library. While any one of these variables may be only loosely correlated with X_1, together they may be highly correlated with X_1. In that case, the covariation of X_1 and Y after partialing out the variation in Y due to other factors in C may be effectively zero. In the present application, multicollinearity can result in failing to identify a true component of the source.

3. The variance of the estimates of a simple regression coefficient is a function of the sample size. The variance of the estimates of a partial regression coefficient is a function of sample size and the number of other candidate causes, or regressors—that is, a function of the cardinality of C. The bigger the sample size and the smaller the number of other regressors, the smaller the variance. Assuming a Normal distribution, the trade off is one for one: adding an extra regressor variable is equivalent in its effect on the variance to reducing the sample size by one unit. In the present application, reducing the number of channels used for data analysis increases the variance of the estimates of regression coefficients. In the extreme case in which the number of variables is greater than the sample size, regression is ill-defined, and standard regression packages will not run at all. In our application, regression procedures will not run using the JPL library as the regressor set C and restricting the data to the channels with wavelengths in the interval $[2.0\mu m, 2.5\mu m]$, since there are 135 variables but only 120 channels.

Several remedies to this last difficulty can be considered. The wavelength interval $[2.0\mu m, 2.5\mu m]$, in this case, is chosen because previous work on carbonate spectra shows that this region has distinctive spectral features for carbonates. We could search for a larger range of wavelengths optimal for regression procedures in this application, but taken as a general rule, this would add a further search problem in every new application and might not improve accuracy of component identification. We could eliminate some of the minerals in the JPL library from the set of possible components of the source, but that would decrease the reliability of the procedure when those components or spectrally similar components are actually present in the source. We could use a stepwise regression procedure, but other experiments with small samples have found stepwise regression less reliable than the procedure used here (Spirtes, Glymour, and Scheines (1993)). A better solution to this problem is available.

Note that all three of the problems cited above with linear regression stem from a single structural feature of the regression procedure. In estimating the influence of a variable X on the outcome Y, regression conditions simultaneously on all other candidate variables—i.e., all of the other members of C. That is, in our (rather conventional, but not textbook) use of regression, we test the null hypothesis that X has no influence on Y (or is not a component of Y) by using the distribution of a test statistic that is conditioned on all other members of C.

There is an alternative procedure that minimizes the number of variables on which we must condition. It takes as input a set of background variables $C = \{X_1, X_2, ..., X_n\}$ together with a target variable Y not in C and dynamically eliminates variables from C using conditional independence facts, calculated from data. Variables are eliminated if they are independent of Y conditional on subsets of other remaining variables in C, where the cardinality m of the subsets increases in size ($m = 0, 1, 2, ...$) until no more variables can be eliminated from C. More formally:

Algorithm 11 (Modified PC Algorithm) *Given set C of background variables and target variable Y:*

1. *For each X_i in C, test the hypothesis that the correlation of X_i with Y is zero*[6]*; if the correlation of X_i with Y is zero, $C := C - \{Xi\}$;*

2. *For each X_i in C, and for each $Xj \neq X_i$ in C, test the hypothesis that the correlation X_i with Y, controlling for X_j , is zero; if the correlation of X_i with Y controlling for X_j is zero, $C := C - \{X_i\}$;*

3. *For each X_i in C and each $X_j, X_k \neq X_i$ in C test the hypothesis that the correlation X_i with Y, controlling for X_j, X_k is zero; if the correlation of X_i with Y controlling for X_j, X_k is zero $C := C - \{X_i\}$;*

...and so on, until no more members of C can be removed. Return C.

Applied to the prediction of mineral class composition in particular, we used the same procedure as for simultaneous regression and stepwise regression, above. We used JHU rock spectra as targets, JPL large grain mineral spectra as background library (all appropriately interpolated and hull-differenced), and we returned as classification for each JHU rock the list of mineral classes represented in the list of JPL spectra returned by the above algorithm.

[6] This test may be carried out in different ways; the method we used was to numerically integrate under the p.d.f. for conditional correlation as a function of sample size n and the number of compared variables k, $f(n,k) = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n-k+1}{2})}{\Gamma(\frac{n-k}{2})}(1-x^2)^{\frac{n-k-2}{2}}$, $(-1 < x < 1)$. (The number of "compared variables" is equal to the two plus the number of conditioning varaibles.) This formula is given in Cramer (1951), p. 412. The shape of this p.d.f. is a sharp spike about $x = 0$, the width of which changes depending on the value of n and k. A sample conditional correlation r may be judged to be nonzero just in case $|r| > d$, where d is chosen so that the area under the p.d.f. $f(n,k)$ for $|x| > d$ is equal to the chosen significance level α, usually 0.05.

5.4.4 Logistic Regression

The general idea of the logistic regression algorithm used by Model 1 is to map input to output variables as expressed by the following equation:

$$y = \frac{1}{1 + e^{-\left(w_0 + \Sigma_{i=1}^{N_{inputs}} w_i x_i\right)}} \tag{5.1}$$

where y is the output of the logistic function, the x_i's are the inputs, and the w_i's are the free parameters or coefficients. The cross-entropy cost function is used to fit the logistic function to a data set:

$$E_k = d_k \cdot \ln(1/y_y) + (1 - d_k) \cdot \ln(1/(1 - y_k)) \tag{5.2}$$

where E_k is the error for the k-th data record, y_k is the output produced with the input vector of the k-th data record, and d_k is the desired output for the k-th data record. Cross entropy over the whole training set is

$$E = \Sigma_{k=1}^{N_{train}} E_k \tag{5.3}$$

which is to be minimized using a gradient descent procedure. The algorithm itself is divided into a training phase and a testing phase.

Algorithm 12 (Logistic Regression) *For testing, for each pattern x_k in the training set, computing the logistic output according to the above formula, the compute the gradient of the entropy error with respect to each weight w_i due to x_k, then compute the change in weights and update the weights accordingly. Repeat until convergence is reached. For testing, compute the logistic output for each pattern x_k in the test set.*

Applied to the problem of predicting mineral class composition of spectra, wavelengths were taken to be variables, with the various rock or mineral spectra as units, and prediction variables were added to the model to reflect the JPL mineral class composition of the various rock or mineral spectra. A logistic regression model was constructed according to the training portion of the above algorithm using the spectra in the JPL large grain mineral library. Predictions were then generated using this model for the spectra in the JHU rock library, and these predictions were then compared to the gold standard constructed by examining petrology in the JHU rock library file descriptors.

5.4.5 Classification and Regression Trees (CART)

As implemented in Model 1, these are binary decision trees which split on a single value of a single variable at each node (e.g., if $V < x$ then first branch, else second branch, for some x). Optimal splitting criteria at any specific node are found as follows:

$$\Phi(s|t) = Max_i(\Phi(s_i/t)) \tag{5.4}$$

$$\Phi(s|t) = 2P_L P_R \Sigma_{j=1}^{classes} |P(j/t_L) - P(j/t_R)| \tag{5.5}$$

After the full tree is grown, branches that reduce the overall effectiveness of the tree are pruned away by computing a strength $g(t)$ for each non-terminal node t, taking into account the misclassification rate and the statistical significance of the node.

Algorithm 13 (CART) *A binary decision tree is constructed with rules of the form "If $V < x$ then first branch, else second branch," where the "x" in each rule is selected* *to maximize $\Phi(s|t)$, defined as follows:*

$$
\begin{aligned}
\Phi(s|t) &= 2P_LP_R\Sigma_{j=1}^{classes}\left|P(j/t_L) - P(j/t_R)\right| \\
t_L &= \textit{left offspring of node } t \\
t_R &= \textit{right offspring of node } t \\
P_L &= \frac{\textit{total \# of patterns at } t_L}{\textit{total \# of patterns in training set}} \\
P_R &= \frac{\textit{total \# of patterns at } t_R}{\textit{total \# of patterns in training set}} \\
P(j|t_L) &= \frac{\textit{total \# of class } j \textit{ patterns at } t_L}{\textit{total \# of patterns at } t} \\
P(j|t_R) &= \frac{\textit{total \# of class } j \textit{ patterns at } t_R}{\textit{total \# of patterns at } t}
\end{aligned}
$$

Once the full tree is constructed, branches are pruned at nodes t that have the smallest score $g(t)$ defined as follows:

$$
\begin{aligned}
g(t) &= \frac{R(t) - R(T_t}{|T_t'| - 1} \\
R(t) &= r(t)p(t) \\
r(t) &= 1 - Max_i\left[\frac{\textit{\# of class } j \textit{ patterns at } t}{\textit{total \# of patterns at } t}\right] \\
p(t) &= \frac{\textit{\# of patterns at } t}{\textit{total \# of patterns in the training set}} \\
R(T_t) &= \sum_{t' \in T_t'} R(t')
\end{aligned}
$$

where $|T_t'|$ is the number of leaves in the subtree headed by node t.

CART models implementing this algorithm were applied to the prediction of mineral class composition of spectra using the same training and testing regime as for the the other Model 1 tests. With the 826 wavelengths of the JPL library taken as variables, the various mineral and rock spectra taken as units, and with prediction variables added to the model to represent mineral class composition according to the 17 JPL mineral categories, a CART model was constructed using the JPL mineral spectral data. This model was then used to predict mineral class composition for each of the 192 JHU samples, and these predictions were compared to the gold standard constructed by examining petrology in file headers.

5.4.6 Naive Bayes Classifiers

The naive Bayes algorithm used by Model 1 assumes that the inputs are independent and then estimates the conditional probability density function for each output class. Since independence is assumed, the marginal densities for the inputs $P(x[i]|y)$ are multiplied together to produce the density estimate $P(x|y)$. For a very simple example, the naive Bayes output for a binary classifier would be:

$$\frac{P(Y=1)P(x|y=1)}{P(y=0)P(x|y=0)+P(y=1)P(x|y=1)} \tag{5.6}$$

This approach is applied to the prediction of mineral class composition for rock spectra in a similar way to the other Model 1 tests above. Wavelengths were taken as variables, spectra for rocks and mineral as units, and additional variables were added to represent mineral class composition according to the 17 JPL mineral categories. A naive Bayes model was then constructed using the JPL mineral spectral data, and this model was then used to predict mineral class composition for the JHU rock spectral data, with results compared to the gold standard constructed by examining petrology of JHU rocks in file headers.

5.4.7 Backpropagation

Backpropagation is a technique for training neural networks to perform classifications. In a typical backpropagation scheme, one sets up a feedforward network of nodes in three layers—the input layer, the hidden layer, and the output layer. Patterns of activation are applied to the nodes in the input layer, and if they exceed a certain threshold, a signal is sent on from each of the nodes in the input layer to each of the nodes in the hidden layer. Each node in the hidden layer, therefore, receives input from each node in the input layer. A weighted sum of these inputs is then sent to a filter function, and if the output of the filter function exceeds a certain threshold, a signal is sent on to all of the nodes in the output layer. Each node in the output layer thus receives a pattern of activation from nodes in the hidden layer, just as nodes in the hidden layer received a pattern of activation from nodes in the input layer. Activation is propagated from the hidden layer to the hidden layer to the output layer, and depending on exactly how inputs at each stage are weighted, patterns at the input layer are transformed into other, specific patterns at the output layer.

Model 1 implements a backpropagation neural net of this sort using the following algorithm, divided into training and testing phases:

Algorithm 14 (Backpropagation) *In the training phase, for each pattern in the training set, and for each node j of layer k (except the output layer), compute its output $y_{jk} = \frac{1-e^{-net_{jk}}}{1+e^{-net_{jk}}}$ where $net_{jk} = \sum_{l=1}^{N_{k-1}+1}$. Compute the average root-mean-square error $E_i = \frac{1}{2}\sum_{j=1} J(d_{ij} - yij)^2$. Change the weight w_{ij} between node j and input node l as follows: $\delta w_{jl} = -\eta \frac{\partial E_i}{\partial w_{jl}} + \alpha(\delta w'_{jl} - \delta w''_{jl})$, where η is the step size of the steepest descent* *and α is the momentum term to which the minimization is "smoothed" over successive descents. In the testing phase, for each pattern in the test set, for each node j of layer k (except the output layer), compute its output $y_{jk} = \frac{1-e^{-net_{jk}}}{1+e^{-net_{jk}}}$ where $net_{jk} = \sum_{l=1}^{N_{k-1}+1}$. For the output layer, compute $y_{jk} = \frac{1}{1+e^{-net_{jk}}}$ where $net_{jk} = \sum_{l=1}^{N_{k-1}+1} w_{lj}x_l$.*

This algorithm was used to predict mineral class composition in the same way as the other Model 1 tests. Wavelengths were taken to be variables, mineral and rocks units, with additional variables added for prediction to represent mineral class composition for the 17 JPL mineral categories. A backpropagation neural net model was constructed using the mineral spectra of the JPL library; the rock spectra of the JHU library were then used in the testing phase to generate predictions of mineral class composition, which were then compared to the gold standard classifications generated by examining petrologies of the JHU samples in file descriptors.

5.5 Experiments

Five experiments are reported. In the first, the expert classification from Chapter 4 is compared directly to classification by the modified PC algorithm on all 17 JPL mineral classes. In the second, various preparations and identification schemes are compared to each other for a single limestone sample measured in the field. In the third, various algorithms, preparations methods, and identification schemes are compared to one another on the task of identifying carbonates from among the JHU rocks. In the fourth, various algorithms, preparation methods, and identification schemes are compared to one another on the task of identifying carbonates from among the Silver Lake field samples. In the fifth, the Model 1 data mining program is used to quickly compare hundreds of other models to the models tested on carbonate identification.

5.5.1 17 Class Comparison of Expert to Modified PC, JHU Rock Samples

In the previous chapter, an expert spectroscopist judged, for each of 191 JHU rock samples and for each of the 17 JPL mineral classes, which classes were present in the composition of which targets. The modified PC algorithm was then run on the same spectra, using the entire $[0.4\mu m - 2.5\mu m]$ range and outputting a JPL class if any representative of that class was found for the sample. For eight of the seventeen JPL mineral classes, no representative, or no more than two representatives, were present in the JHU library; for those classes the experiment is of no significance. For tectosilicates, carbonates, and nesosilicates (i.e., the three most accurate categories for the expert), there was a tradeoff between expert performance and modified PC performance; expert performance was more accurate, whereas modified PC performance had better coverage. For phyllosilicates, inosilicates, and oxides, the expert outperformed the modified PC algorithm in both accuracy and coverage. For elements and sulfides, there was no significant difference between expert and modified PC perormance. The most interesting comparisons are for the three hightest accuracy categories for the expert; for these, the positive identifications of the human expert have significant reliability and can be approximated (somewhat less reliably) by the modified PC algorithm. Numerical results are given in Table 5.3 and shown graphically in Figures 5.2 and 5.3.

Table 5.3: Calculated values presented in Figures 5.2 and 5.3. ("Tr" = truth, according to the exclusive gold standard G_2; "Exp" = subject from expert experiment; "mPC" = modified PC.)

