1995

Data Analysis Using Reconstructability Theory.

Deky Dicky joe Gouw
Louisiana State University and Agricultural & Mechanical College

DATA ANALYSIS
USING RECONSTRUCTABILITY THEORY

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Computer Science

by
Deky Dicky Joe Gouw
Ir., Parahyangan Catholic University, Indonesia, 1987
M.S., Louisiana State University, 1990
May 1995

UMI Number: 9538732

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Bush Jones, for his patience, help, and guidance throughout the research and preparation of this report.

I would also like to thank Professors S. Sitharama Iyengar, Donald Kraft, John Tyler, Jianhua Chen of the Computer Science Department; Ye-Sho Chen (Minor Professor) of the Quantitative Business Analysis Department; and Gregory Hussey (Dean's Representative) of the Physics and Astronomy Department for serving as members of the examining committee.

A special gratitude goes to all my teachers, especially the faculty members in Computer Science Department, for providing me with the knowledge I need to achieve this goal and for providing me with financial support during my study here. Also I wish to express my gratitude to many friends who have given me help and advice.

Finally, I would like to extend my special heartfelt thanks to my wife, Karolina Dhanudimuljo, for her sacrifice, love, understanding, and support during my academic endeavors.

This research was supported in part by the Office of Naval Research under grant number N00014-91-J-1306.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this dissertation, we develop a new *data analysis* method using reconstructability theory, particularly *k*-systems theory. *K*-systems or Klir systems theory is a branch of reconstructability theory and provides a setting wherein common problems in statistics can be solved using the power of information theory by invoking the maximum entropy principle. The method is superior over classical statistical data analysis. The approach starts with the concept of variable interaction in reconstructability analysis (RA). We then use the RA definition of interaction to improve the quality of main and interaction effects in data analysis.

With classical statistical data analysis, we need to assume a model for the data and then check the validity of the model using such techniques as plotting residuals. Questions arise not only on how the model relates to the underlying data but also on the relevance of the assumptions which accompany the model. If these assumptions are grossly violated, the procedures used to draw inferences about the model may be invalid.

Our method overcomes this problem. Unlike classical statistical data analysis, our method assumes no model for the data and works directly with whatever information is available in the data. Thus, no model validity checking or assumption verification is needed. In addition, the results obtained are valid and true for the given data. The method has been used to analyze real experimental design data.

Chapter 1

GENERAL INTRODUCTION

1.1. Overview

As we can see from the title of this dissertation, there are two different fields of science involve in this research, i.e., data analysis and reconstructability theory. This research will explore the application of this relatively new theory to a concept in statistics, namely, variable interaction, and apply that to develop a new generation of data analysis method, a model-less data analysis.

Data analysis is a technique of exploring and analyzing a set of data to gain insights into the data, which includes both summarization of the information in the data and exposure of unanticipated anomalies and phenomena. It has emerged as an important discipline in statistics which emphasizes formal structure of the system of inference that often requires prespecifying many things such as the hypothesis to be tested and the model associated with the underlying data.

Reconstructability analysis (RA) is a process of investigating the possibilities of reconstructing desirable properties of overall systems from the knowledge of the corresponding properties of their subsystems (Klir 1985a). RA contains a powerful set of concepts and procedures dealing with wholes and parts, systems and subsystems, or states and substates. This notion of systems and subsystems provides new important insights into the structure and dynamics of the systems.

The concept of RA, particularly k-systems, has opened a door for new approaches to some problems in statistics. The application of k-systems to the concept of variable interaction in statistics will improve the concept greatly. The employment of substates in RA is a radical departure from the traditional statistical approach with variables. Substates allow us to examine system dynamics at its most refined level. We may

expect to find meanings that are hidden in analysis where variables are the fundamental unit (Jones 1985c).

Another advantage of the RA methodology over traditional statistical methods is that RA does not make any assumption on the structure of the system (data) and never introduces extraneous information to the system. It assumes no structure that does not explicitly exist - structure is discovered, not forced. The analysis on a k-system will never modify information in any way. Solution properties depend only on existing information and never introduce new information of any kind (Jones 1985c, 1986).

The concept of variable interaction is a very important concept in statistics. Interactions among variables are usually considered in data analysis involving two or more variables. The presence of variable interaction in the data, if not treated properly, could lead to a misleading conclusion. When the effect of interaction among variables is large, the main effects of the corresponding variables have little practical meaning, that is, knowledge of the interaction effect is more useful than knowledge of the main effects. A significant interaction will often mask the significance of main effects.

1.2. Some Related Works

Some notable works in the area of reconstructability analysis and k-systems have been reported. RA emerged from ideas formulated by George Klir in the mid-1970s (Klir 1976). These ideas were sparked primarily by Ashby (1964), who initially recognized and introduced the reconstructability problem for many-dimensional relations. Cavallo and Klir (1979, 1981a,b, 1982a,b) have developed to a great extent the concept of RA and introduced its applications in three different systems: relational, probabilistic, and possibilistic systems. Finally, a report on RA containing a description of problems that are studied under RA, a summary of the main results obtained within RA, and an outline of some important issues associated with RA could be found in Klir and Way (1985).

Jones (1982, 1985a-e, 1986, 1989) has improved the concept of RA for probabilistic systems. He has presented a very effective and efficient algorithm for determining unbiased reconstructions. He also has introduced the concept of k-systems and its application for multivariate data.

Works in RA can be found in quite a number of recent research publications, whereas works in variable interactions are rather difficult to find. Most of the initial works in variable interactions were done a long time ago and have not changed too much since then. Scheffe' (1959) has provided a good discussion of the variable interaction concept. Kettenring (1983a,b) has shown the least squares estimates of variable interaction effects. A good number of other authors have also provided detailed discussions about variable interaction in their statistical textbooks (Box et al. 1978; Hicks 1982; Mason et al. 1989; Montgomery 1991).

While the works and references in these two separate subjects, RA and variable interaction, still can be seen in research publications and statistical textbooks, the combined work in interaction concept and RA methodology can not be found anywhere. Nobody else, as far as we know, has done any research combining the two concepts.

1.3. Scope and Limitations of The Dissertation

Since the initial inception of RA in the mid-1970s, applications of RA have been explored in many areas of sciences, including statistics, social science, computer science, medical science, and agricultural science. The discussion in this dissertation will be focussed on problems in statistics, particularly experimental design and data analysis problems.

We attempt to combine the concept of RA methodology with the concept of variable interaction in statistics. That is, we attempt to use the RA definition of interaction to base other definitions in statistics. We use some works described above to incorporate the RA concept of interaction into other concepts in statistics. In other words, we are

trying to utilize the novel concept of RA to improve the quality of variable interaction. The employment of substates in RA allows us to compute true variable interaction effects. This idea, in turn, could be used to develop other concepts in a design of experiments and data analysis, and eventually to improve the quality of the experimental design method and data analysis itself.

RA is predominantly based upon principles of information theory. It is formulated, for the most part, within classical set theory and probability theory, but attempts have been made to extend the formulation to the more general framework of fuzzy set theory (Dubois and Prade 1980). RA has been extended to an area of fuzzy set theory that is referred to as possibility theory which is beyond the scope of this dissertation. This dissertation will discuss RA in the context of probability theory and probabilistic systems only.

1.4. Structure of The Dissertation

The organization of the dissertation is as follows.

Chapter 2 serves as a preliminary chapter to discuss the background theories, including the theory about RA and some important points related to experimental design methods. These are needed as the fundamental blocks to understand the subsequent chapters. The basic concepts of reconstructability theory including systems and subsystems, different types of problems in RA along with the procedures to solve them, and the concept of g-systems and k-systems are explained in this chapter. Also, the basic concepts of experimental design methods, especially factorial experiments, are discussed here.

Chapter 3 presents the fine details of the interaction concept in the k-systems and statistical context. Using a set of sample data, quantitative differences between the results of the two concepts are demonstrated and the comparison shows greatly disparate results.

Chapter 4 describes the practical use of k-systems theory to compute main and interaction effects in data analysis problems. An example of fractional factorial design and its confounding problem is presented quite elaborately. Then, we show how we overcome this problem using the k-systems theory. A new concept for computing main effects is developed, and a general algorithm for design of experiments and data analysis in general using k-systems theory is presented.

Chapter 5 provides a case study using a set of real experimental design data. The data consists of figures of dry weights of sorghum obtained from two 3x3x2 factorial experiments with 4 replications. The primary objective is not only to describe the results but also to present an in-depth analysis of the data using k-systems theory.

Finally, chapter 6 gives the summary and conclusion for the results achieved in this research. A discussion on future works in this research area is also given in this chapter.

Chapter 2

PRELIMINARIES

2.1. Introduction

In this chapter, we present the preliminary theory of reconstructability analysis (RA) and the basic concepts of a design of experiments. They are provided as the foundation for the subsequent chapters. We describe the fundamental theory of RA in section 2.2 and the basic concepts of a design of experiments in section 2.3.

2.2. Reconstructability Analysis

RA has rapidly emerged as a new important discipline in general systems theory. It has been recognized and described by Ross Ashby in the early 1960's (see Ashby 1964) but has been comprehensively investigated only after George J. Klir's research in the mid-1970's (see Klir 1976). RA has provided a powerful tool for the study of the relationships between parts and wholes, the relationships between subsystems and systems, or the relationships between substates and states.

2.2.1. Systems and Subsystems

The most fundamental concept in RA is that of systems and subsystems. A system constitutes a set of states and subsystems in general constitute substates of the system. The status of a system and subsystem is not absolute. Any given system can assume the role of a subsystem (a part) in one context and the role of a system (a whole) in another context. This makes it possible to represent each system by a hierarchy of collections of subsystems, i.e., by a collection whose subsystems are also represented by collections of subsystems. A collection of subsystems in RA is also called a *structure system* or a *reconstruction hypothesis*.

Associated with a system is a finite set of variables $\{v_i\}$. Each non-empty subset of the variables identifies one subsystem of the system. States $\{\alpha\}$ and substates $\{\beta\}$ of the system are determined by particular value assignments to the variables (Jones 1985c,d).

v_1	v_2	v_3	$f(.)$
0	0	0	0.20
0	0	1	0.10
0	1	0	0.15
0	1	1	0.20
1	0	0	0.10
1	0	1	0.05
1	1	0	0.15
1	1	1	0.05

FIGURE 2.1 An Example of A System

To illustrate the concept of system and subsystems, let us consider the example of a system in figure 2.1. The figure shows a probabilistic system with 3 variables and a probabilistic behavior/system function f. All non-empty subsets of the variables of the system produce all possible subsystems (structure system). Figure 2.2 presents all the subsystems for the system in figure 2.1.

v_1	${}^1f(.)$
0	0.65
1	0.35

v_2	${}^2f(.)$
0	0.45
1	0.55

v_3	${}^3f(.)$
0	0.60
1	0.40

v_1	v_2	${}^4f(.)$
0	0	0.30
0	1	0.35
1	0	0.15
1	1	0.20

v_2	v_3	${}^5f(.)$
0	0	0.30
0	1	0.15
1	0	0.30
1	1	0.25

v_1	v_3	${}^6f(.)$
0	0	0.35
0	1	0.30
1	0	0.25
1	1	0.10

FIGURE 2.2 An Example of A Set of Subsystems

Formally, given an overall system B defined with probabilistic system function f and a subsystem, say ${}^{m}B$, then ${}^{m}f$ must satisfy the condition

$${}^{m}f(\beta) = \sum_{\alpha > \beta} f(\alpha) \qquad .....(1)$$

where $\beta \in {}^{m}B$ and $\alpha \in B$. The probability function f is associated with the system, and its marginals ${}^{m}f$ are associated with the subsystems.

In this system, we also note that

$$\sum_{\alpha} f(\alpha) = 1. \qquad .....(2)$$

This might not be the case when dealing with general functions that will be discussed in the next subsection. With general functions, the sum might not be equal to one.

The above linear equations can be viewed as a general reconstruction hypothesis. The complete set of solutions to these equations is the *reconstruction family*, and the unique maximum entropy solution is the *unbiased reconstruction* (Jones 1985a).

Besides probabilistic systems, RA is also relevant to other kinds of systems, e.g., relational systems (see Cavallo and Klir 1979) and fuzzy or possibilistic systems (see Cavallo and Klir 1982). We are not going to discuss those systems here.

2.2.2. Type of Problems in RA

Two problems are involved in RA. The first problem is the *identification problem* where a set of subsystems (parts) is given and the goal is to derive from the information in the subsystems as much information as possible regarding the overall system (Jones 1982, 1985a). The second problem is the *reconstruction problem* where an overall system (whole) is given and the goal is to determine a set of subsystems needed to reconstruct the overall system to an acceptable degree of approximation (Jones 1985b,c,d).

The two problems can be summarized as follows:

- Identification Problem

A set of subsystems (parts)	------------>	Overall system (whole)

- Reconstruction Problem

An overall system (whole)	------------>	A set of subsystems (parts)

Identification Problem

Here, we will briefly discuss the idea on how an overall system is computed from a structure system (a set of subsystems). Given a structure system $S = \{{}^mB\}$, a set of linear equations based on equations (1) and (2) can be easily defined. For example, using the structure system given in figure 2.2, we define ${}^4f(01)$ as follows:

$$f(010) + f(011) = {}^4f(01) = 0.35.$$

Other ${}^4f(.)$ as well as other ${}^mf(.)$ are defined similarly.

Solving the set of linear equations for the unknown $f(\alpha)$ using the *maximum entropy algorithm* (Jones 1985a) will give the unbiased reconstruction of the overall system. This algorithm is also known as the *unbiased reconstruction algorithm.*

The algorithm is simply a particular iterative scheme for solving the equations. First, we rewrite the set of linear equations as

$$\sum_j f_{ij} = a_i \quad \text{for all} \quad a_i = {}^mf(\beta)$$

and then solve them by the maximum entropy algorithm (Jones 1985a).

1) Initialize f_{ij} to a flat distribution.

2) For all i:

$$\text{new } f_{ij} = f_{ij}\,(a_i/\hat{a}_i) \qquad \text{for every } j$$

where $\hat{a}_i$ is derived from the current estimate of f_{ij} and a_i is a true value.

3) Convergence test:

$$|\text{new } f_{ij} - \text{old } f_{ij}| \le \varepsilon \qquad \text{for all } \; i, j$$

If satisfied, stop.
If not satisfied, go to 2.

The crucial part of this algorithm is in initializing the unknown probability distribution to a flat distribution. A flat distribution is a probability distribution when the states are equally likely to occur ($f(\alpha) = 1/n$ for all states α, and n is the number of states). This is the probability distribution that will maximize the *information theoretic entropy* measure (Shannon 1948) and produce the most unbiased solution (Jaynes 1957). It is important to start with a flat distribution to get a unique and unbiased solution. If we start with a different probability distribution, we will not get the most unbiased solution.