	P(Tr | Exp)	P(Exp | Tr)	P(Tr | mPC)	P(mPC | Tr)
tectosilicates	1.00	0.18	0.83	0.56
carbonates	0.96	0.39	0.63	0.46
nesosilicates	0.77	0.19	0.57	0.53
phyllosilicates	0.72	0.69	0.52	0.51
inosilicates	0.59	0.63	0.43	0.29
oxides	0.49	0.38	0.23	0.15
elements	0.12	0.5	0.07	0.05
sulfides	0.12	0.22	0.07	0.22
cyclosilicates	0.00	0.00	0.00	*
halides	0.00	*	0.00	0.00
hydroxides	0.00	0.00	0.00	0.00
sulphates	0.00	*	0.00	*
arsenates	*	*	0.00	0.00
borates	*	*	0.00	0.00
phosphates	*	0.00	0.00	*
sorosilicates	*	0.00	0.00	*
tungstates	*	*	*	*

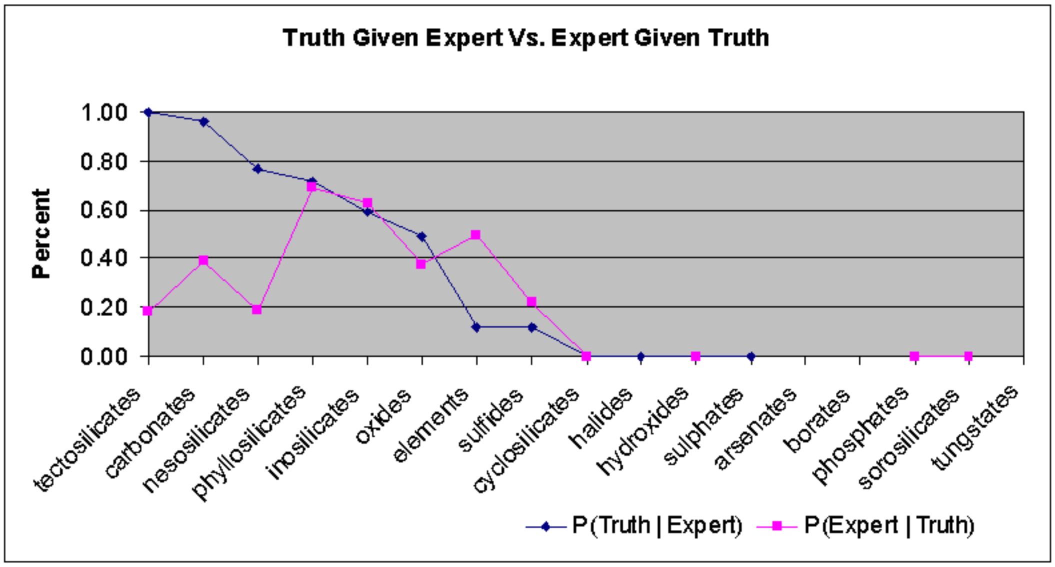


Figure 5.2: A comparison of carbonate identification by an expert examining unlabeled, uniformly formatted graphs presented in random order ("Expert") to true carbonate identification as judged by an expert examining petrological information included in headers of JHU data files ("Truth").

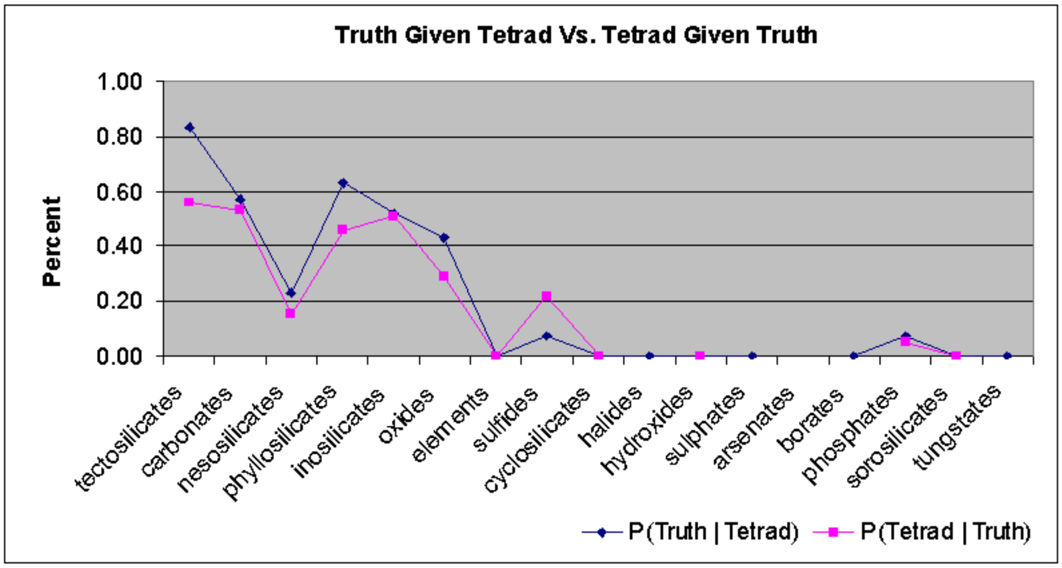


Figure 5.3: A comparison of carbonate identification by the modified PC algorithm ("Tetrad") to true carbonate identification as judged by an expert examining petrological information included in headers of JHU data files ("Truth").

Table 5.4: Analyses of the spectrum of a sample taken near Silver Lake, CA, known to contain dolomite and calcite and described as "dolomite with calcite veins." Only minerals with positive regression coefficients are shown. The mineral 'co3' is dolomite; 'co5' is calcite (a and c suffixed denote different varieties of calcite); 'c10' is a carbonate that is neither calcite nor dolomite; identifiers beginning with'cs' refer to cyclosilicates; 'a10' is an arsenate.

Water lines removed		Data Treatment	Water Lines Included	
modified PC	Stepwise		Stepwise	modified PC
c03d c05a	c03d cs02a cs04a	None	co3d cs02a co5c	c05d
co3d co5c	co5c co3d co5a cs02a a01a	Hull Difference	co3a cs02a co5a cs04a a01a	co5c cs01a cs02a
co5c	co3d co5c c10a	First Difference	co3e	None

5.5.2 Single Sample Limestone Analysis

Prior to receipt of the Silver Lake field data, a single limestone spectrum measured by Roush and Glymour near Silver Lake, California was sent to us for examination. We used this single limestone sample to get a rough estimate of the performance of stepwise regression and the modified PC algorithm as applied to spectra measured in the field. We used the JPL Large Grain Mineral Library as a background library for each algorithm, in three different preparations: (a) as raw spectral intensities, (b) as hull differences of intensities, and (c) as first differences of intensities. As the sample was measured in the field, noise was present in the vicinity of $1.9\mu m$ due to atmospheric moisture ("water lines"). Therefore, we examined the effect of including water lines in the analysis as well as the effect of excluding water lines. Predictions on each method of the mineral composition of the single sample spectrum are shown in Table 5.4.

These experiments, it should be pointed out, were performed "blind." That is, when the experiments were performed, a scientific analysis of the composition of the sample was not available. As Table 5.4 shows, the modified PC algorithm estimates the composition of the sample to contain calcite and dolomite, when water lines are removed, both for data with no preparation and for data which was hull-differenced. We therefore judged, based on this observation, that the sample contained calcite and dolomite. Independent analysis (by acid test) at Washington University showed the specimen to be dolomite with calcite veins. If, however, first differences of spectral intensities were used, no output was obtained (because the correlations with the spectra of the true components, calcite and dolomite, then fall to zero).

5.5.3 Carbonate Identification, Silver Lake Samples

In an experiment designed to broaden the range of procedures compared, we considered a variety of methods for predicting carbonate content in the Silver Lake field samples. The following combinations of procedures were used to analyze the data:

1. The modified PC algorithm, seeking to recognize any carbonates, using a restricted interval of wavelengths with intensity patterns characteristic of carbonates.

2. The modified PC algorithm, seeking to identify carbonates from calcites and dolomites only, using a restricted interval of wavelengths with intensity patterns characteristic of carbonates.

3. The modified PC algorithm, seeking to recognize any carbonates, using all wavelengths available from the instrument.

Table 5.5: Performance of modified PC (with variations) and simultaneous linear regression (with variations) against the 21 Silver Lake field samples, testing the identification of carbonates. Modified PC results are shown in (a); regression results are shown in (b). Results obtained from an expert system experiment examining the same data are also shown in (b).

(a)

	mPC all Carbs [0.4,2.5]	mPC all Carbs [2.0,2.5]	mPC Calcite & Dolomite [0.4,2.5]	mPC Calcite & Dolomite [2.0,2.5]
# Carbs Attempted	19	13	16	12
# Carbs Correct	13	12	13	12
# Non-carbs MisID'd	6	1	3	0
# Carbs MisID'd	0	1	0	1
Total # Errors	6	2	3	1
P(Carb | Carb ID)	0.68	0.92	0.75	1.00
P(Carb ID | Carb)	1.00	0.92	1.00	0.92

(b)

	Regr all Carbs [0.4,2.5]	Regr all Carbs [0.4,2.5] Pos. Coef.	Regr Calcite & Dolomite [0.4,2.5]	Regr Calc & Dol [0.4,2.5], Pos. Coef.	Expert System
# Carbs Attempted	21	20	20	15	9
# Carbs Correct	13	13	13	11	9
# Non-carbs MisID'd	8	7	7	4	0
# Carbs MisID'd	0	0	0	2	4
Total # Errors	8	7	7	6	4
P(Carb | Carb ID)	0.62	0.65	0.65	0.73	1.00
P(Carb ID | Carb)	1.00	1.00	1.00	0.85	0.69

4. The modified PC algorithm, seeking to recognize only calcites and dolomites, using all wavelengths available from the instrument.

5. Linear regression, seeking to recognize the presence of any carbonates using all wavelengths available from the instrument.

6. Linear regression, seeking to recognize the presence of any carbonates using all wavelengths available from the instrument, but reporting only components with positive regression coefficients.

7. Linear regression, seeking to recognize carbonates from calcites or dolomites only, using all wavelengths available from the instrument.

8. Linear regression, seeking to recognize carbonates from calcites or dolomites only, using all wavelengths available from the instrument, but reporting only components with positive regression coefficients.

9. An expert system seeking to recognize the presence of any carbonates.[7]

The data, with sample names in the leftmost column, are given in Appendix B. Note that the number of samples correctly estimated to contain carbonates, given at the bottom of each column, is based on the assumption that the expert field identifications (shown also in Table C.1) represent the truth. This is a reasonable assumption, since expert field identifications rely on both examination of physical samples and examination

[7] The expert system in question used the following procedure: (a) Apply a hull fit to the spectrum to subtract the background; (b) Use first and second derivative heuristics to obtain a list of spectral features; (c) Use a simple noise test to obtain a measurement of noise in the relevant wavelength range; (d) Use the facts from steps (b) and (c) as input to a simple forward-chaining expert system, consisting of a pattern matcher operating in association with a rudimentary math parser. Rules for identifying carbonates included looking for properly shaped features in the 2.0-2.5 μm range.

of measured spectra and, where tested, agree with laboratory analysis of the samples. Assuming instead that remote expert classifications represent the truth would change values only for samples ‘jawa’ and ‘R2D2,’ increasing the scores of two regression procedures by one and the score of the expert system by two. Note also that all data were hull-differenced for all procedures.

5.5.4 Carbonate Identification, JHU Rock Samples

The same collection of procedures from the Silver Lake carbonate identification experiment (with the exception of the expert system) was applied to the JHU large grain rock library. The results are shown in Table 5.6, in a format parallel to the comparative Silver Lake results of Table 5.5. The results from the human expert are included in the same table for comparison, since they were targeted at the same set of spectra.

These results, taken together, suggest that the modified PC procedure, in combination with the use of a restricted set of wavelengths as data preparation, and the identification of carbonates through recognition of calcites and dolomites, outperforms all of the other eight procedures considered and is nearly as good as expert identification in the field using physical examination and spectra. Three of the four regression procedures are essentially useless overfittings which guess that almost all samples contain carbonates and so are correct at about the relative frequency of carbonates in the data set (= 0.62). The fourth regression procedure is little better. The results also suggest that restricting wavelengths and identifying carbonates through calcites and dolomites are important to the reliability of the procedure—without these features, the modified PC procedure overfits almost as badly as regression and is inferior to an expert system

Table 5.6: Performance of modified PC (with variations) and simultaneous linear regression (with variations) against the 192 JHU large grain rock samples, testing the identification of carbonates. Modified PC results are shown in (a); regression results are shown in (b). Relevant data from the Model 1 experiment are also shown in (a), and relevant data from the expert experiment of Chapter 4 are shown in (b).

(a)

	mPC all Carbs [0.4,2.5]	mPC all Carbs [2.0,2.5]	mPC Calcite & Dolomite [0.4,2.5]	mPC Calcite & Dolomite [2.0,2.5]	Model 1
# Carbs Attempted	58	63	42	41	73
# Carbs Correct	38	47	36	38	36
# Non-carbs MisID'd	20	16	6	3	37
# Carbs MisID'd	54	45	56	54	56
Total # Errors	74	61	62	57	93
P(Carb | Carb ID)	0.66	0.60	0.90	0.93	0.49
P(Carb ID | Carb)	0.41	0.51	0.39	0.41	0.39

(b)

	Regr all Carbs [0.4,2.5]	Regr all Carbs [0.4,2.5] Pos. Coef.	Regr Calcite & Dolomite [0.4,2.5]	Regr Calc & Dol [0.4,2.5], Pos. Coef.	Human Expert
# Carbs Attempted	192	191	154	176	25
# Carbs Correct	92	91	79	70	24
# Non-carbs MisID'd	100	100	75	86	1
# Carbs MisID'd	0	1	13	2	68
Total # Errors	100	101	88	88	69
P(Carb | Carb ID)	0.48	0.48	0.51	0.51	0.96
P(Carb ID | Carb)	1.00	0.99	0.86	0.98	0.26

modeling an expert spectroscopist.

In the laboratory test, the human expert correctly identified 24 carbonates, and of these, 20 were limestone or marble samples, principally composed of calcite and dolomite. In effect, the human expert was a calcite or dolomite detector. The modified PC algorithm using truncated spectra but attempting to identify any JPL carbonates in the JHU library, rather than just calcites or dolomites, correctly identified all 28 of the limestone and marble samples in the JHU library. Of the 35 other JHU samples the modified PC algorithm identified as carbonates, 19 were carbonates and 16 were not—in other words, outside of limestones and marbles, the estimated chance that an identification of a carbonate was correct was only a little better than chance. When the modified PC algorithm using truncated data reported carbonate presence only when it first output calcite or dolomite, outside of limestones and marbles the probability that an identification of a carbonate was correct was about 0.82. Put another way, the version of the program outputting carbonate only if calcite or dolomite were identified found 14 of the 19 carbonates found by the unrestricted modified PC algorithm using the same wavelengths, and avoided 13 of the errors of the latter procedure. Two explanations are possible: many of the carbonate samples besides marbles and limestones may have contained calcite or dolomite; alternatively, the calcite and dolomite spectra in the region [2.0 mm, 2.5 mm] may be so highly correlated with the spectra (in that region) of certain other carbonates that a calcite or dolomite detector will detect them as well. The second explanation is certainly true, and the first may also be true.

The correlations among the hull difference truncated to [2.0 mm, 2.5 mm] for the 15 JPL carbonates show that one or another of the calcites (c03a, c03d, c03e) are

Table 5.7: Correlation matrix of the 15 JPL carbonate spectra truncated to the interval $[2.0\mu m, 2.5\mu m]$.

	c01a	c02a	c03a	c03d	c03e	c04a	c05a
c02a	0.671						
c03a	0.922	0.439					
c03d	0.943	0.533	0.970				
c03e	0.959	0.499	0.986	0.957			
c04a	-0.260	-0.286	-0.202	-0.105	-0.251		
c05a	0.736	0.206	0.908	0.799	0.867	-0.227	
c05c	0.686	0.140	0.857	0.721	0.828	-0.280	0.985
c06a	-0.136	-0.251	0.034	0.112	-0.084	0.787	0.073
c07a	0.465	0.309	0.461	0.427	0.448	-0.059	0.582
c08a	0.977	0.746	0.870	0.925	0.907	-0.228	0.641
c09a	0.948	0.513	0.983	0.948	0.985	-0.265	0.907
c10a	-0.291	0.134	-0.332	-0.196	-0.394	0.064	-0.413
c11a	0.941	0.712	0.867	0.936	0.879	-0.059	0.654
c12a	0.528	0.338	0.542	0.572	0.500	0.200	0.576

	c05c	c06a	c07a	c08a	c09a	c10a	c11a
c06a	-0.007						
c07a	0.605	0.159					
c08a	0.574	-0.079	0.403				
c09a	0.870	-0.028	0.552	0.899			
c10a	-0.470	0.336	-0.006	-0.158	-0.340		
c11a	0.571	0.133	0.454	0.976	0.888	-0.086	
c12a	0.540	0.459	0.915	0.507	0.593	0.093	0.616

strongly correlated with c01a (strontianite) (r = 0.943), c09a (siderite) (r = 0.985) and c11a (smithsonite) (r = 0.936), while one of the dolomites (c05a, c05c) was strongly correlated with c09a (siderite) (r = 0.907). Some of the seven remaining carbonate spectra (e.g., trona) look quite different in the relevant region. The correlation matrix of the truncated spectra is given in Table 5.7, and graphs of the hull-differenced spectra (untruncated) are given in Appendix A.

5.5.5 Carbonate Identification, JHU Rock Samples, Model 1

Model 1 is a commercial data mining program which automatically searches through a bank of models for the model which best predicts data in a training set. The best predictors from this search can then be used to perform classifications on a test set. We presented our JHU data to Model 1 and configured the program so that it would automatically build and compare hundreds of different models, ranking the models in the process. The types of models compared were simultaneous linear regression, stepwise linear regression, logistic regression, classification and regression trees (CART), naive Bayes classifiers, probabilistic decision trees, and backpropagation neural nets.[8] Variations of these models were attempted, some of which used cross-validation and some of which did not.

Model 1 found that a cross-validated linear regression procedure performed best on the JPL library (a simple linear regression procedure performed worst). We applied this best-performing procedure to a the samples of the JHU rock library to predict carbonate composition. Model 1 listed the samples of the JHU rock library in order from those most likely to contain carbonate (according to the algorithm tested) to those least likely to contain carbonate (according to the algorithm tested) and reported how far down in the ordering one must go for any specific number of correct identifications to be obtained. The result for 36 correct carbonate identifications is shown in the next to right-most column of Table 5.6. Results for other selections are similarly poor.

[8] See Section 5.4 for descriptions of these algorithms.

5.6 Summary

We found that the modified PC algorithm outperformed all of the other machine algorithms tested and performed comparably to the human expert on the task of identifying carbonates from spectral data. In the 17-class experiment we found that the modified PC algorithms performed comparably to a human expert in the field when compared to a suitably chosen gold standard. In the single-sample experiment, we found that the modified PC algorithm could be an excellent blind predictor of at least one type of mineral class component—viz., carbonates. We therefore followed up with two experiments in which prediction of carbonate content was carried out using a variety of different algorithms. In the carbonate study using Silver Lake data, we found that one particular variation of the modified PC algorithm outperformed all other variations of the modified PC algorithm, as well as all variations of regression tested, when compared to a gold standard of expert classifications in the field. This particular variation also outperformed an expert system. In the carbonate study using JHU data, we found similar results; the same variation of the modified PC algorithm outperformed all other variations of the modified PC algorithms and all variations of regression tested when compared to a suitable gold standard. It performed comparably to the humen expert from Chapter 4. Finally, using the commercial Model 1 package, we showed that the modified PC algorithm outperformed several other types of algorithms as well (multiple variations each of simultaneous linear regression, stepwise linear regression, logistic regression, classification and regression trees, naive Bayes classifiers, probabilistic decision trees, and three-layer feedforward backpropagation nets).

Chapter 6

Conclusions and Prospects

6.1 Discussion of Experiments

The modified PC algorithm was treated in Chapter 5 as one classification algorithm method among many. One possible reason for its success in the mineral classification from spectra problem is its sensitivity to causal structure. This suggests a response to the skeptical worries from Chapter 2 that were addressed experimentally in Chapters 4 and 5—viz., whether human expertise is needed to do causal discovery with real scientific data, and whether Tetrad-style algorithms perform well under conditions of mixing. The experimental results of Chapters 4 and 5 speak directly to these issues, so aside from pointing out the results themselves and casting them in causal terms where appropriate, very little else needs to be said by way of response.

To Freedman and Humphrey's complaint that not all knowledge relevant to causal search can be programmed into a computer (the "Automation Principle"), one need only say that the need never arose for us to program extraordinary amounts of background knowledge into the modified PC algorithm in order for it to be as successful at solving a very difficult causal discovery problem as a domain expert. It is true that for some mineral categories the domain expert outperformed the algorithm, but for other categories the algorithm outperformed the domain expert, so the performances were comparable. Even if there happened to be relevant background knowledge for this problem which could not be programmed into a computer (the existence of which has not be demonstrated), it turned out to be irrelevant for this particular task.

Cartwright's worry that new algorithms have to be invented for each new causal discovery problem requires a slightly more cautious response. On the one hand, we did not use the standard PC and FCI algorithms in our experiments to tackle the mineral class identification from spectra problem, so in that sense we invented a new algorithm to solve a specific problem. We also built important background knowledge into the modified PC model which is not true of every causal discovery problem. However, it would be an overstatement to insist that an algorithm completely different in character from the PC or FCI algorithms was invented. The same theory used in the PC and FCI algorithms was simply applied against a specific set of background assumptions; the modified PC algorithm is essentially just a simplification of the first part of the PC or FCI algorithm, not a new invention. In theory, there is no need to use the modified PC algorithm in place of, say, the FCI algorithm. The FCI algorithm was tested on subsets of the same data (20 to 30 variables) and produced similar results to those produced by the modified PC algorithm. The problem was that the FCI algorithm became too unwieldy when applied to larger variable sets (the number of variables compared simultaneously in the modified PC experiments was well in excess of 100). So the motivation for using the more streamlined algorithm was primarily one of scale.

As for the issue of mixtures, two types of mixtures were referred to in Chapter 2—physical mixtures and mixtures of records. With regard to physical mixtures, the earlier discussion of expert knowledge in the domain of mineral and rock spectroscopy[1] illustrate how the various processes involved in absorption and reflection combine together in a physical sense. The fact that the object of spectroscopic measurement is such a complex physical mixture of processes, so many of which are nonlinear, might lead one to believe that spectral measurements resulting from such heavy mixtures might not generate the kind of data that would allow a Tetrad-style algorithm to produce reliable causal inferences. Yet if we look at the data from Chapters 4 and 5 we see that the results of the algorithmic analysis are just as reliable as the results of the expert analysis. The modified PC algorithm therefore performs well under conditions of heavy physical mixing on this problem.

To see how the modified PC algorithm presents us with a mixture of records of the sort Cartwright points to (and of the sort pointed to in the Kendall and Simpson examples), it helps to draw a causal diagram for the mineral composition from spectra problem. Figure 6.1 illustrates the basic idea. For each of a series of background mineral intensities $I_1, I_2, ..., I_n$, we assume initially that there is a common cause between that intensity and the intensity of some target rock I_t. The goal of the modified PC algorithm in this context is to eliminate as many of these common cause assertions as possible through considerations of conditional independence. We assume that if a rock contains some mineral as a component, then it's much more likely that minerals of the same

[1] See Chapter 3.

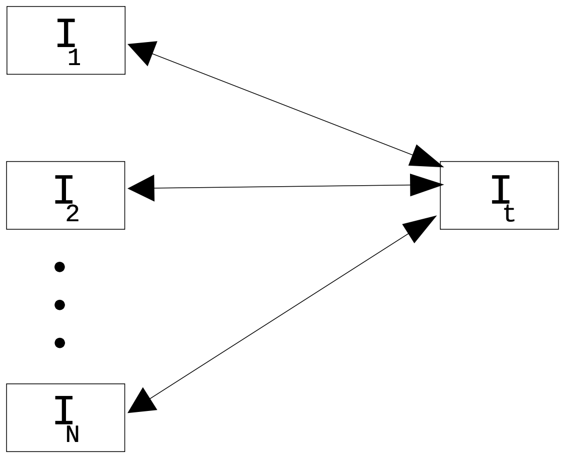


Figure 6.1: A causal interpretation of the initial state of the modified PC algorithm as applied to spectral data before removing any edges. The algorithm removes as many common cause assertions ($\leftrightarrow$) as possible based on conditional independence considerations. The I_i variables represent measured spectral intensities of background minerals at specific wavelengths. (Wavelengths are taken to be units in the analysis.) The I_t variable represents spectral intensity of the target rock at the same wavelengths.

type as that component will have common cause connections with the target rock. By eliminating as many of the common cause connections as possible through the modified PC procedure, background minerals which have strong connections with the target rock (and are therefore very likely to be related to its mineral components) will be discovered.

Notice that the existence of possible common cause connections among the I_i's themselves is irrelevant to the modified PC algorithm. For instance, if there happens to be a common cause connection between I_1 and I_2 and if we discover that I_1 is independent of I_2, it still follows that the common cause connection from I_1 to I_2 should be removed from the graph, because the path $I_1 \leftrightarrow I_2 \leftrightarrow I_t$ contains a collider.[2]

[2] When we tested small subsets of the spectral intensity variables against the FCI algorithm, we discovered that the results typically contained common cause arrows not just between the background

The simple graph in Figure 6.1 can be expanded somewhat, though the degree to which we can incorporate new variables is limited by our understanding of the quantum processes involved and how they react to incident radiation across mineral types. It's tempting to include latent variables that represent the general surface reactivity of particular minerals or rocks to incident light at particular wavelengths, as shown in Figure 6.2. In this diagram, each of the surface reactivity variables, together with incident light (L),[3] is causally responsible for the intensities we measure for each of the background minerals and the target rock. We might think of each $L \rightarrow F_i \rightarrow I_i$ path (as well as the path $L \rightarrow F_t \rightarrow I_t$) as a very simple model for a spectrometer at a particular wavelength. When we measure the intensity of a rock at a particular wavelength, we shine light on the rock at that wavelength and expect a certain intensity of light at the same wavelength to be reflected back.[4] The actual processes by which light is absorbed and reflected at particular wavelengths for rocks and mineral are still largely mysterious. Moreover (and more to the point), what we know about the absorption and reflection of light for certain minerals does not easily transfer to other minerals, so it's unclear how to draw a general causal graph (general across rocks and minerals) in which more specific variables than the ones in Figure 6.2 are included.

Notice that the expanded graph supports the modified PC algorithm as well. If

minerals and the target but also between the background minerals themselves. This appears to confirm the general analysis being given here of how the PC algorithm applies to the mineral spectroscopy data.

[3] The representation of light as a single casual parent for all of the F variables is unrealistic, since different spectra were measured using different light sources, but it doesn't in fact matter from the point of view of the modified PC algorithm whether light is represented as one latent parent of all of the F variables (where we condition on this variable) or as a separate latent parent of each F variable separately.

[4] There are wavelength-shifting effects in spectroscopy as well which are ignored for purposes of this analysis.

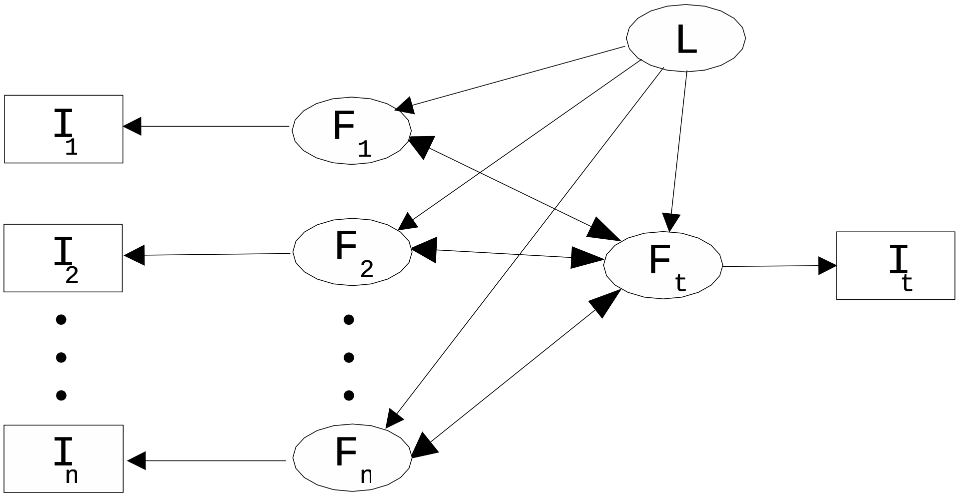


Figure 6.2: The I variables are as in Figure 6.1; surface reactivity for particular minerals or rocks are represented using the F variables. L represents intensity of incident light (values of which we may ideally condition on for spectral measurement).

$I_1 \perp\!\!\!\perp It \mid I_2$, for example, the edge $F_1 \leftrightarrow F_T$ must not be in the graph. In fact, it doesn't matter whether there are any common causal connections among the I_i's or among the F_i's; the algorithm is still supported. For instance, if there is a common cause $I_1 \leftrightarrow I_2$, and we discover that $I_1 \perp\!\!\!\perp I_t$, it still follows that the edge $F_1 \leftrightarrow F_t$ should not be in the graph, because of the collider along the path $I_1 \leftrightarrow I_2 \leftarrow F_2 \leftrightarrow F_t \rightarrow I_t$. Common causes or not, the modified PC algorithm for this diagram can be expected to eliminate edges $F_i \leftrightarrow F_t$ that represent a lack of common causal connection between background library minerals and target rocks.

If we start with the assumption that there are common causal connections to be discovered between intensity data for background minerals and some target rock, it becomes evident fairly quickly that the causal analysis we are engaged in is an extremely complicated mixture in Cartwright's sense. Processes of reflection for minerals and rocks, so far as spectrographic measurement is concerned, act at particular wavelengths—that is, the causal processes acting to produce observed intensities at the given wavelengths measured are wavelength-specific. We could produce a graph like the one in Figure 6.1 at a particular wavelength rather than considering how the graph reacts across wavelengths. Even though we don't have enough data at any particular wavelength to determine with confidence what such a graph would look like (since we typically only have one data point at each wavelength), the observed intensity at each wavelength is certainly the result of a causal process acting at that wavelength, for which many more measurements could in principle be taken. The data we have though is not restricted to a particular wavelength; instead, it spans the range of wavelengths under consideration and is organized (through interpolation and machine design) so that a particular set of wavelengths is used throughout for a variety of different minerals and rocks. These wavelengths themselves are treated as units when applying the modified PC algorithm—a perfectly reasonable way to treat the data under the circumstances.[5] The causal graphs which result are therefore not causal graphs for particular wavelengths but rather causal graphs for data from different wavelengths taken together—a mixture, in exactly Cartwright's sense.

The relevant question, therefore, is whether the causal graphs constructed for each of the target rocks for the modified PC experiment summarized in Chapter 5 constitute good evidence that the modified PC algorithm is a reliable predictor of causal

[5] Given a rectangular array of data of the format shown in Table 5.1, one has to choose in a data analysis whether to take wavelengths as units or minerals/rocks as units. In the modified PC analysis we chose wavelengths; in other analysis, we chose minerals/rocks.

structure under conditions of heavy mixing, in this second sense (mixture of records). There is no question that spectral data used in the experiments are heavily mixed. There is also no question that the modified PC algorithm performs comparably to a human expert on exactly the same task of inferring mineral composition of target rocks. By the criterion that Freedman and Humphreys suggest, this should count as success; the modified PC algorithm performs comparably to a human expert on an extraordinarily difficult causal inference problem. Given that expert performance is the only criterion available against which to compare the results of the modified PC algorithm—for this kind of data, on this problem—there is very little more that can be said by way of objective analysis than this.

The domain expert experiment also constitutes a counterexample to Freedman and Humphreys' position that machine algorithms are inferior to analysis by domain experts for complex and nuanced causal discovery problems taken from normal science. The number of variables involved in the experiments of Chapters 4 and 5 is large, and the data sets contain a sizable number of records. Because of this and because of the nature of the variables involved, the problem is extraordinarily difficult to solve from a causal point of view. The aspect of difficulty is underscored by the amount of time required in Chapter 4 by the domain expert to analyze spectral graphs. The aspect of nuance is underscored by the attitude generally taken by domain experts who give explanations of how to read spectral graphs. Finally, the aspect of grounding of the mineral composition from spectra problem in normal science is underscored by the fact that spectra are measured and spectral graphs interpreted on a regular basis in geology and other disciplines. Our own interest in the problem was driven by a research project through NASA Ames, the goal of which is to develop new methods for data collection and analysis for future Mars Rover missions. By all accounts, this is exactly the kind of problem which Freedman and Humphreys claim cannot be solved effectively through machine methods and which require instead the intervention of domain expertise. And yet the results of Chapters 4 and 5 show that a fairly general machine learning algorithm was capable of solving the problem as well as was a domain expert.

6.2 Prospects

This work is currently being extended in a number of directions, several of which are directly related to the issue of mixtures. As the hull-differenced spectra of carbonates from the JPL library in Appendix A show, certain regions of the spectrum from 0.4 μm to 2.5 μm are more indicative of whether a mineral is a carbonate than other regions. For example, the region $[2.0\mu m, 2.5\mu m]$ is much more informative than the region $[0.4\mu m, 0.8\mu m]$.

One research project has attempted to find the best subset regions (Moody, Silva, and Vanderwaart 2000) of the interval $[0.4\mu m, 2.5\mu m]$ over which the modified PC algorithm best predicts mineral class components of rocks. Two approaches were tested. The first divides up the spectrum in the range $[0.4\mu m, 2.5\mu m]$ into regular subintervals and calculates the entropy of spectra within these subranges for particular types of spectra (carbonates, inosilicates, oxides, and phyllosilicates). The second uses a genetic algorithm to find subintervals and unions of subintervals which are especially helpful for predicting carbonate content of rocks. These algorithms did not improve significantly over the expert results for carbonates reported in Chapter 4; however, there was a marked improvement of modified PC results when using optimal subintervals over using the entire $[0.4\mu m, 2.5\mu m]$ range.