Reconstruction Problem

The problem is to determine a set of substates $\{\beta\}$ whose unbiased reconstruction is within acceptable tolerance to represent the overall system. The idea of the solution is quite straightforward. Given an overall system, first, form a pool of all possible substates of the system. Next, generate a set of independent substates (Jones 1985a) $E = \{\, \beta_1, \beta_2, \ldots, \beta_k \,\}$ from that pool of substates.

What is a set of independent substates? To explain this, we need the concept of a null extension. Let β be a substate. Then a state α is the null extension of β if $\alpha > \beta$ and every variable of α which does not occur in β has the value zero. Two substates are

independent if and only if their null extensions are not the same. So, in a set of independent substates, no two substates have the same null extensions. Basically, a set of independent substates is a set of non-redundant substates.

Then, the algorithm (Jones 1985b) to solve this problem reconstructs the overall system using the substates in E. Every time, the algorithm selects and adds one substate from E and then reconstructs the system using the unbiased reconstruction algorithm explained in the previous section. In any case, the algorithm picks the most desirable substate from E that would reconstruct the system as close as possible to the original overall system. The result of this process is a set of substates that can reconstruct the original overall system within a desirable level of approximation.

In practice, the system function may not be *complete* ($f(\alpha)$ is unknown for certain α) (Jones 1985d). In order to analyze this system by the maximum entropy algorithms, we remedy this situation by assigning the mean of the known function values to each such unknown function value. This is referred to as the *entropy fill*. Also, in practice it is possible that there are two or more system function values for the same state (this can occur as a result of the clustering of variables). In this case, we average the redundant values to obtain a single value for a state (Jones 1985d).

2.2.3. *G*-systems and *K*-systems

The previous discussion has been on systems with probabilistic behavior functions. Jones (1985c) has extended the concepts of RA to general functions. A system with such a function is referred to as a g-system.

If A is the set of all aggregate states of the system, and R^+ is a set of positive real numbers, then

$$f: \quad A \rightarrow R^+$$

is a function that represents information of the system states.

Next, we define a parameter:

$$\tau = \sum_{\alpha \in A} f(\alpha).$$

(As we mentioned earlier, in g-system, the sum of system function values over all states α might not be equal to one. In this case, the sum is τ).

Then, a g-system can be defined as the following tuple:

$$(\tau, \{v_i\}, \{\alpha\}, \{\beta\}, f(.), \{^m f(.)\})$$

where:

- τ	is a parameter
- $\{v_i\}$	is a set of variables
- $\{\alpha\}$	is a set of states
- $\{\beta\}$	is a set of substates
- $f(.)$	is a function of $\{\alpha\}$
- $\{^m f(.)\}$	are functions of $\{\beta\}$

The function $f(.)$ is a measure of some type of information on system states, and commonly has units. In order for RA to be able to work on the system, we need to remove such units from the system by transforming it to a dimensionless system. Such a system is called a Klir-system or k-system. This system was named by Jones (1985c) in honor of the reconstructability analysis founder, George J. Klir.

The following normalization is used for that transformation:

$$k(\alpha) = \frac{f(\alpha)}{\tau}; \quad \text{for every} \quad \alpha,$$

so that

$$0 \leq k(\alpha) \leq 1; \quad \text{for every } \alpha \quad \text{and} \quad \sum_{\alpha} k(\alpha) = 1.$$

Now, a k-system can be defined as the following tuple:

$$(\tau, \{v_i\}, \{\alpha\}, \{\beta\}, k(.), \{^m k(.)\})$$

where:

- τ is a transformation parameter
- $\{v_i\}$ is a set of variables
- $\{\alpha\}$ is a set of overall states
- $\{\beta\}$ is a set of substates
- $k(.)$ is a function of $\{\alpha\}$
- $\{{}^m k(.)\}$ are functions of $\{\beta\}$

G-system and *k*-system induced from the *g*-system are isomorphic (Jones 1985c). They both contain the same system information. System information can be mapped from one system to the other. No information is lost or added from one system or the other. Everything in *g*-system is in *k*-system and everything in *k*-system is in *g*-system.

2.3. Design of Experiments

A design of experiments or experimental design is a method that involves a test or series of tests in which changes are made to the input variables of a process or system so that observations can be performed toward the response variable. In experimental design, the experimenter deliberately controls certain variables that may influence the outcome of the experiment. The experimenter then observes and measures the result.

Experimental design methods have found broad application in many disciplines. Experimental design is a critically important tool in the engineering world for improving the performance of a manufacturing process. It also has extensive application in the development of new processes. Experimental design methods play a major role in engineering design activities, where new products are developed and existing ones improved. The use of experimental design in these areas can result in products that are easier to manufacture, products that have enhanced field performance and reliability, lower product cost, and shorter product design and development time.

There are several techniques of experimental designs. Some of the most common designs are randomized block designs, latin square designs, and factorial designs. The first two designs are used when the experimenter is convinced that the variables in the experiment do not interact with each other. In other words, there is no interaction between variables in the experiment. When interactions among variables may be present in the model, a factorial design is necessary. The next sub-sections will focus on these subjects. But before we go further, we feel necessary to present a list of some common terms used in design and analysis of experiments.

2.3.1. Basic Terminology

This section presents the terms used in experimental design and analysis, which will be used throughout the dissertation. Some explanations of their meaning are also given.

- *Dependent Variable* or *Response Variable*. A variable whose changes we wish to study. It is the measured response of an experiment, the outcome of an experiment. This variable is expressed numerically.
- *Independent Variable* or just *Variable*. A particular force which is varied in the experiment and under the control of the experimenter. An independent variable may be *quantitative* or *qualitative*. A quantitative variable is one whose values can be measured on a numerical scale. A qualitative variable is one whose values are not usually arranged in order of magnitude. The values of qualitative variable cannot usually be measured on a numerical scale.
- *Levels*. The various values at which a variable is tested.
- *Treatment Combination* or *Level Combination*. One of the possible combinations of levels of all variables under investigation.
- *Test Run*. Single combination of levels that yields an observation on the response. The terms, treatment combination and test run are sometimes used interchangeably.

- *Replication.* Repetition of an entire experiment or a portion of an experiment.
- *Design.* An experimental design consists of specifying the number of experiments, the variable level combinations for each experiment, and the number of replications of each experiment.
- *Experimental Unit.* Entity on which the experiment is done. For example, a single animal may be an experimental unit in a feeding experiment with dairy cattle.
- *Effect.* Change in the average response between two variable level combinations or between two experimental conditions.
- *Variable Interaction* or *Interaction.* Existence of joint variable effects in which the effect of each variable depends on the levels of the other variables.
- *Confounding.* One or more effects that cannot unambiguously be attributed to a single variable or interaction.

2.3.2. Factorial Experiments

Many experiments involve the study of several variables simultaneously. All these variables can be studied in one experiment, and factorial design is most efficient for this type of experiment. The meaning of factorial design is that all possible combinations of the levels of the variables may exist in the design. For example, if there are a levels of variable A and b levels of variable B, then each design may contain all ab treatment combinations.

The dimensions of a factorial design are indicated by the number of variables and the number of levels for each variable. In a $p \times q$ factorial design, there are 2 variables, the first has p levels and the second has q levels. A p^n factorial design consists of n variables, each has p levels. A $p^n \times q^m$ factorial design contains n variables at p levels and m variables at q levels.

Basically there are two kinds of factorial designs, *full factorial design* and *fractional factorial design.*

A full factorial design utilizes every possible combination at all levels of the variables. A design with k variables, with the i^{th} variable having n_i levels, requires n experiments, where

$$n = \prod_{i=1}^{k} n_i$$

For example, a full factorial design with 10 variables where each variable has 2 levels would require 2^{10} experiments. The advantage of a full factorial design is that every possible combination is examined. We can find the effect of every variable and all the interactions among variables. Various variable effects could easily be calculated and their contributions to the system could easily be explained. The main problem is the cost of the design. Sometimes the number of experiments required for a full design is too large. This may happen if either the number of variables or their levels is too large. It may not be possible to use a full factorial design due to the expense or the time required. It would take too much time and money to conduct these many experiments. In such cases, the experimenter can use only a fraction of the full factorial design.

Fractional factorial designs save time and expense when compared to full factorial designs since they use considerably fewer experiments. For example, using a fractional factorial design to analyze k variables where each variable has 2 levels, we need 2^{k-p} experiments, where p is a suitably chosen integer, instead of 2^k experiments as for a full factorial design. A 2^{k-1} design requires only half as many experiments as a full factorial 2^k design. Similarly, a 2^{k-2} design needs only one-quarter of the experiments required in a full factorial design. However, there is one problem with fractional factorial designs. The information obtained from a fractional factorial design is less than that obtained from a full factorial design. Some of the variable effects cannot be determined. It may not be possible to get some of the variable interaction effects. Only the

combined influence of two or more effects can be computed. This problem is known as *confounding* and the effects whose influence cannot be separated from each other are said to be *confounded.*

The basic theory of the design and analysis of factorial experiments was first described by Fisher (1926), and was developed to a great extent initially by Yates (1935, 1937) and Bose and Kishen (1940). In subsequent years, a host of other research workers have contributed to this field. Right now, there are numerous books on design and analysis of experiments (in particular, see Box et al. 1978; Hicks 1982; McLean and Anderson 1984; Mason et al. 1989; Montgomery 1991; Lorenzen and Anderson 1993).

2.3.3. Main and Interaction Effects

The effect of a variable is defined to be the change in response produced by a change in the level of the variable. This is called a *main effect.* As an example, consider a 2^2 factorial experiment without interaction represented by the data in table 2.1.

TABLE 2.1 A Factorial Experiment without Interaction

VARIABLE		B b1	b2
A	a1	20	30
	a2	40	50

The main effect of variable A is computed as

$$A = \frac{(40-20)+(50-30)}{2} = \frac{20+20}{2} = 20.$$

That is, increasing variable A from level 1 to level 2 causes an average response increase of 20 units. Similarly, the main effect of B is

$$B = \frac{(30-20)+(50-40)}{2} = 10.$$

Sometimes, we may find that the difference in response between the levels of one variable is not the same at all levels of the other variables. This is due to the presence of an *interaction effect* between the variables. For example, consider the data in table 2.2.

TABLE 2.2 A Factorial Experiment with Interaction

VARIABLE		B b1	b2
A	a1	20	40
	a2	50	12

At the first level of variable B, the A effect is

$$A = 50 - 20 = 30$$

and at the second level of variable B, the A effect is

$$A = 12 - 40 = -28.$$

Since the effect of A depends on the level chosen for variable B, we see that there is interaction between A and B.

These ideas may be illustrated graphically. Figure 2.3 plots the response data in table 2.1 against variable A for both levels of variable B. The b_1 and b_2 lines in that figure are parallel, indicating a lack of interaction between variables A and B. Similarly, figure 2.4 plots the response data in table 2.2. Here, the b_1 and b_2 lines are not parallel. This indicates an interaction between variables A and B.

When an interaction is large, the corresponding main effects have little practical meaning. For the data in table 2.2, we would estimate the main effect of A would be

$$A = \frac{(50-20)+(12-40)}{2} = 1$$

which is very small, and we are tempted to conclude that there is no effect due to A.

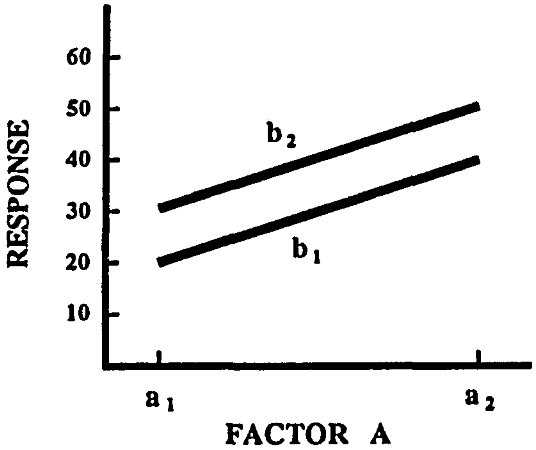

FIGURE 2.3 A Factorial Experiment without Interaction

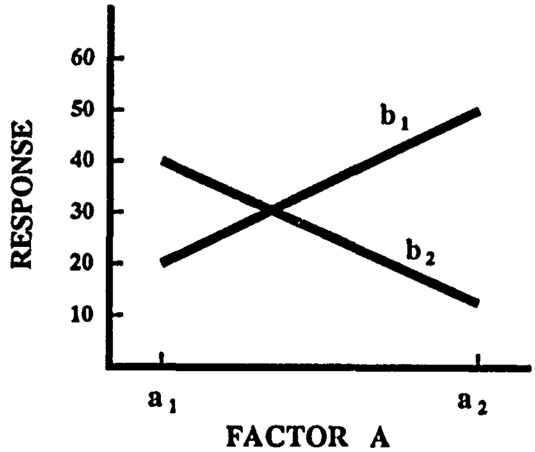

FIGURE 2.4 A Factorial Experiment with Interaction

This is wrong. When we examine the effects of *A* at the first and second levels of *B*, we see that variable *A* does have an effect, but it depends on the level of variable *B*. A significant interaction effect will often mask the significance of main effects. This shows how important the computation of interaction effects in factorial experiment is.

Chapter 3

VARIABLE INTERACTION CONCEPT

3.1. Introduction

Interaction is an additional effect due to the combined influence of two or more variables. For example, in a case involving four variables, there would be four main effects (A, B, C, D), six two-variable interactions (AB, AC, AD, BC, BD, CD) involving the combined effect of two variables, four three-variable interactions (ABC, ABD, ACD, BCD) involving the combined effect of three variables, and one four-variable interaction (ABCD) involving the combined effect of all four variables.

In this chapter, we present the details of the interaction concept in both statistical context and reconstructability analysis theory, particularly k-systems theory. Also, a set of sample data is used to compare and analyze the quantitative results of the two interaction concepts.

3.2. Statistical Interaction

In statistics, interaction is usually discussed in the context of experimental design and the analysis of variance for multiple variables. It is used in finding the confounding factors in fractional factorial design. This allows an experiment for a certain number of parameters using a smaller number of runs.

Let us start with three-way *complete* data with one observation per cell. Complete means that there is at least one observation for every cell; there is no cell without data. If some data values are missing from the cells, then they have to be estimated.

Suppose there are three independent variables A, B, and C with *I levels*, *J levels*, and *K levels* respectively. The observations can be arranged in a *three-way layout* or $I \times J \times K$ table. We denote by y_{ijk} as the expected value of the observation on the i,j,k cell,

i.e. when A is at the *i*th level, B at the *j*th level, and C at the *k*th level. One version of the three-way analysis of variance models takes the following form (Johnson and Graybill 1972; Kettenring 1983b; Scheffe' 1959)

$$y_{ijk} = \bar{y} + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk} + e_{ijk} \ ,$$

where $\bar{y}$ is the overall mean; A_i, B_j, and C_k are the *i*th, *j*th, and *k*th levels of the main effects of A, B, and C respectively; AB_{ij}, AC_{ik}, and BC_{jk} are the corresponding two-variable interactions; ABC_{ijk} is the three-variable interaction term; and e_{ijk} is the residual. The above model holds for the following constraints:

$$A_. = B_. = C_. = 0$$

$$AB_{i.} = AB_{.j} = AC_{i.} = AC_{.k} = BC_{j.} = BC_{.k} = 0$$

$$ABC_{ij.} = ABC_{i.k} = ABC_{.jk} = 0 \ ,$$

where "." denotes a subscript over which an average has been taken.

The least squares estimates of the main and interaction effects of the model are computed as follows:

$$\hat{A}_i = y_{i..} - y_{...}$$

$$\hat{B}_j = y_{.j.} - y_{...}$$

$$\hat{C}_k = y_{..k} - y_{...}$$

$$\hat{AB}_{ij} = y_{ij.} - y_{i..} - y_{.j.} + y_{...}$$

$$\hat{AC}_{ik} = y_{i.k} - y_{i..} - y_{..k} + y_{...}$$

$$\hat{BC}_{jk} = y_{.jk} - y_{.j.} - y_{..k} + y_{...}$$

$$\hat{ABC}_{ijk} = y_{ijk} - y_{ij.} - y_{i.k} - y_{.jk} + y_{i..} + y_{.j.} + y_{..k} - y_{...} \ ,$$

where the "^" denotes an estimate. These estimates satisfy all the constraints mentioned above.

If there are M ($M > 1$) observations on each of the cells (M replications per cell), then we introduce a fictitious "variable" D corresponding to the m-fold replication of the observations in the cells of the three-way data. This becomes a four-way layout. The least squares estimates of the main and interaction effects could easily be extended as follows:

$$\hat{A}_i = y_{i...} - y_{....}$$

$$\hat{B}_j = y_{.j..} - y_{....}$$

$$\hat{C}_k = y_{..k.} - y_{....}$$

$$\hat{D}_m = y_{...m} - y_{....}$$

$$\widehat{AB}_{ij} = y_{ij..} - y_{i...} - y_{.j..} + y_{....}$$

$$\widehat{AC}_{ik} = y_{i.k.} - y_{i...} - y_{..k.} + y_{....}$$

$$\widehat{AD}_{im} = y_{i..m} - y_{i...} - y_{...m} + y_{....}$$

$$\widehat{BC}_{jk} = y_{.jk.} - y_{.j..} - y_{..k.} + y_{....}$$

$$\widehat{BD}_{jm} = y_{.j.m} - y_{.j..} - y_{...m} + y_{....}$$

$$\widehat{CD}_{km} = y_{..km} - y_{..k.} - y_{...m} + y_{....}$$

$$\widehat{ABC}_{ijk} = y_{ijk.} - y_{ij..} - y_{i.k.} - y_{.jk.} + y_{i...} + y_{.j..} + y_{..k.} - y_{....}$$

$$\widehat{ABD}_{ijm} = y_{ij.m} - y_{ij..} - y_{i..m} - y_{.j.m} + y_{i...} + y_{.j..} + y_{...m} - y_{....}$$

$$\widehat{ACD}_{ikm} = y_{i.km} - y_{i.k.} - y_{i..m} - y_{..km} + y_{i...} + y_{..k.} + y_{...m} - y_{....}$$

$$\widehat{BCD}_{jkm} = y_{.jkm} - y_{.jk.} - y_{.j.m} - y_{..km} + y_{.j..} + y_{..k.} + y_{...m} - y_{....}$$

$$\widehat{ABCD}_{ijkm} = y_{ijkm} - y_{ijk.} - y_{ij.m} - y_{i.km} - y_{.jkm} + y_{ij..} + y_{i.k.} + y_{i..m} + y_{.jk.} + y_{.j.m} + y_{..km} - y_{i...} - y_{.j..} - y_{..k.} - y_{...m} + y_{....} \,.$$

For unequal cell numbers, the estimates become complicated. Some standard textbooks, such as (Mason et al. 1989) and (Scheffe' 1959), provide some discussions on that matter. The four-way analysis of variance models also satisfy the conditions similar for three-way models.

This result can be generalized for more variables. In the complete p-way data there are an overall mean, p main effects, n_2^p two-variable interactions, n_3^p three-variable interactions, ..., and 1 p-variable interaction, where n_q^p denotes the coefficient $p![q!(p-q)!]^{-1}$. The q-variable interaction is the sum of 2^q terms. The first term is the cell mean $y_{i,j,etc.}$ with subscripts not related to the q variables replaced by dots, next there are n_1^q terms with minus signs, obtained by replacing the q subscripts by dots one at a time, next n_2^q terms with plus signs, obtained by replacing the q subscripts by dots two at a time, next n_3^q terms with minus signs, etc.

3.3. *K*-systems Theory

In this section, we will review briefly the important concept of the k-systems theory. We will use this concept later as the basis for computing interaction effects.

We first introduce some terminology. In the language of reconstructability analysis, we are concerned with a system with which we associate a finite set of variables $\{v_i\}$ which take either discrete or continuous values. In the case of a continuous variable, we cluster the values into a discrete set of categories. Also, a system has a behavior which is described by a real valued system function $f(.)$ (known as response variable in experimental design). Each nonempty subset of the variables identifies one subsystem of the system, and states $\{\alpha\}$ and substates $\{\beta\}$ of the system are determined by particular value assignments to the variables.

Now, in practice the system function may not be *complete* ($f(\alpha)$ is unknown for certain α) (Jones 1985d). In order to analyze this system by the maximum entropy

algorithms, we remedy this situation by assigning the mean of the known function values to each such unknown function value. This is referred to as the *entropy fill*, and results in minimizing the information added to the subsequent k-system where the $f(.)$ function has been transformed to a (0,1) function and we invoke information theoretic algorithms. Also, in practice it is possible that there are two or more system function values for the same state. This can occur as a result of the clustering of variables (or as a result of replication of some experiments in an experimental design). In this case, we average the redundant values to obtain a single value for a state (Jones 1985d).

The basic theory of the reconstructability analysis method was developed to a great extent initially by Cavallo and Klir (1981a,b). Jones (1985a-e, 1986, 1989) has defined some important concepts of the unbiased reconstruction and k-systems theory of the reconstructability analysis method.

The purpose of this section is not to discuss in detail all the developments that have taken place in the area of reconstructability analysis methodology, rather it is intended to present a brief overview to the interaction concept of the k-systems theory, and then show the use of this theory for computing interaction effects. The main effects can be computed using the same concept as for computing interaction effects. We will see how this is conducted in chapter 4.

First, we define a function ${}^{m}f(.)$ for each subsystem as follows

$${}^{m}f(\beta) = \sum_{\alpha > \beta} f(\alpha).$$

That is, we sum the system function over all states α for which β is a substate. We further define

$$\tau = \sum_{\alpha} f(\alpha).$$

The tuple $(\tau, \{v_i\}, \{\alpha\}, \{\beta\}, f(.), \{{}^m f(.)\})$ constitutes what is called the g-system (Jones 1985c). This is transformed into a Klir system or k-system (Jones 1985c) denoted by $(\tau, \{v_i\}, \{\alpha\}, \{\beta\}, k(.), \{{}^m k(.)\})$ by the transformations

$$k(\alpha) = \frac{f(\alpha)}{\tau} \qquad \text{for every } \alpha,$$

and

$${}^m k(\beta) = \sum_{\alpha > \beta} k(\alpha).$$

This transformation accomplishes a dimensionless (0,1) system, which can be analyzed by entropy mathematics. Such an analysis yields structural information which is isomorphic under this simple mapping to the original g-system.

We summarize the steps:

A) Cluster values of continuous variables.

B) Average redundant system function values.

C) Entropy fill missing system function values.

D) Map the original system into a k-system.

To compute interaction effects in the k-system, we need to be able to compute unbiased reconstructions or maximum entropy approximations of the system function from any set of substates.

Associated with each β is an equation

$${}^m k(\beta) = \sum_{\alpha > \beta} k(\alpha),$$

so that any selected set of β form a set of linear equations wherein we now assume the

$k(\alpha)$ are unknown. We rewrite this set of linear equations as

$$\sum_j k_{ij} = a_i \qquad \text{for all} \quad a_i = {}^m k(\beta)$$

and we solve them by the maximum entropy algorithm (Jones 1985a).

1) Initialize $k(.)$ to a flat distribution.
2) For all i:
   new $k_{ij} = k_{ij}(a_i/a'_i)$ for every j
   where a'_i is derived from the current estimate of k_{ij} and a_i is a true value.
3) Convergence test:
   $|\text{new } k_{ij} - \text{old } k_{ij}| \leq \delta$ for all ij
   If satisfied, stop.
   If not satisfied, go to 2.

3.4. *K*-systems Interactions

Let $V = \{v_1 = c_1, v_2 = c_2 \ , \ldots, v_n = c_n\}$ be a set of n variables with a particular value assignment (a factor or substate); let V_m be the set of all nonempty subsets of V; and let V_p denote the set of all nonempty *proper* subsets of V. We compute the unbiased reconstructions $\mu(V_m)$ and $\mu(V_p)$. Then, for any factor which embodies V, we consider the difference in its effect as computed from these two unbiased reconstructions. This difference represents the contribution to a factor effect that is due to the components acting in unison - the n-variable interaction.

As an example, let us consider a set of 3 variables $V = \{v_1 = c_1, v_2 = c_2, v_3 = c_3\}$. We compute the unbiased reconstruction for $V_m = \{/v_1 = c_1/,\ /v_2 = c_2/,\ /v_3 = c_3/,\ /v_1 = c_1,\ v_2 = c_2/, /v_1 = c_1,\ v_3 = c_3/, /v_2 = c_2,\ v_3 = c_3/, /v_1 = c_1,\ v_2 = c_2,\ v_3 = c_3/\}$ and $V_p = \{/v_1 = c_1/,\ /v_2 = c_2/,\ /v_3 = c_3/,\ /v_1 = c_1,\ v_2 = c_2/,\ /v_1 = c_1,\ v_3 = c_3/,\ /v_2 = c_2,\ v_3 = c_3/\}$.. Then, the difference in its effect as computed from these two unbiased reconstructions represents the interaction.

3.5. An Example of Statistical and *K*-systems Interactions

To illustrate these concepts, let us see a four-variable (*size, pressure, temperature, trial*) experiment. For simplicity, they are renamed as A, B, C, and D respectively. Variable A consists of 3 different values, also called 3 *levels*, i.e. 7.5, 12.5, and 17.5. Variable B has 3 *levels* (5.0, 12.5, 20.0), C has 3 *levels* (1900, 2000, 2300), and D has 2 *levels* (1, 2). The observation values (system function values) are denoted by y_{ijkm}. Table 3.1 illustrates the layout of the data and is used to help understand the concept. Table 3.2 provides the observation data in a four-way layout. Table 3.3 gives some of the interactions as computed by the two methodologies. The complete comparisons of the interaction effects for the data in this study can be found in appendix A. Appendix A also provides the comparisons of the main effects (will be discussed in chapter 4) for the data in this example.

TABLE 3.1 Notation Used for the 3x3x3x2 Experiment

VARIABLE				A							
				a_1			a_2			a_3	
							C				
			c_1	c_2	c_3	c_1	c_2	c_3	c_1	c_2	c_3
	b_1		d_1	y_{1111}	y_{1121}	y_{1131}	y_{2111}	y_{2121}	y_{2131}	y_{3111}	y_{3121}	y_{3131}
			d_2	y_{1112}	y_{1122}	y_{1132}	y_{2112}	y_{2122}	y_{2132}	y_{3112}	y_{3122}	y_{3132}
B	b_2	D	d_1	y_{1211}	y_{1221}	y_{1231}	y_{2211}	y_{2221}	y_{2231}	y_{3211}	y_{3221}	y_{3231}
			d_2	y_{1212}	y_{1222}	y_{1232}	y_{2212}	y_{2222}	y_{2232}	y_{3212}	y_{3222}	y_{3232}
	b_3		d_1	y_{1311}	y_{1321}	y_{1331}	y_{2311}	y_{2321}	y_{2331}	y_{3311}	y_{3321}	y_{3331}
			d_2	y_{1312}	y_{1322}	y_{1332}	y_{2312}	y_{2322}	y_{2332}	y_{3312}	y_{3322}	y_{3332}

3.6. Concluding Remarks

The interactions in table 3.3 differ in many significant cases. For example, at A=17.5, B=20.0, C=1900, and D=1, the statistical interaction is -42 whereas the *k*-

TABLE 3.2 Data of the Experiment

VARIABLE				A								
				7.5			12.5			17.5		
				C								
				1900	2000	2300	1900	2000	2300	1900	2000	2300
B	5.0	D	1	340	316	374	260	388	266	134	146	152
			2	375	386	350	244	304	234	140	194	212
	12.5		1	388	338	334	322	300	234	186	412	194
			2	370	214	366	342	420	258	30	428	208
	20.0		1	378	348	380	330	260	350	40	436	230
			2	378	378	398	298	366	284	210	490	254

system interaction is -60. In some other cases, the k-system interaction is about twice the statistical interaction. In yet others, the statistical interaction is twice the k-systems interaction. The statistical interactions are incorrect because they are based on the assumption of a linear model. The k-systems interactions represent a correct form for interactions because they assume no model; further, they measure the combination effect as an increase in information content due to the combination as measured by information theory.

Consider the equations which are solved in the analysis of variance. If we substitute the exact marginal summation for each of these equations, we get the marginal summations of k-systems analysis. That is, we begin with equations that are correct. Now, we solve for the system function by a maximum entropy algorithm which uses these equations as constraints. We have used a correct model, and an algorithm which introduces minimal extraneous information. By varying the constraint equations, we add interactions into the system function. This is done without assumptions on the form of the system function, and without simplifying assumptions in the solution of the marginal sums.

We conclude that there are significant differences between statistical and *k*-systems interactions, and that these differences are due to the erroneous model and simplifying assumptions of statistical interactions.