Plans are in the works to extend the algorithmic analysis in Chapter 5 into the infrared range ($[3.0\mu m, 30\mu m]$). For many types of minerals, the infrared range is much more diagnostic than the visual to near infrared range ($[0.4\mu m, 3.0\mu m]$). In silicates, for instance, the strongest spectral features are between $8\mu m$ and $12\mu m$,[6] well into the infrared range. These stronger features allow silicates to be distinguished more easily from nonsilicates and from one another. A substantial amount of spectral data is available for rocks and minerals in the infrared range, and more data is in the process of being assembled.

Finally, there are obvious ways to extend the modified PC algorithm (and other algorithms discussed in Chapter 5) to other spectral techniques—e.g., gamma ray spectroscopy, Raman spectroscopy, etc. The modified PC algorithm, especially, is a general technique with very few special assumptions built in about the type of spectra being classified.[7] There is every reason to believe that the modified PC algorithm and other similarly general algorithms would work equally well for other kinds of spectral data.

[6] See Salisbury, Walter, Vergo, and D'Aria (1991, p. xiv).

[7] The measure of conditional independence—conditional correlation—assumes that the data can be modeled linearly, but other measures of conditional independence can be constructed.

Appendix A

JPL Mineral Library

A.1 Introduction

This appendix contains ancillary information about the Jet Propulsion Laboratories (JPL) Mineral Spectral Library. The library itself is published in Grove, Hook, and II (1992). As explained in the text, each data file in the library is measured at one, two, or three different grain sizes. The options are $0\mu m - 45\mu m$, $45\mu m - 125\mu m$, and $125\mu m - 500\mu m$ ("small", "medium", and "large"). There are 160 minerals in the library as a whole, and out of these, 135 are measured at the large grain size. These 135 large grain spectra are used as the background library for relevant experiments in Chapters 5.

The material in the following section provides background information for preparation of the JPL Mineral Library. Table A.1 lists the mineral categories used in the JPL library along with the number of minerals represented in each category. Table A.2 gives a categorized list of the minerals in the library, together with an indication as to whether each mineral was measured at the large grain size. The filenames given are the filenames used in the library; the mineral names are the mineral names included in the data files. Figure A.1 gives a set of hull-differenced graphs of the carbonates in the JPL library.

A.2 Preparation

The following technical description of the JPL Spectral Library is excerpted from C.I. Grove, S.J. Hook, and E.D. Paylor II, "Laboratory Reflectance Spectra of 160 Minerals, 0.4 to 2.5 Micrometers":[1]

> 1.0 INTRODUCTION
>
> This JPL Spectral Library includes laboratory reflectance spectra of 160 minerals in digital form. No attempt has been made to investigate the causes of the spectral features observed. Excellent theoretical studies dealing with the causes of these features are available in the open literature. One of the most comprehensive studies was published by Hunt and his coworkers in a series of papers between 1970 and 1979. Since then, several catalogues of reflectance spectra have been published, including Clark *et al.*, 1990; Lang *et al.*, 1990; and Urai *et al.*, 1989.
>
> Data for 135 of the minerals are presented at three different grain sizes: $125-500\mu m$, $45-125\mu m$, and $< 45\mu m$. This study was undertaken to illustrate the effect of particle size on the shape of the mineral spectra. Ancillary information is provided with each mineral spectrum, including the mineral name, mineralogy, supplier, sampling locality, and our designated sample number. Generalized chemical formulae were obtained from Fleischer (1983) or electron microprobe analysis, when available. The purity of each mineral sample was evaluated by X-ray diffraction (XRD), and identifiable accessory minerals, if present, are noted with the ancillary information.
>
> In the original publication the spectra were separated into classes according to the dominant anion or anionic group present, which is the classification scheme traditionally used in mineralogy. This format has been followed with the organization of the ftp site and each mineral class is a separate subdirectory. Classes include arsenates, borates, carbonates, elements, halides,

[1] The report refers to the measurements made with the Beckman spectrometer. The same sample preparation and measurement procedure was used with the Nicolet spectrometer.

hydroxides, oxides, phosphates, silicates, sulphates, sulphides and tungstates. The silicate class has been subdivided further into subclasses based on the degree of polymerization of the silicon tetrahedra. These subclasses include cyclosilicates, inosilicates, nesosilicates, phyllosilicates, sorosilicates and tectosilicates.

2.0 METHODOLOGY

2.1 Sample Origin

The majority of non-clay minerals used in this report were obtained from Ward's Natural Science Establishment, Rochester, New York; the Burnham Mineral Company (Burminco), Monrovia, California; or from an in-house collection. Most of the clay minerals were obtained from the Source Clay Mineral Repository, University of Missouri, Columbia, Missouri. In all cases, the supplier of the sample is listed in the ancillary information describing each sample.

It was not possible to obtain sufficiently large quantities of all natural minerals, particularly oxides and some clays. In these cases, synthetic minerals were used. In all cases, XRD data for the synthetic mineral agreed with the data reported for the natural mineral in the Mineral Powder Diffraction File Data Book (Joint Committee Powder Diffraction Standards, 1980).

2.2 Sample Preparation

Mineral samples were pulverized with a steel percussion mortar. A magnet was used to remove metallic impurities introduced during this procedure. The pulverized sample was then ground with mortar and pestle and separated into different size fractions by wet-sieving with distilled water or 2-propanol (for water-soluble minerals) in nested sieves. The size fractions selected for reflectance measurements were $125 - 500\mu m$, $45 - 125\mu m$, and $< 45\mu m$, which correspond to fine-to-medium sand, coarse silt to very fine sand, and medium silt to clay, respectively. In certain instances, only one grain size was analyzed due to the nature and/or paucity of the sample (e.g., cristobalite, clay minerals).

2.3 X-ray Diffraction

The purity of each mineral sample was evaluated by using standard XRD methods described in Klug and Alexander (1954). A Norelco water-cooled X-ray diffractometer, equipped with a vertical scan goniometer and focusing monochromater was used for the analysis. All samples were analyzed by Ni-filtered, CuK radiation. Diffraction lines were recorded on a strip chart recorder at a scan rate of 1° (2) per minute over the angular range of 4° to 65° (2).

The Mineral Powder Diffraction File Search Manual and Data Book (Joint Committee on Powder Diffraction Standards, 1980) was used to identify crystalline phases. Additional XRD data were obtained from Borg and Smith (1969) and Berry and Thompson (1962). Identification of clay mineral phases was facilitated by techniques and diffraction data presented by Carroll (1970) and Brindley and Brown (1980).

The XRD criteria used to determine the purity of our samples were based on the number and intensity of diagnostic peaks. If impurities were identified in a sample, a semiquantitative estimate of their abundance was made by XRD. Several limitations of XRD analysis are applicable to our results. Many crystalline substances are strong diffractors of X-rays and can be detected when present in concentrations as small as 1-2 percent. Other materials diffract X-rays less efficiently and yield diffraction patterns of measurable intensity only when they constitute a major portion of the sample. As a result, certain minor constituents cannot be identified by XRD alone. For example, many of the feldspars have suffered minor incipient alteration, which manifests as small features in the Beckman spectra (for example, the sharp feature around $2.2\mu m$ in Sanidine TS-14A). No alteration products were identified by XRD in the Sanidine sample. The problems associated with the detection limits of XRD analyses are discussed in detail in Klug and Alexander (1954).

2.4 Electron Microprobe

Chemical composition data were acquired by electron microprobe analysis for some of our minerals known to deviate significantly from idealized end-member compositions. These analyses were undertaken with the Cameca CAMEBAX electron microprobe at U.C.L.A. Chemical compositions were obtained from polished grain mounts in the wavelength dispersive mode by using 15 keV accelerating potential, 1.5 nAo absorbed current and 20s counting intervals. Compositionally well-characterized silicate minerals from the U.C.L.A. collection were employed as standards. Raw data were reduced according to the ZAF correction scheme (Henog *et al.*, 1982). Results for major elements are believed to be accurate to ±1%. ...

2.5 Spectrophotometer

Hemispherical reflectance measurements in the visible and short-wavelength region of the electromagnetic spectrum (0.4 to $2.5\mu m$) were made by using a Beckman UV5240 spectrophotometer (Price, 1977). The Beckman UV5240 incorporates a single-pass monochromator and utilizes a diffraction grating as its dispersing element. The sampling interval is $.001\mu m$, from 0.4 to $0.8\mu m$, and $.004\mu m$, from 0.8 to $2.5\mu m$. Our instrument has been modified with an integrating sphere rotated 90 degrees, which facilitates measurement of powdered samples and soils by allowing the materials to remain in the sample holder in a horizontal position.

For reflectance measurements, the size-sorted samples were poured into aluminum sample holders that measured 3.2 cm in diameter and 0.5 cm in depth. The upper surface was carefully smoothed with the edge of a stainless steel spatula. Smoothing with a spatula may compact the sample and introduce some preferred orientation of grains. However, every effort was made to minimize these effects. The sample was then placed in the sample compartment where it and a Halon reference standard were illuminated alternately by monochromatic radiation from a high-intensity halogen source lamp. Halon, a trade name for polytetrafluoroethylene powder, has been shown to be a good diffuser of incident radiation over the spectral range of the 0.2 - 2.5 m region (Weidner and Hsia, 1981). However, Halon does have a small absorption feature near $2.2\mu m$. This feature is manifest in spectra with a high reflectance in the $2.0 - 2.5\mu m$ region. In order to correct this Halon artifact, the spectra were multiplied by the reflectance of Halon vs a perfect diffuser given in Weidner and Hsia (1981). The correction largely removed the influence of the Halon absorption feature in our spectra.

Table A.1: The 17 mineral class used in the Jet Propulsion Labs Mineral Spectral Library together with the number of minerals in each class at the largest grain size.

JPL Mineral Class	Count
Arsenates	2
Borates	6
Carbonates	15
Cyclosilicates	4
Elements	2
Halides	5
Inosilicates	12
Nesosilicates	7
Hydroxides	1
Oxides	9
Phosphates	4
Phyllosilicates	20
Sulphides	11
Sulphates	13
Sorosilicates	5
Tectosilicates	18
Tungstates	1
	135

Table A.2. List of rocks in the JPL library. The filename of each sample is given together with the name of the mineral and the mineral class to which the mineral belongs. There are 17 mineral classes altogether.

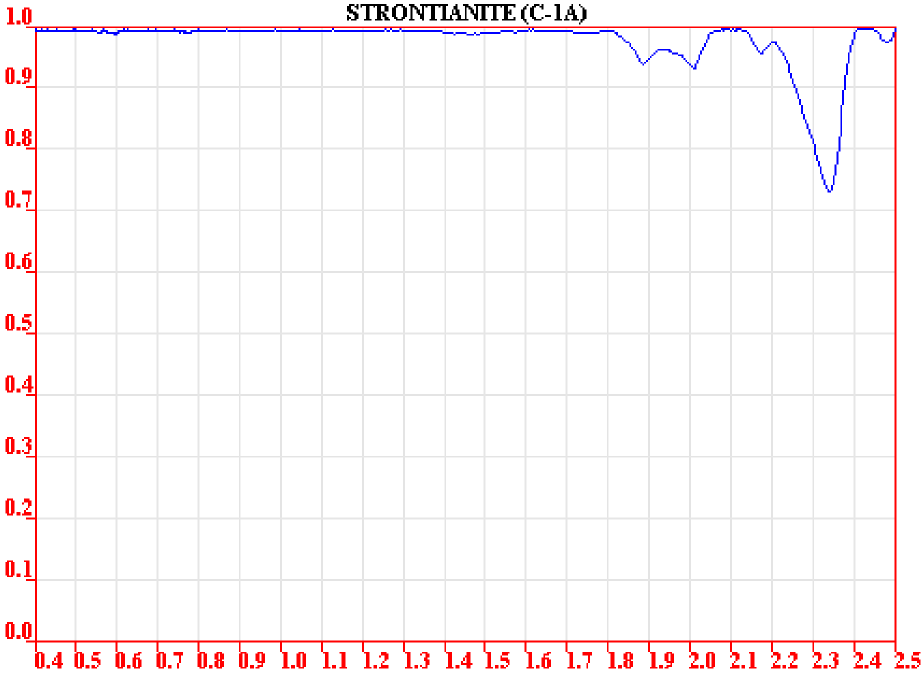
Index	Filename	Mineral Name	Mineral Class
1	a01a	Mimetite	Arsenate
2	a02a	Scorodite	Arsenate
3	b01a	Colemanite	Borate
4	b02a	Kernite	Borate
5	b03a	Ulexite	Borate
6	b04a	Tincalconite	Borate
7	b05a	Howlite	Borate
8	b06a	Borax	Borate
9	c01a	Strontianite	Carbonate
10	c02a	Witherite	Carbonate
11	c03a	Calcite	Carbonate
12	c03d	Calcite	Carbonate
13	c03e	Calcite	Carbonate
14	c04a	Trona	Carbonate
15	c05a	Dolomite	Carbonate
16	c05c	Dolomite	Carbonate
17	c06a	Magnesite	Carbonate
18	c07a	Malachite	Carbonate
19	c08a	Rhodochrosite	Carbonate
20	c09a	Siderite	Carbonate
21	c10a	Cerussite	Carbonate
22	c11a	Smithsonite	Carbonate
23	c12a	Azurite	Carbonate
24	cs01a	Tourmaline, Dravite-S	Cyclosilicate
25	cs02a	Beryl	Cyclosilicate
26	cs03a	Cordierite	Cyclosilicate
27	cs04a	Ferroaxinite	Cyclosilicate
28	e01a	Graphite	Element
29	e02a	Sulfur	Element
30	h01a	Cryolite	Halide
31	h02a	Fluorite, Purple	Halide
32	h02b	Fluorite	Halide
33	h03a	Halite	Halide
34	h04a	Atacamite	Halide
35	in01a	Rhodonite	Inosilicate
36	in02a	Wollastoniate	Inosilicate
37	in03a	Glaucophane	Inosilicate
38	in04a	Actinolite	Inosilicate
39	in05a	Tremolite	Inosilicate

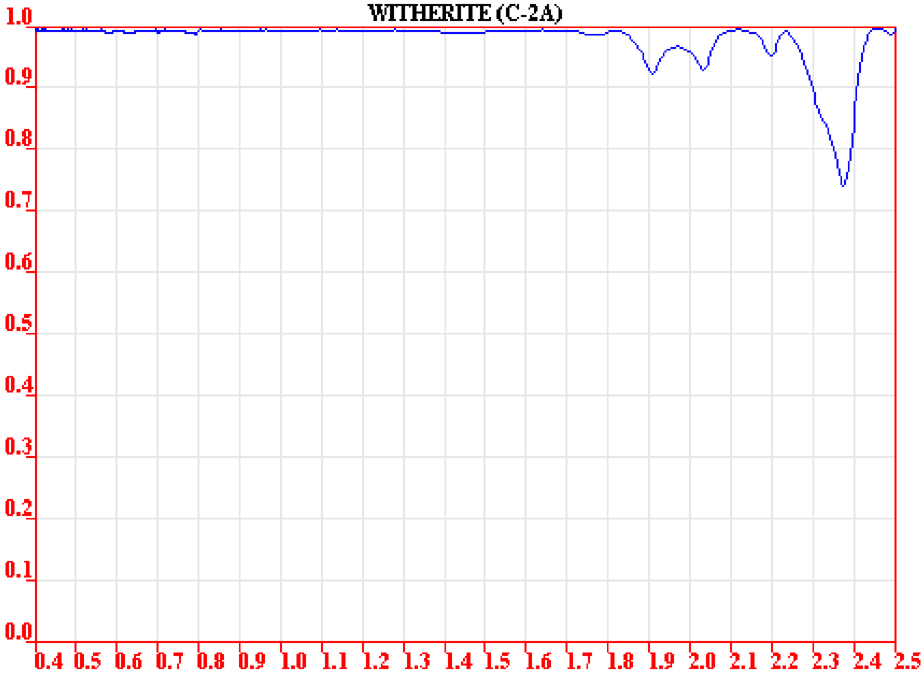
Index	Filename	Mineral Name	Mineral Class
40	in06a	Cummingtonite	Inosilicate
41	in07a	Riebeckite	Inosilicate
42	in08a	Anthophyllite	Inosilicate
43	in09a	Diopside	Inosilicate
44	in10a	Enstatite	Inosilicate
45	in12a	Johannsenite	Inosilicate
46	in13a	Spodumene	Inosilicate
47	in14a	Hypersthene	Inosilicate
48	in15a	Augite	Inosilicate
49	ns01a	Fayalite	Nesosilicate
50	ns02a	Forsterite, Synthetic	Nesosilicate
51	ns03b	Grossular Garnet	Nesosilicate
52	ns04a	Almandine Garnet	Nesosilicate
53	ns06a	Topaz	Nesosilicate
54	ns07a	Titanite	Nesosilicate
55	ns08a	Sillimanite	Nesosilicate
56	ns09a	Zircon	Nesosilicate
57	o01a	Hematite	Oxide
58	o01b	Hematite, Synthetic	Oxide
59	o02a	Rutile	Oxide
60	o03a	Cassiterite	Oxide
61	o04a	Magnetite	Oxide
62	o06a	Pyrolusite	Oxide
63	o07a	Columbite	Oxide
64	o08a	Magnesiochromite	Oxide
65	o11a	Gahnite	Oxide
66	o12a	Anatase, Synthetic	Oxide
67	o13a	Zincite, Synthetic	Oxide
68	o14a	(Unnamed)	Oxide
69	o15a	Corundum, Synthetic	Oxide
70	oh01a	Brucite	Hydroxide
71	oh02a	Goethite	Hydroxide
72	oh03a	Gibbsite	Hydroxide
73	p01a	Apatite	Phosphate
74	p02a	Montebrasite	Phosphate
75	p03a	Amblygonite	Phosphate
76	p04a	Triphylite	Phosphate
77	ps01a	Kaolinite, Well Ordered	Phyllosilicate
78	ps01b	Kaolinite, Disordered	Phyllosilicate
79	ps02b	Montmorillonite, Cal	Phyllosilicate
80	ps02d	Montmorillonite, Sod	Phyllosilicate
81	ps03a	Dickite	Phyllosilicate