TABLE 3.3 Reconstructability Interactions vs Statistical Interactions

FACTOR	INTERACTION EFFECT (Reconstructability Analysis)	INTERACTION EFFECT (Statistical Analysis)
A=12.5 B=5.0 C=1900 D=1	-13.47	-15.26
A=12.5 B=5.0 C=1900 D=2	13.47	15.26
A=12.5 B=5.0 C=2000 D=1	32.80	35.96
A=12.5 B=5.0 C=2000 D=2	-32.80	-35.96
A=12.5 B=5.0 C=2300 D=1	-18.35	-20.70
A=12.5 B=5.0 C=2300 D=2	18.35	20.70
A=12.5 B=12.5 C=1900 D=1	-11.07	-6.26
A=12.5 B=12.5 C=1900 D=2	11.07	6.26
A=12.5 B=12.5 C=2000 D=1	-6.93	-12.87
A=12.5 B=12.5 C=2000 D=2	6.93	12.87
A=12.5 B=12.5 C=2300 D=1	15.97	19.13
A=12.5 B=12.5 C=2300 D=2	-15.97	-19.13
A=12.5 B=20.0 C=1900 D=1	25.20	21.52
A=12.5 B=20.0 C=1900 D=2	-25.20	-21.52
A=12.5 B=20.0 C=2000 D=1	-26.50	-23.09
A=12.5 B=20.0 C=2000 D=2	26.50	23.09
A=12.5 B=20.0 C=2300 D=1	1.06	1.57
A=12.5 B=20.0 C=2300 D=2	-1.06	-1.57
A=17.5 B=5.0 C=1900 D=1	10.75	11.85
A=17.5 B=5.0 C=1900 D=2	-10.75	-11.85
A=17.5 B=5.0 C=2000 D=1	-5.70	-3.93
A=17.5 B=5.0 C=2000 D=2	5.70	3.93
A=17.5 B=5.0 C=2300 D=1	-6.72	-7.93
A=17.5 B=5.0 C=2300 D=2	6.72	7.93
A=17.5 B=12.5 C=1900 D=1	52.00	30.52
A=17.5 B=12.5 C=1900 D=2	-52.00	-30.52
A=17.5 B=12.5 C=2000 D=1	-37.19	-24.09
A=17.5 B=12.5 C=2000 D=2	37.19	24.09
A=17.5 B=12.5 C=2300 D=1	-8.42	-6.43
A=17.5 B=12.5 C=2300 D=2	8.42	6.43
A=17.5 B=20.0 C=1900 D=1	-59.95	-42.37
A=17.5 B=20.0 C=1900 D=2	59.95	42.37
A=17.5 B=20.0 C=2000 D=1	40.05	28.02
A=17.5 B=20.0 C=2000 D=2	-40.05	-28.02
A=17.5 B=20.0 C=2300 D=1	16.81	14.35
A=17.5 B=20.0 C=2300 D=2	-16.81	-14.35

Chapter 4

COMPUTING
TRUE MAIN AND INTERACTION EFFECTS

4.1. Introduction

This chapter presents a new method to compute variable effects, called, main and interaction effects. The method could be used for any type of experimental design problems, but for the purpose of this chapter, we will use an example of fractional factorial design to discuss the superiority of this method over classical statistical method. With fractional factorial design, we could show the confounding problem and how we overcome this problem using this method.

In the next sections, we will see an example of fractional factorial design and its confounded effects computed using classical statistical method. Then, we will present how we will be able to get all the variable effects in the fractional factorial design using k-systems theory of the reconstructability analysis method.

4.2. An Example of Fractional Factorial Design

Before we go into the detail of an aspect of fractional factorial design, i.e. confounding, let us see a simple hypothetical example and its result. Table 4.1 shows the variables and their level assignments in a scheduler design study. There are four variables where each variable has two levels, low/high, denoted by - (low) and + (high). To investigate the effects of all possible combinations of all the levels of the variables, a full 2^4 factorial design with 16 experiments would have to be performed. Instead, a 2^{4-1} fractional factorial design with 8 experiments is used to study the relative importance of these variables. So, instead of running 16 experiments, we only run half of them.

TABLE 4.1 Variables and Levels in the Scheduler Design Study

SYMBOL	VARIABLE	LEVEL -	LEVEL +
A	Preemption	No	Yes
B	Time Slice	Small	Large
C	Queue Assignment	One Queue	Two Queues
D	Fairness	Off	On

Table 4.2 shows the level assignments for the four variables in the 8-experiment and the measured throughputs. The last column in the table denotes the throughputs for a particular workload.

TABLE 4.2 Measured Throughputs for Scheduler Design Study

EXPERIMENT NO.	A	B	C	D	y
1	-	-	-	-	25.0
2	+	-	-	+	41.0
3	-	+	-	+	36.0
4	+	+	-	-	16.0
5	-	-	+	+	64.0
6	+	-	+	-	13.0
7	-	+	+	-	36.0
8	+	+	+	+	23.0

For a two-level factorial design, the main and interaction effects can be computed easily by preparing an orthogonal sign table (Box et al. 1978; Mason et al. 1989; McLean and Anderson 1984; Montgomery 1991). For a more general factorial design, the main and interaction effects can be estimated using least squares method that will produce the same results. Using a sign table, a 2^{k-p} fractional factorial design may be constructed by writing down a *basic design* consisting of the experiments for a *full* 2^{k-p} factorial design and then adding the other p variables by identifying their plus and

minus levels with the plus and minus signs of the interaction effects. Therefore, the 2^{4-1} fractional factorial design is obtained by writing down the full 2^3 factorial design as the basic design and then equating the other variable to one of the interaction effects in the 2^3 design.

TABLE 4.3 A 2^3 Experimental Design

EXPERIMENT NO.	I	A	B	C	AB	AC	BC	ABC
1	+	-	-	-	+	+	+	-
2	+	+	-	-	-	-	+	+
3	+	-	+	-	-	+	-	+
4	+	+	+	-	+	-	-	-
5	+	-	-	+	+	-	-	+
6	+	+	-	+	-	+	-	-
7	+	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+	+

Table 4.3 shows a full 2^3 factorial design as the basic design for the 2^{4-1} fractional factorial design. The first column in the table is labeled I, and consists of all +s. It is used to compute the mean response / throughput. The next three columns, i.e. A, B, and C, contain all possible combinations of - and +. The rest four columns are all possible variable interactions among A, B, and C. AB is the product of the entries in columns A and B, AC is the product of the entries in columns A and C, and so on.

Table 4.4 presents a 2^{4-1} fractional factorial design obtained from table 4.3. From the four columns on the right in table 4.3, we pick the rightmost column and mark it D, the 4^{th} variable. Note that, this pick is based on the plus and minus levels in column D in table 4.2. If we selected other column, say column BC, to be marked as D, then the plus and minus levels for D in table 4.2 would need to be changed according to the plus and minus levels of BC. And certainly, the last column in table 4.2 would have

different values. The design in table 4.4 will allow us to compute main effects A, B, C, and D along with interaction effects AB, AC, and BC.

TABLE 4.4 A 2^{4-1} Experimental Design

EXPERIMENT NO.	I	A	B	C	AB	AC	BC	D
1	+	-	-	-	+	+	+	-
2	+	+	-	-	-	-	+	+
3	+	-	+	-	-	+	-	+
4	+	+	+	-	+	-	-	-
5	+	-	-	+	+	-	-	+
6	+	+	-	+	-	+	-	-
7	+	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+	+

The main and interaction effects can be computed by taking the inner product of y-column and various x-columns and then dividing the sum by 8, as shown in table 4.5. The main effects of variable A through D and the two-variable interaction effects AB, AC, and BC are given in the last line of the table. The design presented in table 4.5 is based on an assumption that the data fit the following linear model:

$$y = \mu + q_A x_A + q_B x_B + q_C x_C + q_D x_D + q_{AB} x_{AB} + q_{AC} x_{AC} + q_{BC} x_{BC}$$

where μ is the overall mean; q_A, q_B, q_C, and q_D are the main effects; and q_{AB}, q_{AC}, and q_{BC} are the two-variable interaction effects.

The result is interpreted as follows. The average throughput of an average scheduler for a particular job is 31.75. The important effects that impact the throughput are A (Preemption), D (Fairness), and AC (interaction between Preemption and Queue Assignment). Variables B (Time Slice) and C (Queue Assignment) and their interaction do not have significant impacts on the throughput. Although C is not an important

variable, the interaction between C and A contributes a significant impact on the throughput. Preemption can decrease or increase the average throughput by 8.50. Fairness, on the other hand, can increase or decrease the average throughput by 9.25. In the mean time, the interaction between Preemption and Queue Assignment might affect the average throughput by 7.50.

TABLE 4.5 Data and Effects in the Scheduler Design Study

EXPERIMENT NO.	I	A	B	C	AB	AC	BC	D	y
1	+	-	-	-	+	+	+	-	25.0
2	+	+	-	-	-	-	+	+	41.0
3	+	-	+	-	-	+	-	+	36.0
4	+	+	+	-	+	-	-	-	16.0
5	+	-	-	+	+	-	-	+	64.0
6	+	+	-	+	-	+	-	-	13.0
7	+	-	+	+	-	-	+	-	36.0
8	+	+	+	+	+	+	+	+	23.0
	254.00	-68.00	-32.00	18.00	2.00	-60.00	-4.00	74.00	Total
	31.75	-8.50	-4.00	2.25	0.25	-7.50	-0.50	9.25	Effect

4.3. Confounding

As mentioned earlier, there is one problem with fractional factorial designs. Some variable effects cannot be determined, only the combined influence of two or more effects can be computed. This problem is known as *confounding*, and the effects whose influence cannot be distinguished from each other are said to be *confounded*.

Now, let us see again the example in the previous section, the 2^{4-1} fractional factorial design. There are four variables: A, B, C, and D; and each has two levels. Out of 16 possible combinations or experiments, we take only the following 8 for investigation:

$$(----,\ +--+,\ -+-+,\ ++--,\ --++,\ +-+-,\ -++-,\ ++++)$$

The relation between the responses from these combinations and the main and interaction effects can be presented as shown in table 4.6. We can see from table 4.6, the same observational contrasts (the plus and minus signs) estimates both μ and four-variable interaction q_{ABCD}, where μ denotes the mean response. Similarly, the same observational contrasts estimates both q_A and q_{BCD}. Also, the same observational contrasts estimates both q_D and q_{ABC}. And so on.

Let us consider main effect q_D and three-variable interaction effect q_{ABC}. If y_i represents the response value in the i^{th} experiment, then the effect of D can be obtained by taking the inner product of column D and column y and dividing the sum by 8. This gives

$$q_D = \sum_i \frac{y_i x_{Di}}{8} = \frac{-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8}{8} .$$

The effect of variable interaction ABC is obtained by multiplying the respective elements of columns A, B, C, and y. This gives

$$q_{ABC} = \sum_i \frac{y_i x_{Ai} x_{Bi} x_{Ci}}{8} = \frac{-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8}{8} .$$

Notice that the expression for q_{ABC} is identical to that for q_D. In fact, the expression is neither q_D nor q_{ABC}; it is the sum of the two:

$$q_D + q_{ABC} = \sum_i \frac{y_i x_{Ai} x_{Bi} x_{Ci}}{8} = \frac{-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8}{8} .$$

Without a full factorial design, it is not possible to separate the estimate of q_D from q_{ABC}. In other words, D is completely confounded with ABC. This confounding can be denoted as

$$\mathrm{D} \equiv \mathrm{ABC} .$$

TABLE 4.6 Relation Between Responses and Effects in a Fraction of a 2^4 Factorial Design

LEVEL COMBINATION	EFFECTS															
	μ	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
−−−−	+	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
+−−+	+	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
−+−+	+	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
++−−	+	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
−−++	+	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+
+−+−	+	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
−++−	+	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
++++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

In this design, D and ABC effects are not the only effects that are confounded. It is easy to see that A and BCD effects are also confounded.

$$q_A + q_{BCD} = \sum_i \frac{y_i x_{Ai}}{8} = \frac{-y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8}{8}$$

that is,

$$A \equiv BCD\,.$$

In fact, every column in the design represents a sum of 2 effects. The complete list of confoundings in this fractional factorial design is as follows:

$$\begin{array}{llll} I \equiv ABCD, & A \equiv BCD, & B \equiv ACD, & C \equiv ABD, \\ D \equiv ABC, & AB \equiv CD, & AC \equiv BD, & AD \equiv BC \end{array}$$

where I represents μ, the mean response.

Thus, the complete estimates of the main and interaction effects in the 2^{4-1} fractional factorial design above can be summarized as follows:

$$\begin{array}{lclcr} \mu & + & q_{ABCD} & = & 31.75 \\ q_A & + & q_{BCD} & = & -8.50 \\ q_B & + & q_{ACD} & = & -4.00 \\ q_C & + & q_{ABD} & = & 2.25 \\ q_D & + & q_{ABC} & = & 9.25 \\ q_{AB} & + & q_{CD} & = & 0.25 \\ q_{AC} & + & q_{BD} & = & -7.50 \\ q_{AD} & + & q_{BC} & = & -0.50 \end{array}$$

Please note that a fractional factorial design is not unique. For the same number of variables k and the same number of experiments 2^{k-p}, there are 2^p possible different fractional factorial designs.

4.4. *K*-systems Main and Interaction Effects

Now, we will see how to compute main and interaction effects using the concept we have mentioned in the previous chapter. First, we present interaction effects and then main effects after that. Main effects can be computed using the same concept as that for computing interaction effects.

Let $V = \{v_1 = c_1, v_2 = c_2, \ldots, v_n = c_n\}$ be a set of n variables with a particular value assignment (a factor or substate); let V_m be the set of all nonempty subsets of V; and let V_p denotes the set of all nonempty *proper* subsets of V. We compute the unbiased reconstructions $\mu(V_m)$ and $\mu(V_p)$. Then, for any factor which embodies V, we consider the difference in its effect as computed from these two unbiased reconstructions. This difference represents the contribution to a factor effect that is due to the components acting in unison - the n-variable interaction.

A main effect can be considered as a 1-variable interaction, interaction between the variable and the flat system. To find a main effect, V_m would be the set of all nonempty subsets of V where V is a set of 1-variable substate, and V_p would be an empty set that represents the flat system. Then, we compute the unbiased reconstruction $\mu(V_m)$ while $\mu(V_p)$ is the flat system. The main effect can be obtained from the difference of the function values of the two unbiased reconstructions.

To illustrate this, let us see how to compute interaction effect ABD where A is at the low level, B at the high level, and D at the high level. If X_- represents variable X at the low (-) level and X_+ represents variable X at the high (+) level, then the interaction effect ABD above can be denoted as ABD_{-++}.

The set V is written as $V = \{A_-, B_+, D_+\}$. The set of all nonempty subsets of V is given by

$$V_m = \{\{A_-\}, \{B_+\}, \{D_+\}, \{A_-, B_+\}, \{A_-, D_+\}, \{B_+, D_+\}, \{A_-, B_+, D_+\}\}$$

and the set of all nonempty proper subsets of V is defined by

$$V_p = \{\{A_-\}, \{B_+\}, \{D_+\}, \{A_-, B_+\}, \{A_-, D_+\}, \{B_+, D_+\}\}.$$

The unbiased reconstructions $\mu(V_p)$ and $\mu(V_m)$ are given in table 4.7. f(−−−−) denotes the reconstructed system function for state (A_-, B_-, C_-, D_-), f(−−−+) for state (A_-, B_-, C_-, D_+), f(−−+−) for state (A_-, B_-, C_+, D_-), and so on. Each value in the table represents a reconstructed system function value for the set of substates V_p or V_m.

TABLE 4.7 The Unbiased Reconstructions for Computing ABD_{-++}

	$\mu(V_p)$	$\mu(V_m)$
f(−−−−)	29.03	28.38
f(−−−+)	47.22	47.88
f(−−+−)	29.03	28.38
f(−−++)	47.22	47.88
f(−+−−)	33.22	33.88
f(−+−+)	34.53	33.88
f(−++−)	33.22	33.88
f(−+++)	34.53	33.88
f(+−−−)	21.72	22.38
f(+−−+)	37.03	36.38
f(+−+−)	21.72	22.38
f(+−++)	37.03	36.38
f(++−−)	24.53	23.88
f(++−+)	26.72	27.38
f(+++−)	24.53	23.88
f(++++)	26.72	27.38

Interaction effect ABD_{-++} is computed as follows:

$$ABD_{-++} = \frac{(33.88 - 34.53) + (33.88 - 34.53)}{2} = -0.65\,.$$

Similarly, interaction effects ABD_{---}, ABD_{--+}, ABD_{-+-}, ABD_{+--}, ABD_{+-+}, ABD_{++-}, and ABD_{+++} could be found the same way. They are -0.65, 0.65, 0.65, 0.65, -0.65, -0.65, and 0.65 respectively.

To compute a main effect, say B_-, the set V is defined as $V = \{B_-\}$. The set of all nonempty subsets of V and the set of all nonempty proper subsets of V are $V_m = \{\{B_-\}\}$ and $V_p = \{\ \}$ respectively. Table 4.8 presents the unbiased reconstructions for V_p and V_m. Column $\mu(V_p)$ in that table represents the system function values for the flat system.

TABLE 4.8 The Unbiased Reconstructions for Computing B_-

	$\mu(V_p)$	$\mu(V_m)$
f(− − − −)	31.75	33.75
f(− − − +)	31.75	33.75
f(− − + −)	31.75	33.75
f(− − + +)	31.75	33.75
f(− + − −)	31.75	29.75
f(− + − +)	31.75	29.75
f(− + + −)	31.75	29.75
f(− + + +)	31.75	29.75
f(+ − − −)	31.75	33.75
f(+ − − +)	31.75	33.75
f(+ − + −)	31.75	33.75
f(+ − + +)	31.75	33.75
f(+ + − −)	31.75	29.75
f(+ + − +)	31.75	29.75
f(+ + + −)	31.75	29.75
f(+ + + +)	31.75	29.75

Main effect B_- is obtained as follows:

$$B_- = \frac{8 \times (33.75 - 31.75)}{8} = 2.00 .$$

Main effect B_+ could be found similarly, and its value is -2.00.

All other main and interaction effects of the above 2^{4-1} fractional factorial design are computed the same way, and could be summarized as follows:

$$
\begin{aligned}
\mu &= 31.75 \\
q_A &= -4.25 \\
q_B &= -2.00 \\
q_C &= 1.13 \\
q_D &= 4.63 \\
q_{AB} &= -0.14 \\
q_{AC} &= -3.60 \\
q_{AD} &= 0.37 \\
q_{BC} &= -0.18 \\
q_{BD} &= -3.46 \\
q_{CD} &= -0.04 \\
q_{ABC} &= 4.37 \\
q_{ABD} &= 0.65 \\
q_{ACD} &= -1.46 \\
q_{BCD} &= -4.09 \\
q_{ABCD} &= -1.65
\end{aligned}
$$

The above main and interaction effects are taken when each variable in the substate is at the high (+) level. Appendix B gives the complete effects computed using this method.

Now, we can write the algorithm to find all the main and interaction effects of a design of experiments using k-systems theory.

1) Preparation steps.
   This includes averaging redundant system function values (or replicated response values), entropy fill missing system function values, and mapping the original system into a k-system.
2) Generate S, the set of all substates (nonempty subsets of the variables at all levels).
3) For every substate V in S,
   a) Let V_m be the set of all nonempty subsets of V and
      V_p be the set of all nonempty proper subsets of V.
   b) Compute the unbiased reconstructions $\mu(V_m)$ and $\mu(V_p)$ (express them in g-system). In the case that V_p is empty, $\mu(V_p)$ is the flat system.
   c) Compute the main/interaction effect.
      For any factor which embodies V, take the difference in its effect as computed from these two unbiased reconstructions. This difference represents the main or interaction effect.

4.5. Concluding Remarks

The concept of k-systems theory has opened a door for a new approach to experimental design problems. The employment of substates in k-system is a radical

departure from the classical statistical approach with variables. Substates allow us to examine system dynamics at its most refined level. We may expect to find information that are hidden in analysis where variables are the fundamental unit.

When using classical statistical method for a fractional factorial design, an assumption regarding the absence of certain interaction effects has to be made in order to obtain the unbiased estimates of other effects. This could be a drawback of classical statistical method since sometimes it is unrealistic to assume an absence of certain interaction effects. As we have seen in the previous design, although a certain variable might not be important in the design, the interaction effect between that variable and other variable could be significant enough for the overall design. Also, not all the effects in a fractional factorial design can be determined. Only the combined influence of two or more effects can be computed. This could cause loss of information and eventually could lead us to a wrong conclusion.

On the other hand, with k-systems theory, we do not need to make any assumption regarding the absence of certain effects. No loss of information happens, all main and interaction effects can be computed. Interaction effects, as well as main effects, are computed from the existing data and the results are true for the given information.

K-systems theory is a general methodology and can be used for any type of experimental design, not only fractional factorial design. This could eliminate the need for different types of experimental designs with all kind of their unique characteristics and restrictions. Also, it can be easily used for design with unusual number of experiments, for example, a design to investigate four two-level variables with 11 experiments or even 5 experiments which is very difficult to analyze with a classical statistical method.

Another advantage of k-systems theory over classical statistical method is that k-systems theory does not make any assumption on the structure of the system (data). It assumes no structure that does not explicitly exist. The analysis on a k-system will never modify information in any way and solution properties depend only on existing

information. On the other hand, classical statistical methods assume some kind of model on the data. For example, factorial experiments require that certain assumptions be satisfied. These assumptions are that the data are adequately described by a certain linear model and that the errors are normally and independently distributed (Montgomery 1991; Petersen 1985; Raktoe et al. 1981; Scheffe' 1959; Snedecor and Cochran 1967). Any type of model inadequacy and violations of the underlying assumptions will make the analysis invalid.

Chapter 5

A CASE STUDY

5.1. Introduction

Data analysis has emerged as an important discipline in statistics which emphasizes formal structure of the system of inference that often requires prespecifying many things such as the hypothesis to be tested and the model associated with the data. This chapter describes an in-depth analysis of a set of real experimentation data using a more practical approach which emphasizes more on the descriptive nature of the approach.

Our primary goal is to show how data analysis would be done using k-systems theory. Since k-systems theory works directly with existing data and does not make any assumption on the structure of data; the results obtained from the analysis using this method are true for the given data (Gouw and Jones in press2), and at the same time, the analysis is also simpler than the analysis using traditional statistical method.

In this context, we present an analysis using a set of data collected from a design of experiments at LSU Plant Pathology Department. The primary objective of the analysis is to search for the significant patterns in the data and the relationships among variables in the experiments.

5.2. The Data

The data consist of figures of dry weights of sorghum obtained from a design of experiments which was conducted in two consecutive years. Each year, a 3x3x2 design of experiments with four replications was performed. The dry weights are broken down into three parts, i.e., root dry weights, stalk dry weights, and head dry weights. It is important to separate them that way since each part has its own characteristics and might be subject to different variables in the experiments.

TABLE 5.1 Variables and Levels in the Sorghum Experiments

SYMBOL	VARIABLE	LEVEL 1	LEVEL 2	LEVEL 3
M	M. Phaseolina	0 cfu/g soil	10 cfu/g soil	100 cfu/g soil
N	Nematode	0 nematode	1554 nem./16-kg soil	3108 nem./16-kg soil
H	Hybrid	Dekalb-Pfizer 50	Pioneer 8333	

There are three variables under investigation in the experiments: *macrophomina phaseolina*, *nematode*, and *hybrid*. Macrophomina phaseolina (M. phaseolina) is a type of soil borne fungus that rots the root and stalk of sorghum. Nematode destroys root cells and reduces uptake of water and nutrients into the plant. Severely infected plants by these pathogens are smaller as the result of extensive root dysfunction (Mughogho and Pande 1984). The effects of M. phaseolina and nematode were investigated for two different sorghum hybrids. Table 5.1 provides the details of the variables under study and their corresponding levels.

Tables 5.2 and 5.3 show the dry weights data for the first and the second year respectively. The tables show all possible combinations of levels of variables under study. The last three columns show the root, stalk, and head dry weights; and each cell contains the four replications for a level combination. A "--" entry in table 5.3 indicates a missing data for a particular level combination at a particular replication.

A problem has been passed over in arriving at these data. Actually, the original dry weights data were prepared from a three-year experiments. So, in addition to the data in table 5.2 and table 5.3 above, we have another table for the earlier year (year-0) experiments. But after considerable preliminary analysis, it was found that these data were very much different from the other two. For some reasons, the dry weights data from year-0 were considerably smaller than the ones obtained from the subsequent two years. The analysis including year-0 data would have masked the effects of the

TABLE 5.2 Measured Dry Weights for Year-1

Var. Levels M	N	H	Root Dry Weight (grams)	Stalk Dry Weight (grams)	Head Dry Weight (grams)
1	1	1	(119, 98, 100, 102)	(56, 61, 54, 58)	(63, 50, 55, 54)
1	1	2	(121, 109, 130, 112)	(57, 62, 43, 50)	(55, 52, 56, 54)
1	2	1	(92, 94, 99, 93)	(47, 48, 44, 47)	(45, 40, 42, 46)
1	2	2	(92, 94, 93, 93)	(50, 44, 49, 47)	(31, 45, 40, 44)
1	3	1	(88, 91, 89, 90)	(49, 49, 46, 49)	(32, 40, 40, 43)
1	3	2	(89, 92, 84, 88)	(50, 42, 46, 46)	(30, 38, 42, 32)
2	1	1	(79, 90, 82, 77)	(46, 46, 44, 45)	(31, 34, 41, 26)
2	1	2	(78, 88, 76, 78)	(46, 46, 45, 46)	(15, 28, 24, 25)
2	2	1	(69, 80, 69, 70)	(45, 45, 44, 47)	(21, 29, 22, 22)
2	2	2	(70, 78, 63, 73)	(47, 47, 45, 45)	(32, 24, 23, 23)
2	3	1	(64, 70, 65, 67)	(42, 46, 40, 47)	(26, 26, 19, 20)
2	3	2	(64, 70, 64, 66)	(46, 40, 44, 45)	(24, 24, 20, 21)
3	1	1	(63, 66, 64, 65)	(47, 44, 46, 47)	(22, 21, 21, 20)
3	1	2	(64, 63, 63, 66)	(48, 45, 45, 44)	(23, 22, 22, 22)
3	2	1	(66, 65, 64, 64)	(44, 45, 46, 46)	(22, 21, 22, 23)
3	2	2	(63, 64, 65, 64)	(45, 44, 44, 45)	(20, 22, 25, 19)
3	3	1	(64, 66, 64, 64)	(46, 47, 40, 46)	(20, 21, 22, 22)
3	3	2	(62, 62, 63, 60)	(44, 46, 45, 44)	(20, 20, 23, 19)

TABLE 5.3 Measured Dry Weights for Year-2

Var. Levels M	N	H	Root Dry Weight (grams)	Stalk Dry Weight (grams)	Head Dry Weight (grams)
1	1	1	(122, 146, 85, 114)	(141, 145, 91, 112)	(59, 57, 59, 73)
1	1	2	(–, 59, 56, 59)	(99, 71, 90, 111)	(–, 51, 69, 70)
1	2	1	(145, 65, 89, 128)	(129, 108, 147, 171)	(–, 71, 80, 121)
1	2	2	(43, 40, 79, 95)	(54, 88, –, 121)	(36, 51, 64, 122)
1	3	1	(115, 81, 75, 107)	(97, 119, 119, 144)	(46, 55, 61, 83)
1	3	2	(140, 146, 57, 66)	(89, 97, –, 106)	(61, 59, 50, 119)
2	1	1	(80, 50, 42, 71)	(155, 121, 113, –)	(96, 62, 81, 50)
2	1	2	(104, 47, 34, 56)	(104, 75, 80, 70)	(52, 57, 69, 50)
2	2	1	(55, 60, 65, 60)	(77, 75, 139, 133)	(60, 71, 84, 108)
2	2	2	(49, 50, 63, 41)	(66, 76, 88, 64)	(54, 48, 56, 49)
2	3	1	(117, 54, 41, 89)	(52, 171, 90, 119)	(59, 58, 51, 78)
2	3	2	(49, 43, 43, 68)	(63, 77, 60, 88)	(47, 56, 37, 53)
3	1	1	(38, 59, 55, 65)	(94, 79, 88, 113)	(73, 76, 39, 55)
3	1	2	(66, 51, –, 38)	(95, 74, –, 90)	(81, 35, –, 51)
3	2	1	(51, 74, 60, 36)	(156, 128, 122, 152)	(25, 78, 60, 108)
3	2	2	(50, 50, 44, 56)	(86, 66, 67, 96)	(38, 50, 54, 65)
3	3	1	(35, 49, 63, 86)	(110, 93, 94, 116)	(60, 46, 42, 61)
3	3	2	(42, 32, 67, 84)	(74, 70, 76, 102)	(46, 49, 57, 66)

variables under study since the *year* would appear as the dominant factor instead of the three variables in the experiments. Therefore, the year-0 data were thrown out. Although, year-1 and year-2 data are not very close either, they are a lot closer than year-0 data, and hence are used here to illustrate this study.

5.3. The Model?

In any data analysis, usually we want to know the patterns that exist in the data. This includes the significant variables that build the data and how much influence they have on the overall structure of the data. The most common solution is to compute the variable effects and the effects resulted from the interactions among variables. They are called main and interaction effects respectively. By knowing these effects, we could find out the significant variables that form the data and how they influence the data.