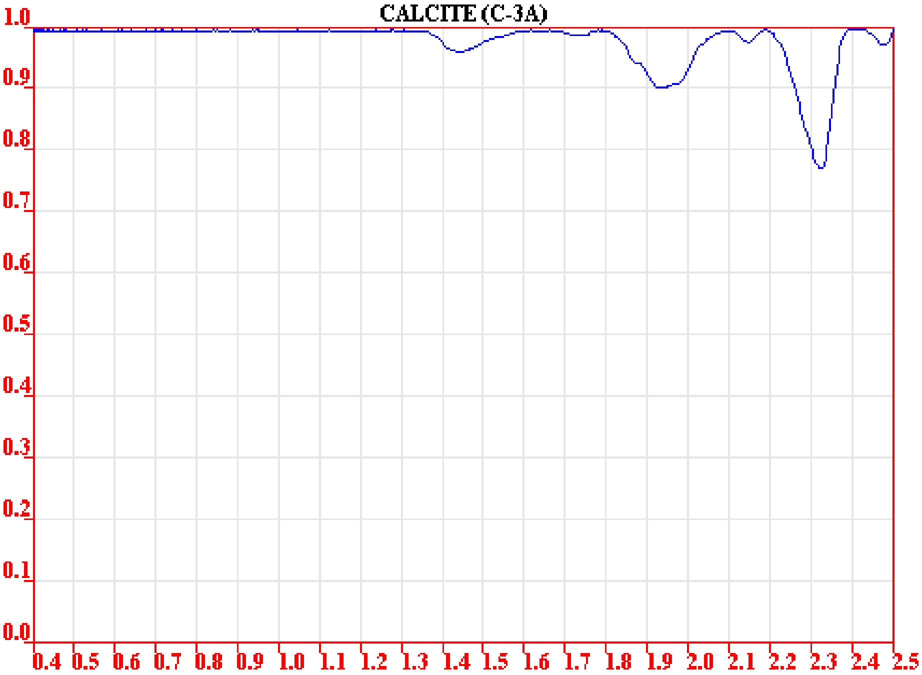
Index	Filename	Mineral Name	Mineral Class
82	ps04a	Palygorskite	Phyllosilicate
83	ps05a	Sepiolite	Phyllosilicate
84	ps06a	Nontronite	Phyllosilicate
85	ps06b	Nontronite	Phyllosilicate
86	ps06d	Nontronite	Phyllosilicate
87	ps07a	Pyrophyllite	Phyllosilicate
88	ps09a	Cookeite	Phyllosilicate
89	ps10a	Corrensite	Phyllosilicate
90	ps11a	Illite	Phyllosilicate
91	ps12a	Chlorite (Ripidolite)	Phyllosilicate
92	ps12c	Chlorite	Phyllosilicate
93	ps12e	Chlorite (Pyrochlorite)	Phyllosilicate
94	ps12f	Chlorite (Thuringite)	Phyllosilicate
95	ps13a	Lepidolite, Yellow	Phyllosilicate
96	ps13b	Lepidolite, Lavender	Phyllosilicate
97	ps14a	Talc	Phyllosilicate
98	ps16a	Muscovite	Phyllosilicate
99	ps18a	Vermiculite	Phyllosilicate
100	ps18b	Vermuculite	Phyllosilicate
101	ps19a	Glauconite	Phyllosilicate
102	ps20a	Serpentine	Phyllosilicate
103	ps21a	Prehnite	Phyllosilicate
104	ps22a	Hydroxyapophyllite	Phyllosilicate
105	ps23a	Biotite	Phyllosilicate
106	ps24a	Saponite	Phyllosilicate
107	s01a	Sphalerite	Sulfide
108	s01a	Pyrite	Sulfide
109	s03a	Realgar	Sulfide
110	s04a	Chalcopyrite	Sulfide
111	s05a	Arsenopyrite	Sulfide
112	s06a	Stibnite	Sulfide
113	s07a	Galena	Sulfide
114	s08a	Chalcocite	Sulfide
115	s09a	Bornite	Sulfide
116	s10a	Marcasite	Sulfide
117	s11a	Molybdenite	Sulfide
118	s12a	Pyrrhotite	Sulfide
119	so01a	Anhydrite	Sulfate
120	so02b	Gypsum	Sulfate
121	so03a	Barite	Sulfate
122	so04a	Alunite	Sulfate
123	so05a	Celestite	Sulfate

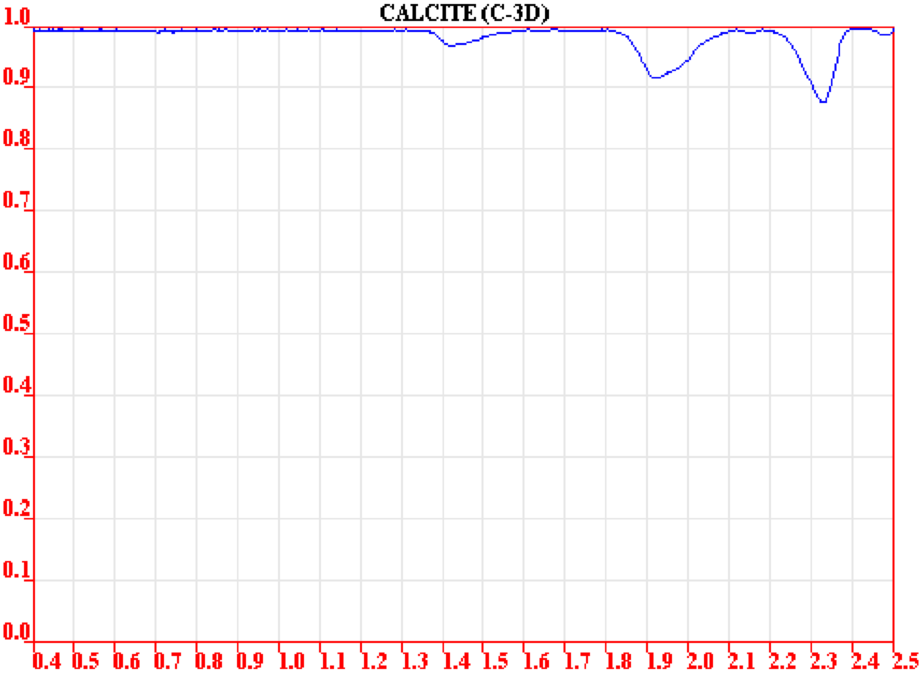
Index	Filename	Mineral Name	Mineral Class
124	so06a	Tschermigite	Sulfate
125	so07a	Jarosite	Sulfate
126	so07b	Plumbojarosite	Sulfate
127	so07c	Natrojarosite	Sulfate
128	so08a	Glauberite	Sulfate
129	so09a	Aphthitalite	Sulfate
130	so10a	Anglesite	Sulfate
131	so11a	Antlerite	Sulfate
132	ss01a	Epidote	Sorosilicate
133	ss01c	Epidote	Sorosilicate
134	ss02a	Hemimorphite	Sorosilicate
135	ss03a	Vesuvianite	Sorosilicate
136	ss04a	Clinozoisite	Sorosilicate
137	t01a	Scheelite	Tunstate
138	ts01a	Quartz, Rock Crystal	Phyllosilicate
139	ts01b	Quartz, Smoky	Phyllosilicate
140	ts01c	Quartz, Rose	Phyllosilicate
141	ts01d	Quartz, Milky	Phyllosilicate
142	ts01e	Quartz, Chrosoprase	Phyllosilicate
143	ts02a	Labradorite	Phyllosilicate
144	ts02b	Labradorite	Phyllosilicate
145	ts03a	Oligoclase	Phyllosilicate
146	ts04a	Andesine	Phyllosilicate
147	ts05a	Anorthite	Phyllosilicate
148	ts06a	Albite	Phyllosilicate
149	ts07a	Cristobalite	Phyllosilicate
150	ts08a	Natrolite	Phyllosilicate
151	ts09a	Stilbite	Phyllosilicate
152	ts10a	Sodalite	Phyllosilicate
153	ts11a	Buddingtonite, Feldspar	Phyllosilicate
154	ts12a	Orthoclase	Phyllosilicate
155	ts13a	Bytownite	Phyllosilicate
156	ts14a	Sanidine	Phyllosilicate
157	ts15a	Chabazite	Phyllosilicate
158	ts16a	Nepheline	Phyllosilicate
159	ts17a	Microcline	Phyllosilicate
160	ts18a	Analcime	Phyllosilicate

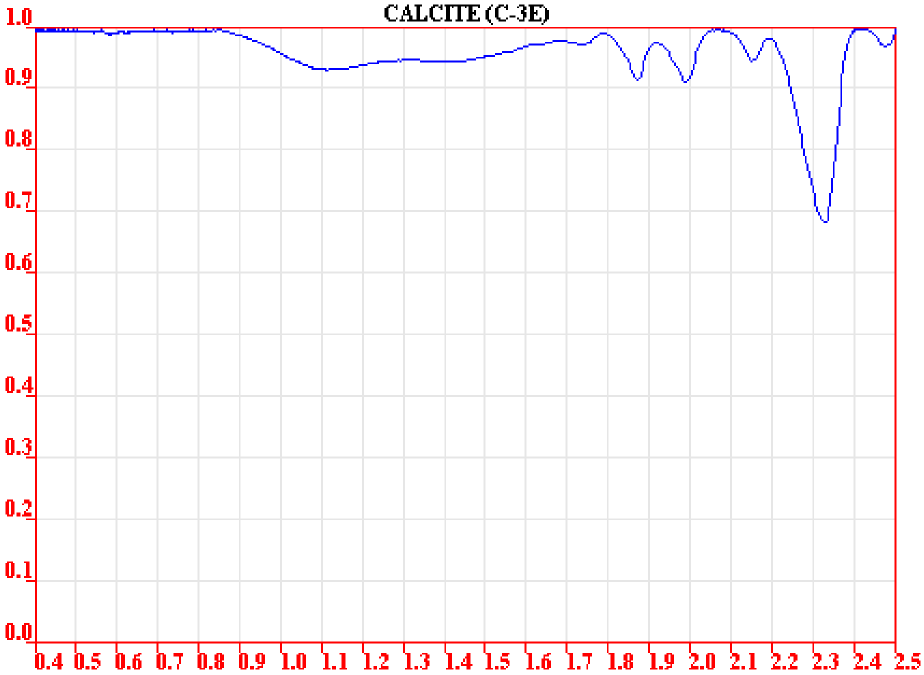
Figure A.1: Hull-difference graphs of 15 JPL Mineral Library carbonates.