K-systems theory computes these main and interaction effects using the power of information theory by invoking the maximum entropy principle. Unlike traditional statistical method which assumes a certain model associated with the data, k-systems theory assumes no model for the data. It assumes no structure that does not explicitly exist (Jones 1985c,d). It works directly with whatever information is available in the data. Thus, the analysis on a k-system will never modify information in any way and solution properties depend only on existing information.

An n-variable interaction in the context of k-systems theory represents an effect due to the n variables acting in unison. It measures the combination effect as an increase in information content as measured by information theory (Gouw and Jones in press1). A main effect, as has been shown in chapter 4, is a special case of n-variable interaction. It can be considered as a 1-variable interaction, interaction between the variable and the flat or average system.

5.4. Results and Analysis

We have performed the analysis in several stages. First, we check the replication. Then, the year. And finally, the variables under study. But before we go further into the analysis, it is appropriate at this time to discuss the tool that we use to measure the importance of an effect. As mentioned earlier, the bulk of the analysis is intended to be a practical one which emphasizes the descriptive nature of the approach. In this spirit, we find that computing the percentages of variation of the effects is very useful in determining the significant effects.

5.4.1. Percentage of Variation

Virtually, all experimental (and also observational) data are subject to a variety of sources which induce variation in the data. This variation could occur because of differences on variables or external variables in the experiment, and also differences due to measurement errors. Measurement error deals with data collection process, and hence will not be discussed here.

The importance of a variable is measured by the proportion of the total variation in the data that is explained by the variable. Thus, if two variables explain 80% and 5% of the total variation in the data, the second variable may be considered insignificant in many practical situations.

Suppose we have 3 variables A, B, and C and each has a, b, and c levels respectively. Let α_i be the effect of A at level i, β_j be the effect of B at level j, γ_k be the effect of C at level k, $\alpha\beta_{ij}$ be the interaction between A and B at levels i and j, ..., and $\alpha\beta\gamma_{ijk}$ be the interaction between A, B, C at levels i, j, k. Then, the variation explained by each factor (variable or variable combination) can be summarized as follows (Mason et al. 1989):

Variation explained by A, $V_A = bc \sum_{i=1}^{a} \alpha_i^2$

Variation explained by B, $V_B = ac \sum_{j=1}^{b} \beta_j^2$

Variation explained by C, $V_C = ab \sum_{k=1}^{c} \gamma_k^2$

Variation explained by AB, $V_{AB} = c \sum_{i=1}^{a} \sum_{j=1}^{b} \alpha\beta_{ij}^2$

Variation explained by AC, $V_{AC} = b \sum_{i=1}^{a} \sum_{k=1}^{c} \alpha\gamma_{ik}^2$

Variation explained by BC, $V_{BC} = a \sum_{j=1}^{b} \sum_{k=1}^{c} \beta\gamma_{jk}^2$

Variation explained by ABC, $V_{ABC} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \alpha\beta\gamma_{ijk}^2$.

The total variation in the data is the sum of variations explained by each factor above.

Total variation, $V_{TOTAL} = V_A + V_B + V_C + V_{AB} + V_{AC} + V_{BC} + V_{ABC}$.

When a variation explained by a factor is expressed as percentage of total variation, this proportion provides an easy way to gauge the importance of a factor. The factors which explain a high percentage of variation are considered important.

5.4.2. Checking the Replication

Each year, a 3x3x2 experiment with four replications was performed. Each replication was prepared in a separate block in the field. Although the experimenter has tried to keep all the blocks as uniform as possible, they may still be subject to different levels of external factors such as watering and sunlight exposure. As we know, in a good experimental design, all factors other than the variables under study should be

made as negligible as possible, so that the experimental results are really due to the variables being experimented, not because of the external factors. Hence, it is of our interest to check this replication to assure the validity of the analysis.

In order to do this, the 3x3x2 experiment with four replications is analyzed as a 4x3x3x2 experiment without replication, where the first variable is the block or replication. Table 5.4 provides the variations explained for root dry weight, stalk dry weight, and head dry weight for year-1 and year-2.

TABLE 5.4 Variation Explained in the 4x3x3x2 Experiments

FACTOR	PERCENTAGE OF VARIATION					
	YEAR-1			YEAR-2		
	RDW	SDW	HDW	RDW	SDW	HDW
B	0.4	5.7	0.4	5.6	5.4	18.0
M	78.8	30.6	77.7	33.9	6.9	5.9
N	10.9	15.1	8.4	1.5	1.3	3.6
H	0.0	0.6	0.6	6.9	36.3	5.1
BM	1.9	2.2	0.9	5.9	4.5	12.2
BN	0.4	3.3	0.4	3.5	11.8	24.3
BH	0.0	3.0	0.2	0.4	0.3	0.6
MN	3.4	15.3	3.0	0.1	3.7	2.1
MH	0.5	1.3	0.2	1.9	0.4	3.9
NH	0.5	0.8	0.2	1.6	2.6	4.3
BMN	1.2	9.2	3.0	15.4	13.2	6.0
BMH	0.3	1.1	0.8	1.6	1.9	4.7
BNH	0.3	4.9	0.7	5.1	2.9	2.9
MNH	0.7	1.5	1.7	7.4	2.7	0.5
BMNH	0.7	5.6	1.8	9.4	6.0	5.6

Now, how to read and interpret this result? Let us consider the variation for factor B (block or replication) for root dry weight in year-1. This number says that from all the variability that exists in the data, 0.4% of them are due to the replication factor. As we mentioned earlier, the effects of external factors other than the variables under study should be kept as minimum as possible. This number is very small and negligible. Thus, this indicates that the experiment in this year was conducted quite successfully.

The percentages of variation for B in the first year are much smaller than the ones in the second year. It means the blocks in the first year are much more uniform than the blocks in the second year. The blocks or replications in year-1 do not contribute a lot to the variability in the data. In average (ignoring the variations contributed by B in 2, 3, and 4-variable interactions involving B), they cause about 2.2% of the variability in the data.

On the other hand, the blocks in year-2 have quite an influence to the variability that exists in the data, especially when it comes to the head dry weight. In average they contribute about 9.7% to the variability in the data. When considering the variations contributed by B in 2, 3, and 4-variable interactions as well, the contribution of B to the variability in the data would be much more than 9.7%. The exact number would be impossible to compute.

This shows that the experiment in year-2 was not as successful as the one conducted in year-1, since some external factors other than the three variables being experimented were affecting the outcome of the experiment. This would raise a validity question to the analysis of year-2 data.

5.4.3. Checking the Year

The next factor that we want to check here is the *year*. Since the experiments were conducted in two different years, then we feel necessary to investigate the year factor to the results of the experiments. Due to different weather conditions from year to year, the outcome of the experiments almost certainly are different. The question is how much difference? If the difference is reasonably small then the analysis from the two years experiments could be combined into one, otherwise it would be wise to do separate analysis.

In order to do this, the two 3x3x2 experiments with four replications are analyzed as a 3x3x2x2 experiment with four replications, where the last variable is the year.

Table 5.5 shows the variations explained for root dry weight, stalk dry weight, and head dry weight for the combined data.

TABLE 5.5 Variation Explained in the 3x3x2x2 Experiments with Four Replications

FACTOR	PERCENTAGE OF VARIATION		
	RDW	SDW	HDW
M	63.1	2.1	14.7
N	1.5	0.4	2.1
H	3.8	8.6	2.1
Y	7.1	81.9	67.4
MN	0.6	1.0	0.2
MH	0.7	0.1	1.1
MY	1.1	0.0	7.0
NH	0.5	0.5	1.1
NY	4.0	0.2	1.6
HY	4.8	3.1	0.2
MNH	3.0	0.8	0.2
MNY	0.7	0.7	1.3
MHY	2.0	0.1	0.2
NHY	1.4	0.3	0.7
MNHY	5.7	0.3	0.3

It is interesting to see that the stalk dry weight and head dry weight data from year-1 are very different from the ones obtained from year-2, while it is not the case with root dry weight. From all the variability that exists in stalk dry weight data, at least 81.9% of them are caused by the year factor; while for head dry weight, the year contributes at least 67.4% to the variability in the data. On the other hand, only 7.1% of the variability in the root dry weight data are contributed by the year factor. This strange phenomena could be induced from the fact that stalk and head of sorghum are above the ground and hence are more susceptible to weather condition, while the root is under the ground and is less affected by different weather conditions.

Thus, we could safely conclude that the data from these two different years of experiments, in general, are very different. Weather conditions in the first and the

second year are not similar; and they, in turn, would create different growth rates of sorghum in the experiments.

With this knowledge, it would be imprudent to make an analysis using the combined data, since the outcome of such analysis would be dominated by the year factor. The year would appear as the significant factor and mask the effects of the three variables under study, which is not what we want to happen. Thus following this analysis, the next analysis should be concentrated on two 3x3x2 experiments with four replications. They should be analyzed separately, year by year.

5.4.4. Analyzing the Variables in the Experiment

There are three variables that were being controlled in the experiment. They are *M. phaseolina*, *nematode*, and *hybrid*. In addition, we also have at least two external factors, *block* and *year*. Of these two, year seems to be the most dominant factor in the data. It masks the effects of other variables. This indicates that year-1 and year-2 data are very much different and can not be combined together for analysis. Therefore, the analysis should be done separately.

TABLE 5.6 Variation Explained in the 3x3x2 Experiments with Four Replications

FACTOR	PERCENTAGE OF VARIATION					
	YEAR-1			YEAR-2		
	RDW	SDW	HDW	RDW	SDW	HDW
M	83.1	47.0	84.7	61.3	9.7	23.9
N	11.5	23.3	9.2	3.0	2.0	16.0
H	0.0	0.9	0.7	14.1	68.6	20.8
MN	3.6	23.4	3.2	0.3	8.7	8.4
MH	0.5	1.9	0.2	4.1	1.1	11.5
NH	0.5	1.2	0.2	3.1	4.6	17.5
MNH	0.8	2.3	1.8	14.1	5.4	1.9

We analyze the data twice, once for each year. Thus, the 3x3x2 experiments of year-1 and year-2 are analyzed individually. Table 5.6 presents the variation explained by each factor for root, stalk, and head dry weight data of year-1 and year-2.

Analysis of Year-1 Data

From year-1 data, it appears that M. phaseolina is the most dominant factor in the data. It contributes at least 83.1%, 47.0%, and 84.7% to the variability in the root, stalk, and head dry weight data respectively. Nematode comes a distant second and explains 11.5%, 23.3%, and 9.2% of all the variability in the root, stalk, and head dry weight data respectively. Meanwhile, hybrid does not have too much influence on the outcome of the experiments. The two and three-variable interactions are negligible and could be ignored, with the exception of the two-variable interaction between M. phaseolina and nematode which explains 23.4% of all the variability in the stalk dry weight data.

Table 5.7 provides a look at the average weights and the details of the individual main effects and an important interaction effect for year-1 data. Other two and three-variable interaction effects are negligible and hence are not reproduced in the table.

Let us consider the root dry weight. The average weight is 78.3 grams. The main effect M1 (M. phaseolina at level 1) indicates that without the existence of M. phaseolina in the soil, the average dry weight of roots in the experiment could increase by 19.7 grams or 25.2% of the average. Meanwhile, adding 10 cfu M. phaseolina per gram soil (M. phaseolina at level 2) decreases the average weight by 5.4 grams or 6.8%. Similarly, adding 100 cfu M. phaseolina per gram soil (M. phaseolina at level 3) decreases the average weight even more, by 14.4 grams or 18.3%.

For stalk dry weight, the average weight is 46.6 grams. Having 10 cfu M. phaseolina per gram soil or 100 cfu M. phaseolina per gram soil does not seem to be different. In fact, having 100 cfu M. phaseolina in every gram of soil decreases the average dry weight less than having 10 cfu M. phaseolina per gram soil does.

TABLE 5.7 Main and Interaction Effects for Year-1 Data

FACTOR	Root Dry Weight		Stalk Dry Weight		Head Dry Weight	
	EFFECT	% AVERAGE	EFFECT	% AVERAGE	EFFECT	% AVERAGE
Avg	78.3		46.6		30.3	
M1	19.7	25.2	3.1	6.7	14.2	46.9
M2	-5.4	-6.8	-1.7	-3.5	-5.3	-17.5
M3	-14.4	-18.3	-1.5	-3.2	-8.9	-29.4
N1	7.3	9.3	2.2	4.7	4.5	14.9
N2	-1.7	-2.2	-0.8	-1.7	-1.0	-3.4
N3	-5.5	-7.1	-1.4	-3.0	-3.5	-11.5
H1	-0.2	-0.2	0.3	0.7	0.9	3.0
H2	0.2	0.2	-0.3	-0.7	-0.9	-3.0
MN11			3.0	6.5		
MN12			-1.9	-4.1		
MN13			-1.1	-2.4		
MN21			-1.6	-3.4		
MN22			1.4	3.0		
MN23			0.1	0.3		
MN31			-1.5	-3.2		
MN32			0.5	1.1		
MN33			1.0	2.1		

MN-interaction effects for stalk dry weight reveals a strange relationship between M. phaseolina and nematode. Having M. phaseolina or nematode alone in the soil would decrease the average dry weight of stalks. For example, see MN21, the interaction between M. phaseolina at level 2 and nematode at level 1. The average weight is decreased by 1.6 grams (3.4%) when we have 10 cfu M. phaseolina per gram soil and no nematode at all. On the other hand, having M. phaseolina and nematode together would increase, instead of decrease, the average weight. For example, having 10 cfu M. phaseolina per gram soil and 1554 nematodes per 16 kg soil together (interaction MN22) would increase the average weight by 1.4 grams (3.0%).

The average dry weight for head is 30.3 grams. The main effects for head dry weight is dominated by M. phaseolina. Having no M. phaseolina at all in the soil would

increase the average dry weight by 14.2 grams (46.9%). Ten cfu M. phaseolina per gram soil would decrease the average weight by 5.3 grams (17.5%). Meanwhile, having 100 cfu M. phaseolina per gram soil (ten times 10 cfu M. phaseolina per gram soil) decreases the average weight only by 8.9 grams (29.4%).

Analysis of Year-2 Data

From table 5.6, we can see that M. phaseolina is the dominant factor in the root dry weight data which explains at least 61.3% of all the variability in the data. Hybrid and three-variable interaction M. phaseolina-Nematode-Hybrid contribute 14.1% each to the total variability in the root dry weight data. Unlike in year-1, nematode does not have too much impact on the variability of the root dry weight data. If we pay more attention to the analysis, we can see that this is an example of a system where its two-variable interactions are not really significant, instead the three-variable interaction is more important to the system.