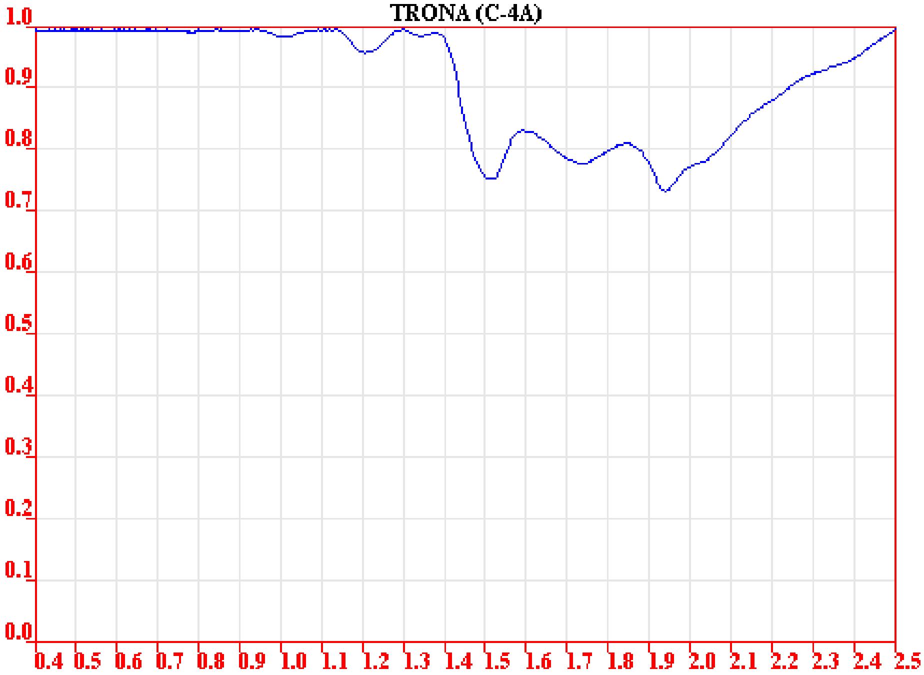


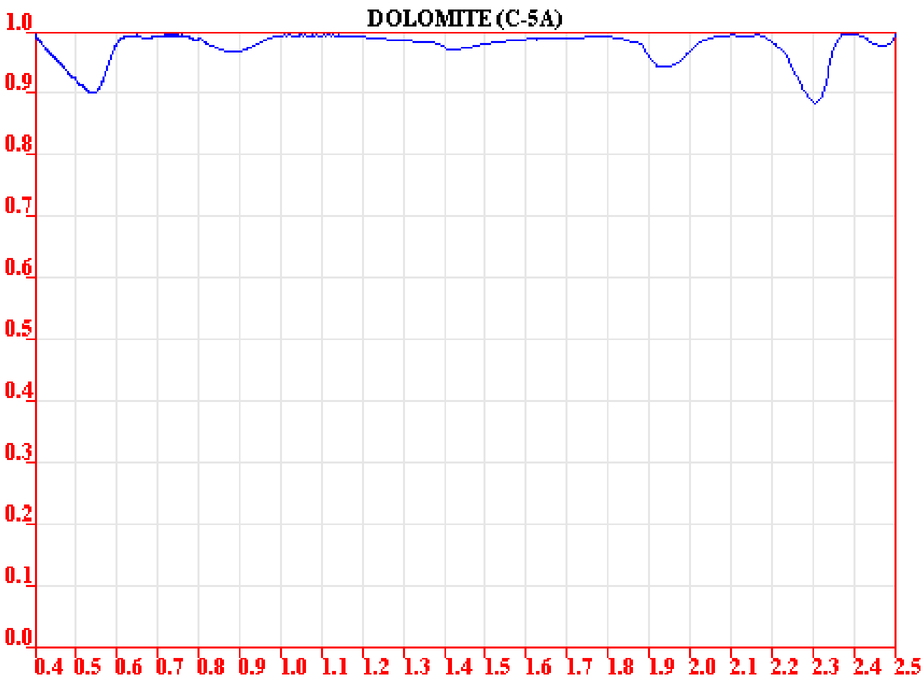


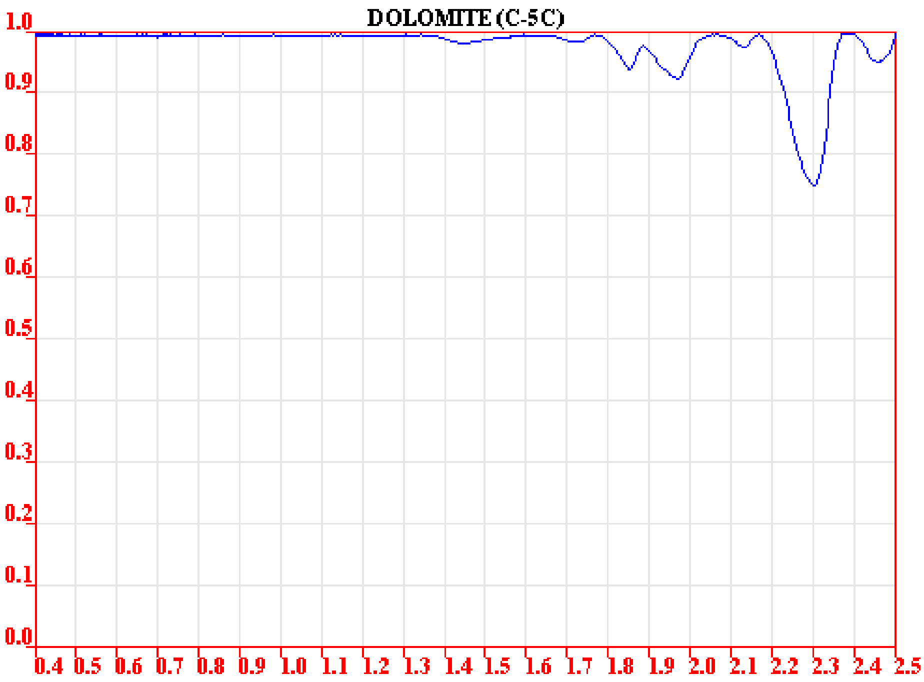

CALCITE (C-3A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

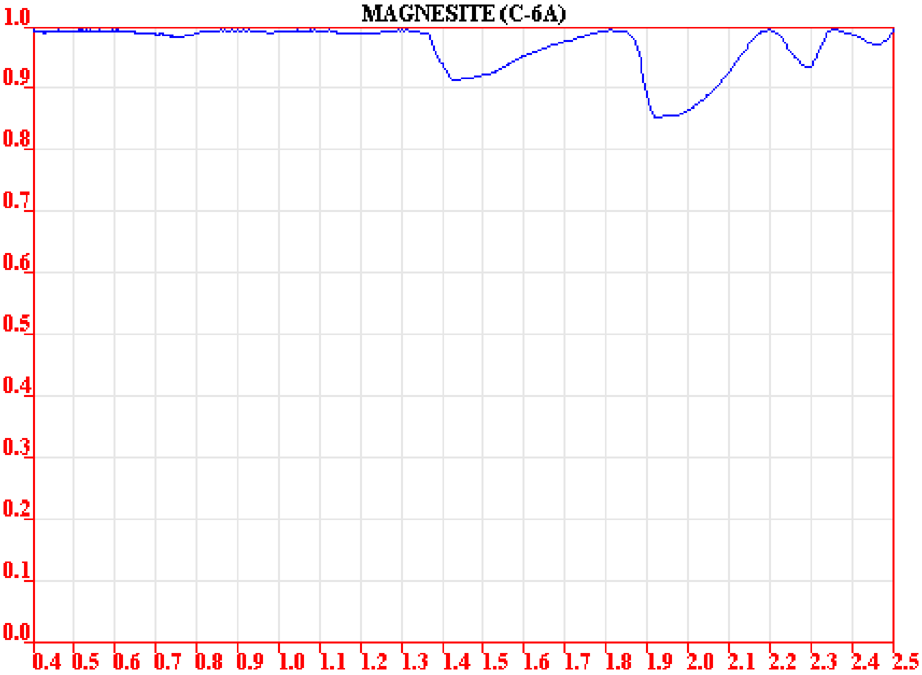

CALCITE (C-3D)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

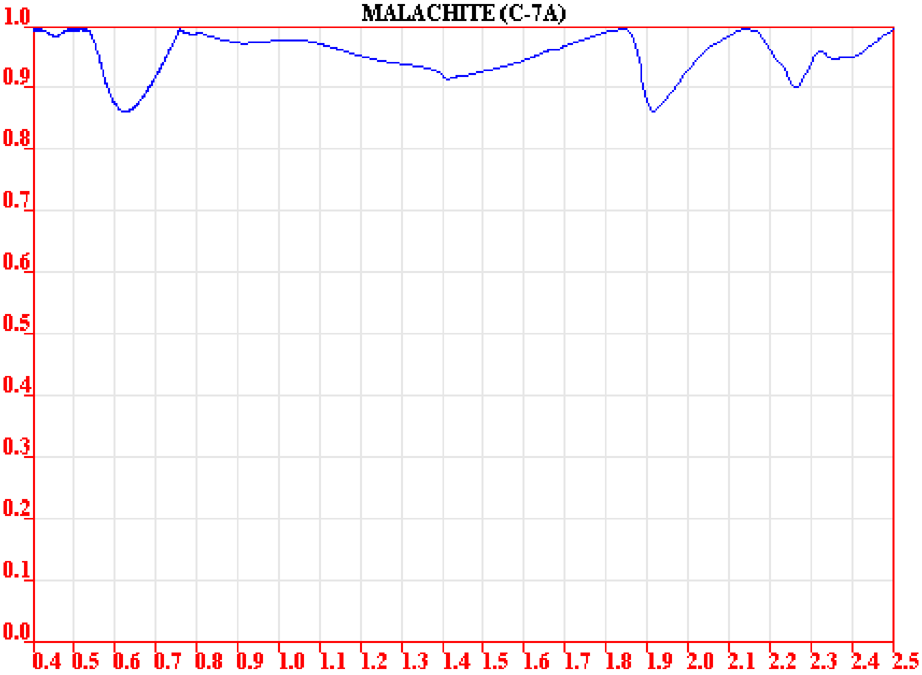

CALCITE (C-3E)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

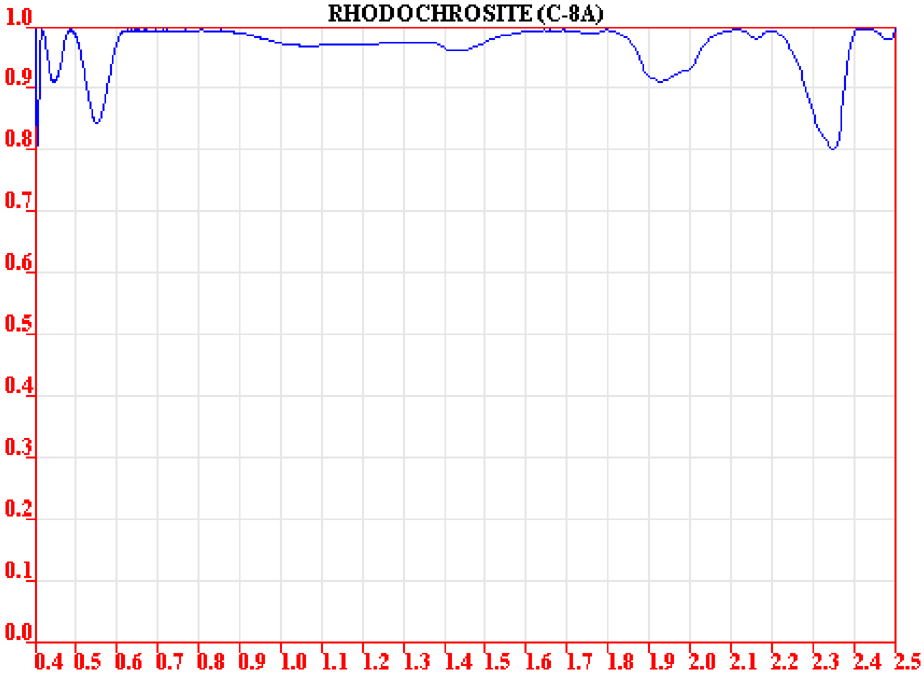

TRONA (C-4A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

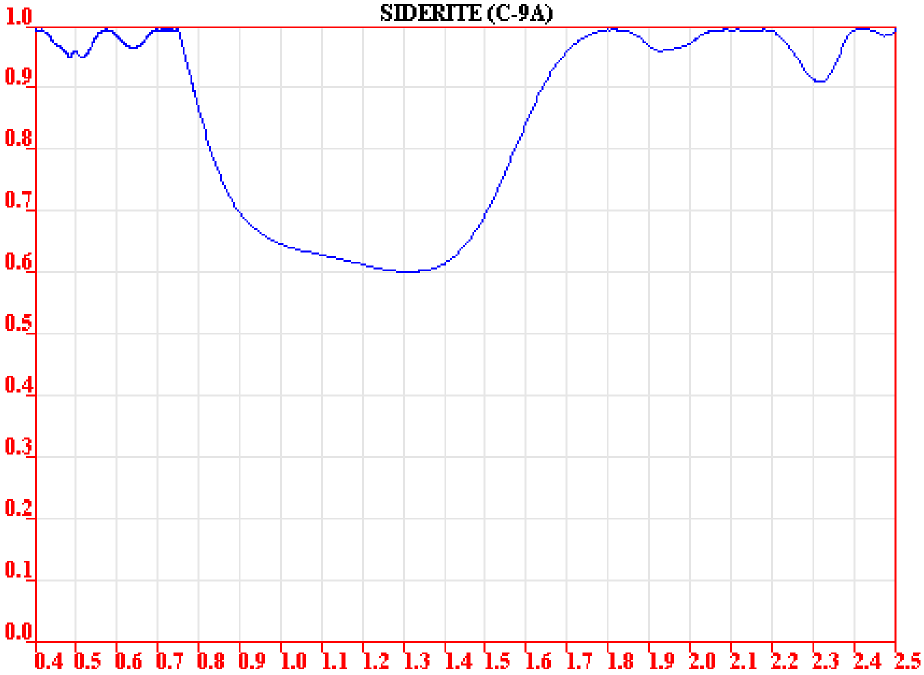

DOLOMITE (C-5A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

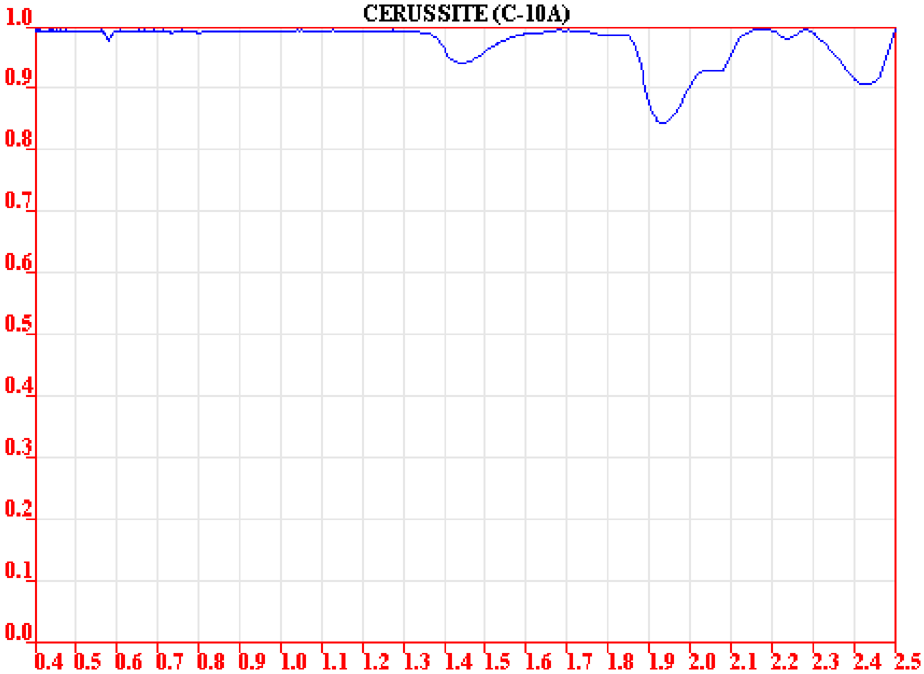

DOLOMITE (C-5C)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

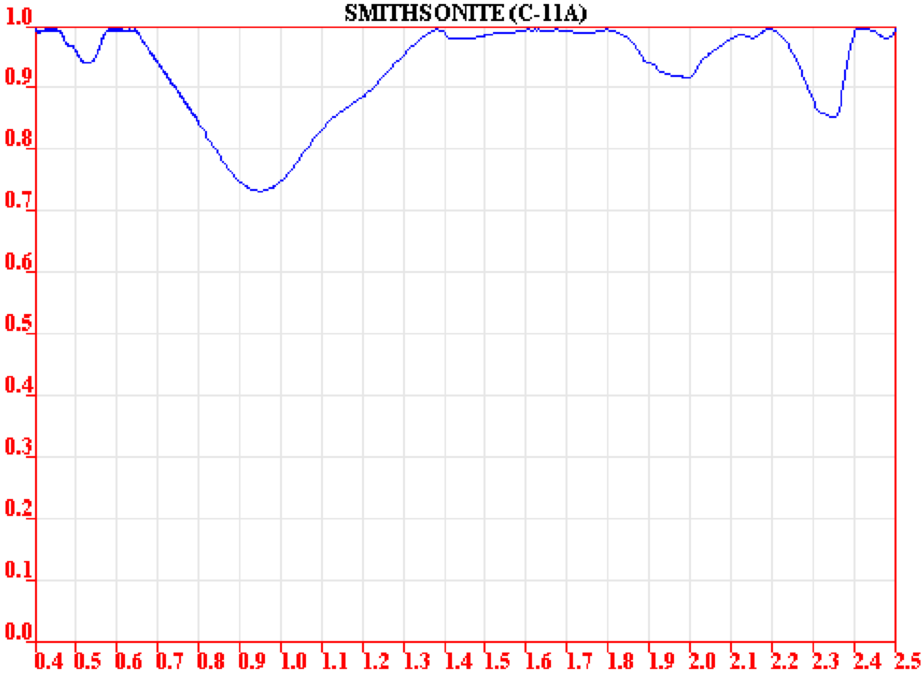

MAGNESITE (C-6A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

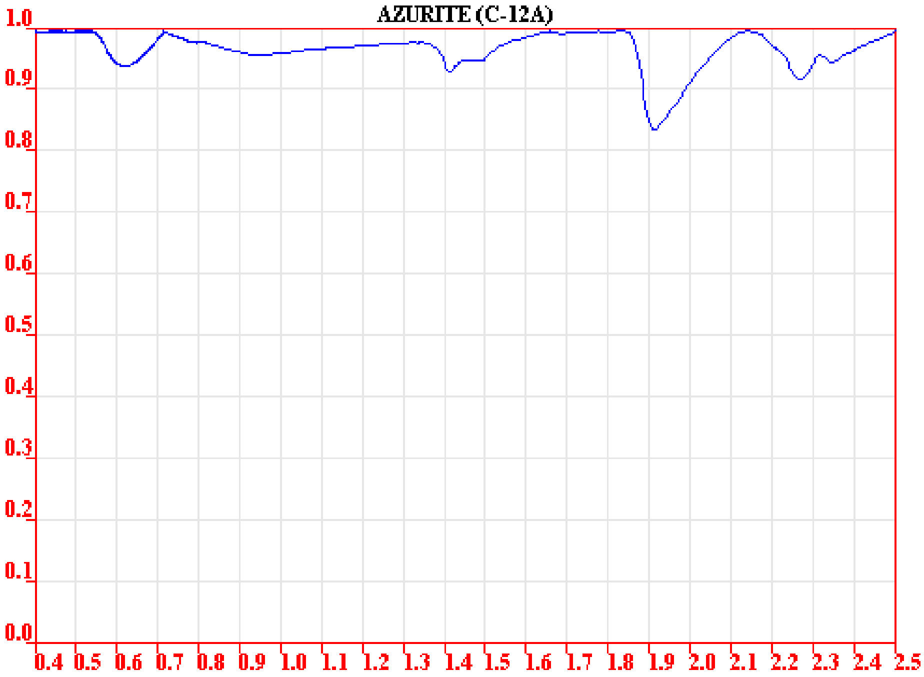

MALACHITE (C-7A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5


RHODOCHROSITE (C-8A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5


SIDERITE (C-9A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5


CERUSSITE (C-10A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5


SMITHSONITE (C-11A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5


AZURITE (C-12A)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5

Appendix B

JHU Mineral Library

B.1 Introduction

This appendix contains a list of rock samples from the Johns Hopkins University (JHU) Spectral Library. This library is currently available as part of the Aster Spectral Library at http://speclib.jpl.nasa.gov; the rock names for the spectra used in Chapters 4 and 5 are listed below in Table B.2. We will refer to the spectra for the 192 rocks in this list as the "JHU Rock Library."