For stalk dry weight, the dominant factor is hybrid which explains at least 68.8% of the variability in the data. M. phaseolina and nematode contribute only 9.7% and 2.0% respectively to the variability in the data. This result is surprisingly different from the one obtained from year-1. The difference could have been resulted from the difference in weather conditions between year-1 and year-2. The weather condition in year-2 might have been too inconvenient for M. phaseolina and nematode to live or to do damaging activities to the plant.

For head dry weight, it seems all variables have about the same shares to the variability in the data. Two-variable interactions M. phaseolina-Hybrid and Nematode-Hybrid are also worth considering. Both of them involve hybrid. Once again, it looks like hybrid is a significant factor in this year-2 data. This result is also different from the result for year-1.

TABLE 5.8 Main and Interaction Effects for Year-2 Data

FACTOR	Root Dry Weight		Stalk Dry Weight		Head Dry Weight	
	EFFECT	% AVERAGE	EFFECT	% AVERAGE	EFFECT	% AVERAGE
Avg	68.1		100.7		62.9	
M1	22.3	32.8	9.1	9.0	6.7	10.7
M2	-8.5	-12.5	-5.4	-5.4	-1.0	-1.5
M3	-13.8	-20.3	-3.7	-3.7	-5.8	-9.2
N1	-1.2	-1.7	0.5	0.5	-1.1	-1.7
N2	-3.6	-5.3	3.4	3.4	5.6	8.9
N3	4.8	7.0	-3.8	-3.8	-4.6	-7.2
H1	7.6	11.2	17.2	17.1	4.8	7.6
H2	-7.6	-11.2	-17.2	-17.1	-4.8	-7.6
MH11					-3.6	-5.8
MH12					3.6	5.8
MH21					4.9	7.7
MH22					-4.9	-7.7
MH31					-1.2	-1.9
MH32					1.2	1.9
NH11					-1.6	-2.5
NH12					1.6	2.5
NH21					6.0	9.6
NH22					-6.0	-9.6
NH31					-4.5	-7.1
NH32					4.5	7.1
MNH111	10.6	15.6				
MNH112	-10.6	-15.6				
MNH121	3.6	5.2				
MNH122	-3.6	-5.2				
MNH131	-14.1	-20.7				
MNH132	14.1	20.7				
MNH211	-8.0	-11.8				
MNH212	8.0	11.8				
MNH221	-2.6	-3.9				
MNH222	2.6	3.9				
MNH231	10.6	15.6				
MNH232	-10.6	-15.6				
MNH311	-2.5	-3.7				
MNH312	2.5	3.7				
MNH321	-0.9	-1.3				
MNH322	0.9	1.3				
MNH331	3.3	4.9				
MNH332	-3.3	-4.9				

Table 5.8 shows the average weights and the details of the individual main effects and some interaction effects for year-2 data. Only MNH-interaction effects for root dry weight, and MH-interaction effects and NH-interaction effects for head dry weight are reproduced in the table. Other interaction effects are not reproduced since they are negligible based on the variation explained by them and their influence could be ignored.

Now let us check the root dry weight. The average weight is 68.1 grams. Having no M. phaseolina in the soil could increase the average dry weight by 22.3 grams (32.8%). Meanwhile, the existence of 10 cfu M. phaseolina per gram soil (M. phaseolina at level 2) would cause a decrease of 8.5 grams (12.5%) to the average dry weight and 100 cfu M. phaseolina per gram soil (M. phaseolina at level 3) would decrease the average weight by 13.8 grams (20.3%). The choice of hybrid could increase or decrease the average dry weight by 7.6 grams (11.2%). The important interaction for root dry weight is the MNH-interaction, while the two-variable interactions are not really significant. Among the MNH-interaction effects, particularly, the ones where M. phaseolina at level 1 or 2 and nematode at level 3 have a quite big impact on the average dry weight.

The average stalk dry weight is 100.7 grams which is more than twice the average stalk dry weight in year-1. It seems that the plants in year-2 experiment are a lot bigger than the ones in year-1. The choice of hybrid is very important to the average dry weight. Hybrid might affect the average dry weight by 17.2 grams (17.1%). As in year-1, having more M. phaseolina in the soil would decrease the stalk dry weight less, instead of more.

The average head dry weight is 62.9 grams which is also more than twice the average head dry weight in year-1. It confirms that the plants in this year are about twice the size of the ones in the previous year. The main and MH or NH-interaction effects are about in the same magnitude. Individually, they might affect the average dry weight by less than 10%. One interesting thing is having 1554 nematodes per 16 kg soil

(nematode at level 2) would increase, instead of decrease, the average dry weight by 8.9%, which is unlikely to be anticipated. Among the more important interaction effects between M. phaseolina or nematode and hybrid are when M. phaseolina is at level 1 or 2, and when nematode is at level 2 or 3.

5.4.5. Recapitulation

In analyzing a set of data, we start by computing the percentage of variation and then check the individual effects. The percentage of variation measures the variability that exists in the data that is caused by the deviation of data from the mean value. This percentage provides an easy way to gauge the significance of a factor (main or interaction effect). After knowing the significant factors as explained by their high percentage of variation, then we investigate the details of the individual effects of those factors.

The analysis was done in three stages. In the first stage, we investigate the replication factor. From this analysis, we have found out that the experiment in year-1 was conducted quite successfully while the experiment in year-2 was not. Some external factors other than the variables under study were affecting the outcome of the experiment in year-2.

In the next stage, we investigate the year factor. Here, we come to the conclusion that the year-1 data are very much different from the year-2 data especially for stalk and head dry weights. This leads us to perform separate analysis for each year data.

Finally, in the last stage, we analyze the data separately for each year. The analysis reveals a variety of detailed patterns. We start by looking at the percentage of variation for each factor followed by the details of individual effects. From year-1 data, the result seems to be more consistent. M. phaseolina is the most significant variable affecting the whole outcomes of the experiment. With one exception, the interaction effects seem to be trivial in year-1. From year-2 data, the dominant factors seem to be more arbitrary and more distributed among variables. M. phaseolina is the important variable for root

dry weight, while hybrid is the dominant variable for stalk and head dry weights in year-2. Other variables are also equally dominant for head dry weight in year-2.

Comparing the analysis for both years, one could not infer and say with enough confidence what variables are really important in reducing sorghum's production since the analysis from the two years have given different results. The conclusion for each year is valid for that year. But when considering them together, a clear conclusion would be difficult to obtain. As was evident from the average weights differences between year-1 and year-2 data and the pattern of data in table 5.2 and table 5.3, it seems there are some fundamental differences between these two data that make it impossible for a single final conclusion to be drawn.

5.5. Concluding Remarks

The style of this analysis was certainly a practical one and the bulk of the findings were presented in a purely descriptive manner. Numerical summaries were used to record and portray particular results. Three stages of analysis were presented to extract as much information as possible because the weight of accumulated evidence is essential for making a convincing case.

One cannot deny that the course of a data analysis will reflect the background and biases of the analysts themselves. It is no accident that *k*-systems theory played a major role in this work. Our involvement with this arose from our recognition of its potential for this study. The use of *k*-systems theory reflects our own upbringing. If the analysis using this theory is done with proper care, then the most essential aspects of what the data have to tell us ought to emerge in a clear and forceful way.

There are several approaches that can be used for data analysis. One of the most common approach is the analysis of variance. With this method, we need to assume a model for the underlying data and then check for the validity of the model using such techniques as plotting residuals and identifying outliers. Often it is tempting to apply

this method to not-quite-appropriate situations. In a sense, the analysis of variance model is forced to fit the data. Questions arise not only on how the model relates to the underlying data but also on the relevance of the assumptions which accompany the model. If these assumptions are grossly violated, the procedures used to draw inferences about the model may be invalid.

K-systems theory overcomes this problem. It works directly with whatever information is available in the data and does not assume any model for the data. Thus, no model validity checking or assumption verification is needed. In addition, the results obtained are valid and true for the given data.

Chapter 6

CONCLUSIONS

6.1. Summary

The focus of this dissertation is data analysis and reconstructability analysis (RA). As mentioned in the first chapter, this research might be the first one which concentrates on developing a new data analysis method using reconstructability theory. The approach starts with the concept of variable interaction in RA. We attempt to use the RA definition of interaction to improve the quality of main and interaction effects in data analysis.

Data analysis methodologies have been studied and improved for many years. Now, it is one of the most important field in statistics and has found broad applications in many disciplines. There are several approaches that can be used for data analysis. From all the approaches, one common characteristic is that they all share the idea of a model for the underlying data. Model validity checking is then performed to ensure the validity of the analysis. In this context, we propose to use RA methodology, particularly k-systems theory, as a new approach to compute variable effects in data analysis. RA works directly with whatever information is available in the data, and most importantly, it does not assume any model for the data. Thus, no model validity checking is needed. Besides, the results obtained are valid and true for the given data.

Although we use a lot of experimental design (measured) data to illustrate the concept in this dissertation, the concept also applies to observational (non-experimental design) data. In other words, the concept applies to any data analysis problems, including experimental design or observational data.

We have presented some specific issues in this research. First, we introduced the concept of variable interaction in classical statistical context and RA methodology in

chapter 3. An illustrative example was provided, and the results from these two approaches were then compared and analyzed. Next, in chapter 4, we developed a general algorithm for data analysis using reconstructability theory. We also supplied a hypothetical example of fractional factorial design to illustrate the approach and to describe the confounding problem and how we overcame this problem using the new approach. In chapter 5, we implemented the technique to a set of real experimental design data in a form of a case study. The data set are collected from a design of experiments at LSU Plant Pathology Department.

6.2. Principle Contributions of The Dissertation

In this section, we summarize the principle contributions of the dissertation as follows:

- A study of comparison between classical statistical interaction and k-systems interaction is presented.
- The concept of k-systems main effect is developed and its detail is discussed.
- An algorithm to compute true main and interaction effects using k-systems theory is developed.
- One of the advantages of our method in overcoming confounding problem which constitutes a major problem in fractional factorial experiments is described quite elaborately.
- Finally, a case study implementing the new method is given. The study provides an in-depth analysis of a set of real experimental design data and shows how data analysis would be done using this method.

6.3. Future Research

Based on our research, there are some interesting questions or problems still remain to be investigated.

- We realize that the method is still yet to be fully developed even though it has shown its potential for solving statistical problems. Thus, while the method has been successfully applied to a variety of data analysis problems, there is still a great deal to be learned about its properties.
- A general sequential algorithm for computing true variable effects using k-systems theory has been discussed. Detailed investigations could lead to an improvement in the implementation part, and also parallelism of the algorithm is a possibility.
- This method could potentially be used in pattern recognition problems, especially in the reconstructions of images from a set of data.

REFERENCES

Ashby, W. R. 1964. Constraint analysis of many-dimensional relations. *General Systems Yearbook* 9:99-105.

Bose, R. C. and K. Kishen. 1940. On the problem of confounding in general symmetrical factorial design. *Sankhya* 5:21-36.

Box, G. E. P., W. G. Hunter, and J. S. Hunter. 1978. *Statistics for Experimenters.* New York: John Wiley & Sons.

Cavallo, R. E. and G. J. Klir. 1979. Reconstructability analysis of multi-dimensional relations: a theoretical basis for computer-aided determination of acceptable systems models. *International Journal of General Systems* 5:143-171.

Cavallo, R. E. and G. J. Klir. 1981a. Reconstructability analysis: overview and bibliography. *International Journal of General Systems* 7:1-6.

Cavallo, R. E. and G. J. Klir. 1981b. Reconstructability analysis: evaluation of reconstruction hypotheses. *International Journal of General Systems* 7:7-32.

Cavallo, R. E. and G. J. Klir. 1982a. Decision making in reconstructability analysis. *International Journal of General Systems* 8:243-255.

Cavallo, R. E. and G. J. Klir. 1982b. Reconstruction of possibilistic behavior systems. *Fuzzy Sets and Systems* 8:175-197.

Cox, D. R. 1966. *Planning of Experiments.* New York: John Wiley & Sons.

Dempster, A. P. 1983. Purposes and limitations of data analysis. In *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and C. F. Wu, 117-133. New York: Academic Press.

Dubois, D. and H. Prade. 1980. *Fuzzy Sets and Systems.* New York: Academic Press.

Edwards, D. 1990. Hierarchical interaction models. *Journal of the Royal Statistical Society.* Series B 52:3-20.

Fisher, R. A. 1926. The arrangement of field experiments. *Journal of Ministry of Agriculture* 33:503-513.

Freeman, D. H. Jr. 1987. *Applied Categorical Data Analysis.* New York: Marcel Dekker.

Gouw, D. and B. Jones. In press1. The interaction concept of K-systems theory. *International Journal of General Systems.*

Gouw, D. and B. Jones. In press2. Using k-systems theory to compute true main and interaction effects in a design of experiments. *Cybernetics and Systems: An International Journal.*

Gouw, D., I. Wenefrida, and B. Jones. Submitted. A case study in experimental design data analysis using k-systems theory. *Cybernetics and Systems: An International Journal.*

Hicks, C. R. 1982. *Fundamental Concepts in the Design of Experiments.* New York: Holt, Rinehart and Winston.

Jaynes, E. T. 1957. Information theory and statistical mechanics. *Physical Reviews* 106:620-630, 108:171-190.

Jaynes, E. T. 1979. Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, eds. R. Levine and M. Tribus, 15-117. Cambridge, MA: MIT Press.

Johnson, D. E. and F. A. Graybill. 1972. An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association* 67:862-868.

Jones, B. 1982. Determination of reconstruction families. *International Journal of General Systems* 8:225-228.

Jones, B. 1985a. Determination of unbiased reconstructions. *International Journal of General Systems* 10:169-176.

Jones, B. 1985b. A greedy algorithm for a generalization of the reconstruction problem. *International Journal of General Systems* 11:63-68.

Jones, B. 1985c. Reconstructability analysis for general functions. *International Journal of General Systems* 11:133-142.

Jones, B. 1985d. Reconstructability considerations with arbitrary data. *International Journal of General Systems* 11:143-151.

Jones, B. 1985e. The cognitive content of system substates. *IEEE Workshop on Languages for Automation*, pp. 11-13. Palma de Mallorca, Spain, June 28-29.

Jones, B. 1986. K-systems versus classical multivariate systems. *International Journal of General Systems* 12:1-6.

Jones, B. 1989. A program for reconstructability analysis. *International Journal of General Systems* 15:199-205.

Kettenring, J. R. 1983a. Components of interactions in analysis of variance models with no replications. In *Contributions to Statistics: Essays in Honor of Norman L. Johnson*, ed. P. K. Sen, 282-297. Amsterdam:North-Holland.

Kettenring, J. R. 1983b. A case study in data analysis. In *Statistical Data Analysis*, ed. R. Gnanadesikan, 105-139. Proceedings of Symposia in Applied Mathematics vol. 28. Rhode Island: American Mathematical Society.