Six types of samples are represented in the JHU Rock Library, as shown in Table B.1; there are three general categories of rocks (igneous, metamorphic, and sedimentary), and each of these is presented in two forms. Igneous rocks are presented in either finely powdered form or solid form; metamorphic and sedimentary rocks are presented in either coarsely or finely powdered form. Table B.2 gives a list of all of the rock samples in the library (all 192 of them), together with their type and form. An additional column is included in the table to show the gold standard classification of these rocks as carbonates

Rock Type	**Sample Type**	**# Samples**
Igneous	Solid	35
Igneous	Coarsely Powdered	35
Metamorphic	Coarsely Powdered	37
Metamorphic	Finely Powdered	37
Sedimentary	Coarsely Powdered	24
Sedimentary	Finely Powdered	24
	Total	192

Table B.1: Types of spectra contained in the Johns Hopkins University Spectral Library.

or not carbonates.

B.2 Preparation

As is noted in the technical description included with the Johns Hopkins University Spectral Library, all of the rock samples contained in the library "were measured under the direction of John W. (Jack) Salisbury," with most measurements being made "by Dana M. D'Aria, either at Johns Hopkins University in Baltimore, MD, or at the U.S. Geological Survey in Reston, VA." The following is an excerpt from this technical description relevant to the JHU Rock Library:

> MEASUREMENT TECHNIQUE
>
> ...The apparently seamless reflectance spectra from 0.4 to 14 micrometers of rocks and soils were generated using two different instruments, both equipped with integrating spheres for measurement of directional hemispherical reflectance, with source radiation impinging on the sample from a centerline angle 10 degrees from the vertical.
>
> Unless specified otherwise (see relevant introductory texts for generic snow and vegetation spectra, and spectra of man-made materials), all visible/near-infrared (VNIR) spectra were recorded using a Beckman Instruments model UV 5240 dual-beam, grating spectrophotometer at the U.S. Geological Survey, Reston, VA. The data were obtained digitally and corrected for both instrument function and the reflectance of the Halon reference using standards traceable to the U. S. National Institute of Science and Technology.

Measurements of such standards indicate an absolute reflectance accuracy of plus or minus 3 percent. Wavelength accuracy was checked using a holmium oxide reference filter and is reproducible and accurate to within plus or minus 0.004 micrometers, or 4 nm (one digitization step). Spectral resolution is variable because the Beckman uses an automatic slit program to keep the energy on the detector constant. The result is a spectral bandwidth typically less than 0.008 micrometers over the 0.4 to 2.5 micrometers spectral range measured, but slightly larger at the two extremes of the range of the lead sulfide detector (0.8-0.9 micrometers and 2.4- 2.5 micrometers). This instrument has a grating change at 0.8 micrometers, which sometimes results in a spectral artifact (either a small, sharp absorption band, or a slight offset of the spectral curve) at that wavelength.

Two similar instruments were used to record reflectance in the infrared range (2.08 to 15 micrometers). Briefly, both are Nicolet FTIR spectrophotometers and both have a reproducibility and absolute accuracy better than plus or minus 1 percent over most of the spectral range. Early measurements of igneous rocks with an older detector were noisy in the 13.5-14 micrometers range and do not quite meet this standard in that region. Because FTIR instruments record spectral data in frequency space, both wavelength accuracy and spectral resolution are given in wave numbers (reciprocal centimeters). Wavelength accuracy of an interferometer type of instrument is limited by the spectral resolution, which yields a data point every 2 wave numbers for these measurements. The X-axis was changed from wave numbers to micrometers for all of these data before the infrared segment was joined to the VNIR data from the Beckman.

Spectra from the Beckman and the FTIR instruments were compared in the overlap range of 2.08-2.5 micrometers. If the difference was greater than 3 percent, measurements were repeated. Typically, however, the agreement was within the 3 percent limit. In view of the greater accuracy of the FTIR measurements, any small discrepancy between the two spectral segments was resolved by adjusting the Beckman data to fit the reflectance level of the segment measured by the FTIR instruments.

Table B.2. List of rocks in the JHU library. The final column shows the gold standard classification for the class 'carbonate,' obtained by examining the file descriptors of each sample. For this column, "A" = absent, "P" = present, "PP" = possibly present, and "N" = not classified.

	Filename	Mineral Name	Type	Form	Carbonate
1	andesi1f	Augite-hypersthene Andesite	Igneous	Fine	A
2	andesi1s	Augite-hypersthene Andesite	Igneous	Solid	A
3	andesi2f	Augite-hypersthene Andesite	Igneous	Fine	PP
4	andesi2s	Augite-hypersthene Andesite	Igneous	Solid	PP
5	andesi4f	Basaltic Andesite	Igneous	Fine	A
6	andesi4s	Basaltic Andesite	Igneous	Solid	A
7	anorth1f	Anorthosite	Igneous	Fine	PP
8	anorth1s	Anorthosite	Igneous	Solid	PP
9	aplite1f	Aplite	Igneous	Fine	PP
10	aplite1s	Aplite	Igneous	Solid	PP
11	basal10f	Basalt	Igneous	Fine	A
12	basal10s	Basalt	Igneous	Solid	A
13	basal1f	Basalt	Igneous	Fine	A
14	basal1s	Basalt	Igneous	Solid	A
15	basal2f	Basalt	Igneous	Fine	A
16	basal2s	Basalt	Igneous	Solid	A
17	basal5f	Basalt	Igneous	Fine	A
18	basal5s	Basalt	Igneous	Solid	A
19	basal7f	Basalt	Igneous	Fine	A
20	basal7s	Basalt	Igneous	Solid	A
21	basal9f	Basalt	Igneous	Fine	A
22	basal9s	Basalt	Igneous	Solid	A
23	diabas1f	Diabase	Igneous	Fine	A
24	diabas1s	Diabase	Igneous	Solid	A
25	diabas2f	Diabase	Igneous	Fine	A
26	diabas2s	Diabase	Igneous	Solid	A
27	diorit1f	Diorite	Igneous	Fine	A
28	diorit1s	Diorite	Igneous	Solid	A
29	dunit1f	Dunite	Igneous	Fine	A
30	dunit1s	Dunite	Igneous	Solid	A
31	gabbro1f	Gabbro	Igneous	Fine	A
32	gabbro1s	Gabbro	Igneous	Solid	A
33	gneiss1c	Chloritic Gneiss	Metamorphic	Coarse	A
34	gneiss1f	Chloritic Gneiss	Metamorphic	Fine	A
35	gneiss2c	Garnet Gneiss	Metamorphic	Coarse	A
36	gneiss2f	Garnet Gneiss	Metamorphic	Fine	A
37	gneiss3c	Felsitic Gneiss	Metamorphic	Coarse	A
38	gneiss3f	Felsitic Gneiss	Metamorphic	Fine	A

	Filename	Mineral Name	Type	Form	Carbonate
39	gneiss4c	Syenite Gneiss	Metamorphic	Coarse	A
40	gneiss4f	Syenite Gneiss	Metamorphic	Fine	A
41	gneiss5c	Albite Gneiss	Metamorphic	Coarse	A
42	gneiss5f	Albite Gneiss	Metamorphic	Fine	A
43	gneiss6c	Hornblende Gneiss	Metamorphic	Coarse	A
44	gneiss6f	Hornblende Gneiss	Metamorphic	Fine	A
45	gneiss7c	Diorite Gneiss	Metamorphic	Coarse	PP
46	gneiss7f	Diorite Gneiss	Metamorphic	Fine	PP
47	gneiss8c	Augen Gneiss	Metamorphic	Coarse	A
48	gneiss8f	Augen Gneiss	Metamorphic	Fine	A
49	granit1f	Alkalic Granite	Igneous	Fine	A
50	granit1s	Alkalic Granite	Igneous	Solid	A
51	granit2f	Granite	Igneous	Fine	PP
52	granit2s	Granite	Igneous	Solid	PP
53	granit3f	Granite	Igneous	Fine	PP
54	granit3s	Granite	Igneous	Solid	PP
55	granit5f	Granite	Igneous	Fine	PP
56	granit5s	Granite	Igneous	Solid	PP
57	granod1f	Granodiorite	Igneous	Fine	A
58	granod1s	Granodiorite	Igneous	Solid	A
59	granod2f	Granodiorite	Igneous	Fine	PP
60	granod2s	Granodiorite	Igneous	Solid	PP
61	greywa1c	Greywacke Sandstone	Sedimentary	Coarse	A
62	greywa1f	Greywacke Sandstone	Sedimentary	Fine	A
63	hornfe1c	Banded Hornfels	Metamorphic	Coarse	P
64	hornfe1f	Banded Hornfels	Metamorphic	Fine	P
65	hornfe2c	Hornfels	Metamorphic	Coarse	PP
66	hornfe2f	Hornfels	Metamorphic	Fine	PP
67	hornfe3c	Spotted Hornfels	Metamorphic	Coarse	A
68	hornfe3f	Spotted Hornfels	Metamorphic	Fine	A
69	ijolit1f	Ijolite	Igneous	Fine	PP
70	ijolit1s	Ijolite	Igneous	Solid	PP
71	lampro1f	Lamprophyre	Igneous	Fine	A
72	lampro1s	Lamprophyre	Igneous	Solid	A
73	limest1c	Fossiliferous Limestone	Sedimentary	Coarse	P
74	limest1f	Fossiliferous Limestone	Sedimentary	Fine	P
75	limest2c	Dolomitic Limestone	Sedimentary	Coarse	P
76	limest2f	Dolomitic Limestone	Sedimentary	Fine	P
77	limest3c	Limestone	Sedimentary	Coarse	P
78	limest3f	Limestone	Sedimentary	Fine	P
79	limest4c	Oolitic Limestone	Sedimentary	Coarse	P
80	limest4f	Oolitic Limestone	Sedimentary	Fine	P

	Filename	Mineral Name	Type	Form	Carbonate
81	limest5c	Lithographic Limestone	Sedimentary	Coarse	P
82	limest5f	Lithographic Limestone	Sedimentary	Fine	P
83	limest6c	Argillaceous Limestone	Sedimentary	Coarse	P
84	limest6f	Argillaceous Limestone	Sedimentary	Fine	P
85	limest7c	Oolitic Limestone	Sedimentary	Coarse	P
86	limest7f	Oolitic Limestone	Sedimentary	Fine	P
87	marble1c	Dolomitic Marble	Metamorphic	Coarse	P
88	marble1f	Dolomitic Marble	Metamorphic	Fine	P
89	marble2c	Serpentine Marble	Metamorphic	Coarse	P
90	marble2f	Serpentine Marble	Metamorphic	Fine	P
91	marble3c	Marble	Metamorphic	Coarse	P
92	marble3f	Marble	Metamorphic	Fine	P
93	marble4c	Dolomitic Marble	Metamorphic	Coarse	P
94	marble4f	Dolomitic Marble	Metamorphic	Fine	P
95	marble5c	Serpentine Marble	Metamorphic	Coarse	P
96	marble5f	Serpentine Marble	Metamorphic	Fine	P
97	marble6c	White Marble	Metamorphic	Coarse	P
98	marble6f	White Marble	Metamorphic	Fine	P
99	marble7c	Pink Marble	Metamorphic	Coarse	P
100	marble7f	Pink Marble	Metamorphic	Fine	P
101	monzon1f	Monzonite	Igneous	Fine	PP
102	monzon1s	Monzonite	Igneous	Solid	PP
103	norite1f	Norite	Igneous	Fine	A
104	norite1s	Norite	Igneous	Solid	A
105	norite2f	Norite	Igneous	Fine	A
106	norite2s	Norite	Igneous	Solid	A
107	obsidi1f	Rhyolitic Obsidian	Igneous	Fine	A
108	obsidi1s	Rhyolitic Obsidian	Igneous	Solid	A
109	philli1c	Phyllite	Metamorphic	Coarse	A
110	phylli1f	Phyllite	Metamorphic	Fine	A
111	picrit1f	Picrite	Igneous	Fine	A
112	picrit1s	Picrite	Igneous	Solid	A
113	picrit2f	Picrite	Igneous	Fine	A
114	picrit2s	Picrite	Igneous	Solid	A
115	qmonzo1f	Quartz Monzonite	Igneous	Fine	PP
116	qmonzo1s	Quartz Monzonite	Igneous	Solid	PP
117	qrtzit1c	Red Quartzite	Metamorphic	Coarse	P
118	qrtzit1f	Red Quartzite	Metamorphic	Fine	P
119	qrtzit2c	Green Quartzite	Metamorphic	Coarse	A
120	qrtzit2f	Green Quartzite	Metamorphic	Fine	A
121	qrtzit3c	Pink Quartzite	Metamorphic	Coarse	A
122	qrtzit3f	Pink Quartzite	Metamorphic	Fine	A

	Filename	Mineral Name	Type	Form	Carbonate
123	qrtzit4c	Purple Quartzite	Metamorphic	Coarse	A
124	qrtzit4f	Purple Quartzite	Metamorphic	Fine	A
125	qrtzit5c	Gray Quartzite	Metamorphic	Coarse	A
126	qrtzit5f	Gray Quartzite	Metamorphic	Fine	A
127	qrtzit6f	Green Quartzite	Metamorphic	Fine	N
128	rhyoli1f	Rhyolite	Igneous	Fine	A
129	rhyoli1s	Rhyolite	Igneous	Solid	A
130	sandst1c	Arkosic Sandstone	Sedimentary	Coarse	P
131	sandst1f	Arkosic Sandstone	Sedimentary	Fine	P
132	sandst2c	Glauconitic Sandstone	Sedimentary	Coarse	P
133	sandst2f	Glauconitic Sandstone	Sedimentary	Fine	P
134	sandst3c	Sandstone (Micacious Red)	Sedimentary	Coarse	P
135	sandst3f	Sandstone (Micacious Red)	Sedimentary	Fine	P
136	sandst4c	Ferruginous Sandstone	Sedimentary	Coarse	P
137	sandst4f	Ferruginous Sandstone	Sedimentary	Fine	P
138	sandst6c	Purple-banded Sandstone	Sedimentary	Coarse	A
139	sandst6f	Purple-banded Sandstone	Sedimentary	Fine	A
140	sandst7c	Sandstone (Red)	Sedimentary	Coarse	A
141	sandst7f	Sandstone (Red)	Sedimentary	Fine	A
142	schis10c	Graphite Schist	Metamorphic	Coarse	P
143	schis10f	Graphite Schist	Metamorphic	Fine	P
144	schist1c	Green Schist	Metamorphic	Coarse	P
145	schist1f	Green Schist	Metamorphic	Fine	P
146	schist2c	Hornblende Schist	Metamorphic	Coarse	A
147	schist2f	Hornblende Schist	Metamorphic	Fine	A
148	schist3c	Mica Schist	Metamorphic	Coarse	A
149	schist3f	Mica Schist	Metamorphic	Fine	A
150	schist4c	Tourmaline Schist	Metamorphic	Coarse	A
151	schist4f	Tourmaline Schist	Metamorphic	Fine	A
152	schist5c	Graphite Schist	Metamorphic	Coarse	A
153	schist6c	Tremolite Schist	Metamorphic	Coarse	A
154	schist6f	Tremolite Schist	Metamorphic	Coarse	A
155	schist7c	Chlorite Schist	Metamorphic	Coarse	A
156	schist7f	Chlorite Schist	Metamorphic	Fine	A
157	schist8c	Anthophyllite Mica Schist	Metamorphic	Course	A
158	schist8f	Anthophyllite Mica Schist	Metamorphic	Fine	A
159	schist9c	Hornblende Schist	Metamorphic	Coarse	P
160	schist9f	Hornblende Schist	Metamorphic	Fine	P
161	shale1c	Shale (Arenacious)	Sedimentary	Coarse	A
162	shale1f	Shale (Arenacious)	Sedimentary	Fine	A
163	shale2c	Shale (Phosphatic)	Sedimentary	Coarse	A
164	shale2f	Shale (Phosphatic)	Sedimentary	Fine	A

	Filename	Mineral Name	Type	Form	Carbonate
165	shale3c	Shale (Calcareous)	Sedimentary	Coarse	P
166	shale3f	Shale (Calcareous)	Sedimentary	Fine	P
167	shale4c	Black Shale	Sedimentary	Coarse	P
168	shale4f	Black Shale	Sedimentary	Fine	P
169	shale5c	Illite-bearing Shale	Sedimentary	Coarse	A
170	shale5f	Illite-bearing Shale	Sedimentary	Fine	A
171	shale6c	Carbonaceous Shale	Sedimentary	Coarse	P
172	shale6f	Carbonaceous Shale	Sedimentary	Fine	P
173	shale7c	Argillaceous Shale	Sedimentary	Coarse	P
174	shale7f	Argillacious Shale	Sedimentary	Fine	P
175	siltst1c	Siltstone	Sedimentary	Coarse	A
176	siltst1f	Siltstone	Sedimentary	Fine	A
177	siltst2c	Limestone Siltstone	Sedimentary	Coarse	P
178	siltst2f	Limestone Siltstone	Sedimentary	Fine	P
179	slate1c	Gray Slate	Metamorphic	Coarse	P
180	slate1f	Gray Slate	Metamorphic	Fine	P
181	slate2c	Green Slate	Metamorphic	Coarse	P
182	slate2f	Green Slate	Metamorphic	Fine	P
183	slate3c	Chiastolic Slate	Metamorphic	Coarse	A
184	slate3f	Chiastolic Slate	Metamorphic	Fine	A
185	syenit1f	Alkalic Syenite	Igneous	Fine	PP
186	syenit1s	Alkalic Syenite	Igneous	Solid	PP
187	syenit2f	Nepheline Syenite	Igneous	Fine	PP
188	syenit2s	Nepheline Syenite	Igneous	Solid	PP
189	tonali1f	Tonalite	Igneous	Fine	PP
190	tonali1s	Tonalite	Igneous	Solid	PP
191	traver1c	Travertine	Sedimentary	Coarse	P
192	traver1f	Travertine	Sedimentary	Fine	P

Appendix C

Silver Lake Samples

C.1 Introduction

The Silver Lake samples consist of 21 spectra measured in the field at Silver Lake near Baker, CA. Table C.1 lists the names of these spectra and shows expert judgments of carbonate content for them. Tables C.2 and C.3 show the raw data from the experiment, along with the results of analyzing it using the modified PC, simultaneous linear regression, and an expert system.

C.2 Preparation

The following technical description of the preparation of the Silver Lake field samples was provided by Ted Roush, Senior Geologist at NASA Ames:

> 1. *Spectrometer Description.*
>
> The FieldSpecFRä (Analytical Spectral Devices, Inc., ASD) is a fiber optic spectrometer operating over the 350-2500 nm wavelength range. This device uses 3 detectors operating over 3 wavelength domains. In the 350-1000 nm region (VIS) a fixed grating is used to disperse the wavelengths across a

Si-Photodiode detector array. In the 1000-1800 nm (SWIR1) and 1800-2500 nm (SWIR2) regions rotatable gratings are used to disperse the wavelengths onto single point InGaAs detectors. Instrument preparation consists of optimization, which sets the gain and integration time of the VIS region and the gain and DC offset of the SWIR1 and SWIR2 regions. After optimization a dark current measurement is made for subsequent processing, and additional dark current measurements may be made at various times throughout the data collection processes. A typical data collection sequence consists of measuring the reflectance of a bright, spectrally neutral reference target followed by measurements of the sample of interest. The spectrometer was operated in an automated mode such that the dark current is subtracted from both measurements and the ratio of the sample to reference was calculated for immediate display.

2. *Data Collection Approach.*

The reflectance measurements obtained at the Silver Lake site were a combination of data collected using the fiber optic cable mounted in a hand-held pointing device and with the fiber optic cable attached to a 1-degree field of view foreoptic telescope. In both cases, a typical data collection sequence consists of measuring the reflectance reference target followed by measurements of the sample of interest.[1]

3. *Spectral Artifacts.*

Instrumental. Obvious in many spectra is a distinctive reflectance difference between the end of the VIS and beginning of the SWIR1 wavelength regions. Upon closer inspection one may find a similar difference between the end of the SWIR1 and beginning of the SWIR2 wavelength regions. These are due to changing detector sensitivities when they are used under different ambient conditions, both internal and external to the spectrometer. Literature from ASD suggests that the SWIR1 region does not suffer from such variable sensitivities, providing a potential mechanism of correcting the VIS and SWIR2 regions if desired. No corrections have been applied to the Silver Lake data.

Atmospheric. Telluric water vapor has several strong absorptions in the 350-2500 nm wavelength region. The two strongest are centered near 1350-1450 and 1800-1950 nm regions and are readily apparent in the spectra obtained at Silver Lake. In addition, there are weaker features located near 800, 900, and 1150 nm that are occasionally present in the Silver Lake data.

Albedo. Although reflectances are recorded in the data files, one must be very careful regarding the absolute values reported. This is due to a variety of potential sources of error that include differences in viewing geometry, distances, and environmental effects between the time that the reference

[1] To compensate for the varying power function of sunlight, digitalized spectra, taken in the field, were automatically divided by the spectrum of a white reference surface placed near the target, and these ratios are recorded.

and sample measurements were obtained. For example, while the reference may be located in the vicinity of the sample, it may be oriented relative to the illumination source quite differently than the portion of the subsequent sample that is being measured. An example of changing environmental effects would be the presence of clouds or shade during collection of reference data that are absent when the sample data is acquired, or vice versa.

C.3 Expert Identifications

Expert judgments of carbonate content for 21 field spectra collected near Silver Lake, CA. "Sample name" refers to the code name given to each sample used (for blind testing). "Field Expert ID" refers to the judgment of carbonate content ("C" = contains carbonate, "NC" = does not contain carbonate) by experts in the field with access to fields samples and spectra. "Laboratory ID" refers to the results of chemical testing for carbonate content for selected samples ("C" and "NC" are as above; "NA" = the sample was not tested)

Table C.1: Expert carbonate identifications of Silver Lake field samples.

Sample Name	Field Expert ID	Laboratory ID
Emperor #1	C	C (90%) NC(10%)
Emperor #1	C	C (90%) NC(10%)
T 103	NC	NA
T 105	NC	NA
T 106	C	NA
Endolith	C	C (93%) NC (7%)
Tubular-tabular	NC	NC (100%)
Arroyo disturbed	C	C (20%) NC (80%)
Arroyo undisturbed	C	C (25%) NC (75%)
C3PO	C	NA
Chewie	NC	NA
Jabba	C	NA
Jawa	C	NA
Lando	C	C (93%) NC(7%)
Luke	C	NA
R2D2	C	C (78%) NC (22%)
Solo	C	NA
Tarken	NC	NA
Vader	NC	NA
Valentine	NC	NC (100%)
Yoda	NC	NA

Table C.2: Carbonate identifications of field spectra from Silver Lake near Baker, California using the modified PC algorithm.

Column 1: Nickname given to each sample for purposes of analysis.
Column 2: Carbonate ID of the rock by an expert in the field.
Column 3: Modified PC ID, reporting all carbonates, wavelength range $2.0\mu m, 2.5\mu m$.
Column 4: Modified PC ID, reporting calcites and dolomites only, wavelength range $2.0\mu m, 2.5\mu m$.
Column 5: Modified PC ID, reporting calcites and dolomites only, wavelength range $0.4\mu m, 2.5\mu m$.
Column 6: Modified PC ID, reporting all carbonates, wavelength range $0.4\mu m, 2.5\mu m$.
Column 7: Laboratory ID.

1	**2**	**3**	**4**	**5**	**6**	**7**
Emperor #1	C	C	C	C	C	C (90%) NC (10%)
Emperor #2	C	C	C	C	C	C (90%) NC (10%)
T 103	NC	NC	NC	NC	NC	NA
T 105	NC	NC	NC	C	C	NA
T 106	C	C	C	C	C	NA
Endolith	C	C	C	C	C	C (93%) NC (7%)
Tubular-tabular	NC	NC	NC	NC	C	NC (100%)
Arroyo disturbed	C	NC	NC	C	C	C (20%) NC (80%)
Arroyo undisturbed	C	C	C	C	C	C (25%) NC (75%)
C3PO	C	C	C	C	C	NA
Chewie	NC	C	NC	NC	C	NA
Jabba	C	C	C	C	C	NA
Jawa	C	C	C	C	C	NA
Lando	C	C	C	C	C	C (93%) NC (7%)
Luke	C	C	C	C	C	NA
R2D2	C	C	C	C	C	C (78%) NC (22%)
Solo	C	C	C	C	C	NA
Tarken	NC	NC	NC	NC	NC	NA
Vader	NC	NC	NC	C	C	NA
Valentine	NC	NC	NC	C	C	NC (100%)
Yoda	NC	NC	NC	NC	C	NA
Total Correct:		19	20	18	15	

Table C.3: Carbonate identifications of field spectra from Silver Lake near Baker, CA using the simultaneous linear regression algorithm in Minitab v. 10..

Column 1: Nickname given to each sample for purposes of analysis.
Column 2: Carbonate ID of the rock by an expert in the field.
Column 3: Carbonate ID using simultaneous linear regression, using wavelength range $0.4\mu m, 2.5\mu m$.
Column 4: Carbonate ID using simultaneous linear regression, using wavelength range $0.4\mu m, 2.5\mu m$, reporting carbonates with positive regression coefficients only.
Column 5: Carbonate ID using simultaneous linear regression, using wavelength range $0.4\mu m, 2.5\mu m$, reporting calcites and dolomites only.
Column 6. Carbonate ID using simultaneous linear regression, using wavelength range $0.4\mu m, 2.5\mu m$, reporting calcites and dolomites only with positive regression coefficients.
Column 7. Expert system ID.
Column 8. Laboratory ID.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Emperor #1	C	C	C	C	C	C	C (90%) NC (10%)
Emperor #2	C	C	C	C	C	C	C (90%) NC (10%)
T 103	NC	C	C	C	C	NC	NA
T 105	NC	C	C	C	C	NC	NA
T 106	C	C	C	C	C	C	NA
Endolith	C	C	C	C	C	C	C (93%) NC (7%)
Tubular-tabular	NC	C	C	C	NC	NC	NC (100%)
Arroyo disturbed	C	C	C	C	C	NC	C (20%) NC (80%)
Arroyo undisturbed	C	C	C	C	C	C	C (25%) NC (75%)
C3PO	C	C	C	C	C	C	NA
Chewie	NC	C	C	NC	NC	NC	NA
Jabba	C	C	C	C	NC	NC	NA
Jawa	C	C	C	C	NC	C	NA
Lando	C	C	C	C	C	C	C (93%) NC (7%)
Luke	C	C	C	C	C	C	NA
R2D2	C	C	C	C	C	NC	C (78%) NC (22%)
Solo	C	C	C	C	C	NC	NA
Tarken	NC	C	NC	C	NC	NC	NA
Vader	NC	C	C	C	C	NC	NA
Valentine	NC	C	C	C	NC	NC	NC (100%)
Yoda	NC	C	C	C	C	NC	NA
Total Correct:		13	14	14	15	17	

Bibliography

Bernath, Peter F. 1995. *Spectra of Atoms and Molecules.* New York: Oxford University Press.

Bollen, Kenneth A. 1995. *Structural Equations with Latent Variables.* New York: John Wiley & Sons.

Cartwright, Nancy. 1999a. "Causal Diversity and the Markov Condition." *Synthese* 121:3–27.

———. 1999b. *The Dappled World: A Study of the Boundaries of Science.* New York: Cambridge University Press.

Clark, Roger N. 1999. "Spectroscopy of Rocks and Minerals and Principles of Spectroscopy." Chapter 1 of *Remote Sensing for the Earth Sciences*, edited by Andrew N. Rencz, Volume 3 of *Manual of Remote Sensing*, 3–58. New York: John Wiley & Sons.

Clark, R. N. and T. L. Roush. 1984. "Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications." *J. Geophys. Res.* 89:6329–6340.

Cooper, G. and E. Herskovits. 1991. "A Bayesian Method for Constructing Bayesian

Belief Networks from Databases." *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 86–94.

———. 1992. "A Bayesian Method for the Induction of Probabilistic Networks from Data." *Machine Learning* 9:309–347.

Cramer, Harald. 1951. *Mathematical Methods of Statistics.* Princeton: Princeton Univerity Press.

Fisher, R. 1951. *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Freedman, David and Paul Humphreys. 1999. "Are There Algorithms That Discover Causal Structure?" *Synthese* 121:29–54.

Gaffey, S. J. 1987. "Spectral reflectance of carbonate minerals in the visible and near-infrared (0.35-2.55 μm): anhydrous carbonate minerals." *J. Geophys. Res.* 92:1429–1440.

Gaines, R., C. Skinner, E. Foord, B. Mason, and A. Rosenzweig, eds. 1997. *Dana's New Mineralogy.* New York: John Wiley & Sons.

Gazis, Paul R. and Ted L. Roush. 1999. "Autonomous identification of carbonates using near-IR reflectance spectra during the February 1999 Marsokhod Field Tests." *J. Geophys. Res.* in press.

Glymour, Clark. 1999. "Rabbit Hunting." *Synthese* 121:55–78.

Grove, C. I., S. J. Hook, and E. D. Paylor II. 1992. Laboratory Reflectance Spectra of 160 Minerals, 0.4 to 2.5 Micrometers. JPL-Publication 92-2.

Hamilton, Lawrence C. 1992. *Regression with Graphics: A Second Course in Applied Statistics.* Belmont, CA: Duxbury Press.

Hapke, B. 1993. *Theory of Reflectance and Emittance Spectroscopy.* New York: Cambridge University Press.

Johnson, J. R. et al. 1999. "Geological characterization of remote field sites using visible and infrared spectroscopy: Results from the 1999 Marsokhod Field Tests." *J. Geophys. Res.* in press.

Merenyi, E. 2000. ""Precision Mining" of High-Dimensional Patterns with Self-Organizing Maps: Interpretation of Hyperspectral Images." In *Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence*, edited by Peter Sincak and Jan Vascak, Volume 54 of *Studies in Fuzziness and Soft Computing.*

Miller, Randolph. 1982. "INTERNIST-1, An Experimental Computer-based Diagnostic Consultant for General Internal Medicine." *New England Journal of Medicine* 15:525–43.

Minitab Release 10. Minitab Inc. State College, PA.

Model 1 32 bit Pro Plus Edition, Version 3.1.304. Unica Technologies, Lincoln, MA.

Moody, Jonathan, Ricardo Silva, and Joseph Vanderwaart. 2000, November. "Feature Selection for Rock Classification from Spectra." Final report for course project.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann.

Pieters, C. M. and P. A. J. Englert, eds. 1993. *Remote Geochemical Analysis: Elemental and Mineralogical composition.* New York: Cambridge University Press.

Rencz, Andrew, ed. 1999. *Remote Sensing for the Earth Sciences.* Third. Volume 3 of *Manual of Remote Sensing.* New York: John Wiley & Sons.

Robins, J., R. Scheines, P. Spirtes, and L. Wasserman. 1999. "The limits of causal knowledge." Technical Report, Carnegie Mellon University Philosophy Department.

Salisbury, John W., Louis S. Walter, Norma Vergo, and Dana M. D'Aria. 1991. *Infrared (2.1 - 25 μm) Spectra of Minerals.* Baltimore: Johns Hopkins University Press.

Scheines, Richard, Peter Spirtes, and Christopher Meek. 1994. TETRAD II. Laurence Erlbaum Publishers.

Scheines, Richard, Peter Spirtes, Clark Glymour, and Christopher Meek. 1994. *Tetrad II: User's Manual.* Hillsdale, N.J.: Lawrence Erlbaum.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search.* New York: Springer-Verlag.

———. 2000. *Causation, Prediction, and Search.* Second. Cambridge, MA: MIT Press.

Stoker, C. et al. 1999. "The 1999 Marsokhod rover mission simulation at Silver Lake California: Mission overview, data sets, and summary of results." *J. Geophys. Res.* in press.

Vidal, Marc. 2001. "A Biological Atlas of Functional Maps." *Cell* 104:333–339.

Wendlandt, Wesley Wm. and Harry G. Hecht. 1966. *Reflectance Spectroscopy.* New York: John Wiley & Sons.