Klir, G. J. 1976. Identification of generative structure in empirical data. *International Journal of General Systems* 3:89-104.

Klir, G. J. 1985a. Cognitive aspects of reconstructability analysis. *IEEE Workshop on Languages for Automation*, pp. 6-10. Palma de Mallorca, Spain, June 28-29.

Klir, G. J. 1985b. *Architecture of Systems Problem Solving*. New York: Plenum Press.

Klir, G. J. and E. C. Way. 1985. Reconstructability analysis: aims, result and open problems. *Systems Research* 2:141-163.

Kullback, S. 1961. *Information Theory and Statistics*. New York: Wiley.

Lorenzen, T. J. and V. L. Anderson. 1993. *Design of Experiments: A No-Name Approach*. New York: Marcel Dekker.

Mallows, C. L. 1983. Data description. In *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and C. F. Wu, 135-151. New York: Academic Press.

Mason, R. L., R. F. Gunst, and J. L. Hess. 1989. *Statistical Design and Analysis of Experiments*. New York: John Wiley & Sons.

McLean, R. A. and V. L. Anderson. 1984. *Applied Factorial and Fractional Designs*. New York: Marcel Dekker.

Montgomery, D. C. 1991. *Design and Analysis of Experiments*. New York: John Wiley & Sons.

Mughogho, L. K. and S. Pande. 1984. Charcoal Rot of Sorghum. In *Sorghum Root and Stalk Rot: A Critical Review*, 11-24. Proceedings of the Consultative Group Discussion on Research Needs and Strategies for Control of Sorghum Root and Stalk Rot Diseases, Bellagio, Italy, 27 Nov - 2 Dec 1983. Patancheru, India: ICRISAT.

Petersen, R. G. 1985. *Design and Analysis of Experiments*. New York: Marcel Dekker.

Raktoe, B. L., A. Hedayat, and W. T. Federer. 1981. *Factorial Designs*. New York: John Wiley & Sons.

Scheffe', H. 1959. *The Analysis of Variance*. New York: John Wiley & Sons.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27:379-423,623-656.

Snedecor, G. W. and W. G. Cochran. 1967. *Statistical Methods*. 6th edition. Ames, Iowa: The Iowa State University Press.

Trivedi, S. K. 1993. *Reconstructability Theory for General Systems and Its Application to Automated Rule Learning*. Louisiana State University: Ph.D. Dissertation.

Yates, F. 1935. Complex experiments. *Journal of the Royal Statistical Society*. Supplement 2:181-247.

Yates, F. 1937. The design and analysis of factorial experiments. *Imperial Bureau of Soil Science*. Technical Communication 35.

Appendix A

RECONSTRUCTABILITY EFFECTS VS STATISTICAL EFFECTS

FACTOR	EFFECT (Reconstructability Analysis)	EFFECT (Statistical Analysis)
A=7.5	60.481	60.481
A=12.5	7.648	7.648
A=17.5	-68.130	-68.130
B=5.0	-28.185	-28.185
B=12.5	1.204	1.204
B=20.0	26.982	26.982
C=1900	-30.963	-30.963
C=2000	44.537	44.537
C=2300	-13.574	-13.574
D=1	-5.463	-5.463
D=2	5.463	5.463
A=7.5 B=5.0	34.62	28.85
A=7.5 B=12.5	-22.62	-22.37
A=7.5 B=20.0	-12.00	-6.48
A=12.5 B=5.0	8.25	7.52
A=12.5 B=12.5	8.10	8.13
A=12.5 B=20.0	-16.35	-15.65
A=17.5 B=5.0	-42.86	-36.37
A=17.5 B=12.5	14.52	14.24
A=17.5 B=20.0	28.35	22.13
A=7.5 C=1900	52.63	46.30
A=7.5 C=2000	-79.81	-70.70
A=7.5 C=2300	27.18	24.41
A=12.5 C=1900	27.76	26.96
A=12.5 C=2000	-9.36	-8.20
A=12.5 C=2300	-18.41	-18.76
A=17.5 C=1900	-80.39	-73.26
A=17.5 C=2000	89.17	78.91
A=17.5 C=2300	-8.78	-5.65
A=7.5 D=1	5.53	4.41
A=7.5 D=2	-5.53	-4.41
A=12.5 D=1	3.38	3.24
A=12.5 D=2	-3.38	-3.24
A=17.5 D=1	-8.91	-7.65
A=17.5 D=2	8.91	7.65
B=5.0 C=1900	9.34	12.30
B=5.0 C=2000	-18.79	-23.04

(table con'd.)

FACTOR	EFFECT (Reconstructability Analysis)	EFFECT (Statistical Analysis)
B=5.0 C=2300	9.45	10.74
B=12.5 C=1900	7.20	7.07
B=12.5 C=2000	10.39	10.57
B=12.5 C=2300	-17.59	-17.65
B=20.0 C=1900	-16.55	-19.37
B=20.0 C=2000	8.40	12.46
B=20.0 C=2300	8.15	6.91
B=5.0 D=1	1.44	1.96
B=5.0 D=2	-1.44	-1.96
B=12.5 D=1	9.49	9.46
B=12.5 D=2	-9.49	-9.46
B=20.0 D=1	-10.93	-11.43
B=20.0 D=2	10.93	11.43
C=1900 D=1	4.39	4.96
C=1900 D=2	-4.39	-4.96
C=2000 D=1	-6.83	-7.65
C=2000 D=2	6.83	7.65
C=2300 D=1	2.43	2.68
C=2300 D=2	-2.43	-2.68
A=7.5 B=5.0 C=1900	-23.70	-26.96
A=7.5 B=5.0 C=2000	37.34	43.37
A=7.5 B=5.0 C=2300	-16.26	-16.41
A=7.5 B=12.5 C=1900	18.03	21.59
A=7.5 B=12.5 C=2000	-42.48	-43.41
A=7.5 B=12.5 C=2300	25.03	21.81
A=7.5 B=20.0 C=1900	5.34	5.37
A=7.5 B=20.0 C=2000	6.48	0.04
A=7.5 B=20.0 C=2300	-9.84	-5.41
A=12.5 B=5.0 C=1900	-36.80	-38.96
A=12.5 B=5.0 C=2000	48.74	50.04
A=12.5 B=5.0 C=2300	-12.31	-11.07
A=12.5 B=12.5 C=1900	15.97	16.26
A=12.5 B=12.5 C=2000	-0.94	0.43
A=12.5 B=12.5 C=2300	-15.95	-16.68
A=12.5 B=20.0 C=1900	20.24	22.70
A=12.5 B=20.0 C=2000	-47.45	-50.46
A=12.5 B=20.0 C=2300	28.47	27.76
A=17.5 B=5.0 C=1900	47.43	65.93
A=17.5 B=5.0 C=2000	-73.20	-93.41
A=17.5 B=5.0 C=2300	27.07	27.48
A=17.5 B=12.5 C=1900	-29.18	-37.85
A=17.5 B=12.5 C=2000	38.46	42.98
A=17.5 B=12.5 C=2300	-8.02	-5.13
A=17.5 B=20.0 C=1900	-17.92	-28.07
A=17.5 B=20.0 C=2000	33.28	50.43
A=17.5 B=20.0 C=2300	-18.08	-22.35

(table con'd.)

FACTOR	EFFECT (Reconstructability Analysis)	EFFECT (Statistical Analysis)
A=7.5 B=5.0 D=1	-13.90	-14.41
A=7.5 B=5.0 D=2	13.90	14.41
A=7.5 B=12.5 D=1	8.04	9.93
A=7.5 B=12.5 D=2	-8.04	-9.93
A=7.5 B=20.0 D=1	5.99	4.48
A=7.5 B=20.0 D=2	-5.99	-4.48
A=12.5 B=5.0 D=1	22.61	22.26
A=12.5 B=5.0 D=2	-22.61	-22.26
A=12.5 B=12.5 D=1	-34.85	-34.57
A=12.5 B=12.5 D=2	34.85	34.57
A=12.5 B=20.0 D=1	11.85	12.31
A=12.5 B=20.0 D=2	-11.85	-12.31
A=17.5 B=5.0 D=1	-10.07	-7.85
A=17.5 B=5.0 D=2	10.07	7.85
A=17.5 B=12.5 D=1	27.29	24.65
A=17.5 B=12.5 D=2	-27.29	-24.65
A=17.5 B=20.0 D=1	-17.03	-16.80
A=17.5 B=20.0 D=2	17.03	16.80
A=7.5 C=1900 D=1	-6.91	-6.74
A=7.5 C=1900 D=2	6.91	6.74
A=7.5 C=2000 D=1	11.76	12.70
A=7.5 C=2000 D=2	-11.76	-12.70
A=7.5 C=2300 D=1	-6.01	-5.96
A=7.5 C=2300 D=2	6.01	5.96
A=12.5 C=1900 D=1	2.27	1.93
A=12.5 C=1900 D=2	-2.27	-1.93
A=12.5 C=2000 D=1	-14.30	-13.80
A=12.5 C=2000 D=2	14.30	13.80
A=12.5 C=2300 D=1	11.80	11.87
A=12.5 C=2300 D=2	-11.80	-11.87
A=17.5 C=1900 D=1	2.12	4.81
A=17.5 C=1900 D=2	-2.12	-4.81
A=17.5 C=2000 D=1	4.93	1.09
A=17.5 C=2000 D=2	-4.93	-1.09
A=17.5 C=2300 D=1	-6.04	-5.91
A=17.5 C=2300 D=2	6.04	5.91
B=5.0 C=1900 D=1	-4.95	-5.63
B=5.0 C=1900 D=2	4.95	5.63
B=5.0 C=2000 D=1	4.10	5.48
B=5.0 C=2000 D=2	-4.10	-5.48
B=5.0 C=2300 D=1	0.57	0.15
B=5.0 C=2300 D=2	-0.57	-0.15
B=12.5 C=1900 D=1	17.63	16.70
B=12.5 C=1900 D=2	-17.63	-16.70
B=12.5 C=2000 D=1	0.36	1.65
B=12.5 C=2000 D=2	-0.36	-1.65

(table con'd.)

FACTOR	EFFECT (Reconstructability Analysis)	EFFECT (Statistical Analysis)
B=12.5 C=2300 D=1	-17.88	-18.35
B=12.5 C=2300 D=2	17.88	18.35
B=20.0 C=1900 D=1	-13.06	-11.07
B=20.0 C=1900 D=2	13.06	11.07
B=20.0 C=2000 D=1	-4.46	-7.13
B=20.0 C=2000 D=2	4.46	7.13
B=20.0 C=2300 D=1	17.69	18.20
B=20.0 C=2300 D=2	-17.69	-18.20
A=7.5 B=5.0 C=1900 D=1	3.04	3.41
A=7.5 B=5.0 C=1900 D=2	-3.04	-3.41
A=7.5 B=5.0 C=2000 D=1	-29.43	-32.04
A=7.5 B=5.0 C=2000 D=2	29.43	32.04
A=7.5 B=5.0 C=2300 D=1	27.31	28.63
A=7.5 B=5.0 C=2300 D=2	-27.31	-28.63
A=7.5 B=12.5 C=1900 D=1	-30.55	-24.26
A=7.5 B=12.5 C=1900 D=2	30.55	24.26
A=7.5 B=12.5 C=2000 D=1	38.91	36.96
A=7.5 B=12.5 C=2000 D=2	-38.91	-36.96
A=7.5 B=12.5 C=2300 D=1	-8.75	-12.70
A=7.5 B=12.5 C=2300 D=2	8.75	12.70
A=7.5 B=20.0 C=1900 D=1	27.84	20.85
A=7.5 B=20.0 C=1900 D=2	-27.84	-20.85
A=7.5 B=20.0 C=2000 D=1	-8.79	-4.93
A=7.5 B=20.0 C=2000 D=2	8.79	4.93
A=7.5 B=20.0 C=2300 D=1	-18.74	-15.93
A=7.5 B=20.0 C=2300 D=2	18.74	15.93
A=12.5 B=5.0 C=1900 D=1	-13.47	-15.26
A=12.5 B=5.0 C=1900 D=2	13.47	15.26
A=12.5 B=5.0 C=2000 D=1	32.80	35.96
A=12.5 B=5.0 C=2000 D=2	-32.80	-35.96
A=12.5 B=5.0 C=2300 D=1	-18.35	-20.70
A=12.5 B=5.0 C=2300 D=2	18.35	20.70
A=12.5 B=12.5 C=1900 D=1	-11.07	-6.26
A=12.5 B=12.5 C=1900 D=2	11.07	6.26
A=12.5 B=12.5 C=2000 D=1	-6.93	-12.87
A=12.5 B=12.5 C=2000 D=2	6.93	12.87
A=12.5 B=12.5 C=2300 D=1	15.97	19.13
A=12.5 B=12.5 C=2300 D=2	-15.97	-19.13
A=12.5 B=20.0 C=1900 D=1	25.20	21.52
A=12.5 B=20.0 C=1900 D=2	-25.20	-21.52
A=12.5 B=20.0 C=2000 D=1	-26.50	-23.09
A=12.5 B=20.0 C=2000 D=2	26.50	23.09
A=12.5 B=20.0 C=2300 D=1	1.06	1.57
A=12.5 B=20.0 C=2300 D=2	-1.06	-1.57
A=17.5 B=5.0 C=1900 D=1	10.75	11.85
A=17.5 B=5.0 C=1900 D=2	-10.75	-11.85

(table con'd.)

FACTOR	EFFECT (Reconstructability Analysis)	EFFECT (Statistical Analysis)
A=17.5 B=5.0 C=2000 D=1	-5.70	-3.93
A=17.5 B=5.0 C=2000 D=2	5.70	3.93
A=17.5 B=5.0 C=2300 D=1	-6.72	-7.93
A=17.5 B=5.0 C=2300 D=2	6.72	7.93
A=17.5 B=12.5 C=1900 D=1	52.00	30.52
A=17.5 B=12.5 C=1900 D=2	-52.00	-30.52
A=17.5 B=12.5 C=2000 D=1	-37.19	-24.09
A=17.5 B=12.5 C=2000 D=2	37.19	24.09
A=17.5 B=12.5 C=2300 D=1	-8.42	-6.43
A=17.5 B=12.5 C=2300 D=2	8.42	6.43
A=17.5 B=20.0 C=1900 D=1	-59.95	-42.37
A=17.5 B=20.0 C=1900 D=2	59.95	42.37
A=17.5 B=20.0 C=2000 D=1	40.05	28.02
A=17.5 B=20.0 C=2000 D=2	-40.05	-28.02
A=17.5 B=20.0 C=2300 D=1	16.81	14.35
A=17.5 B=20.0 C=2300 D=2	-16.81	-14.35

Appendix B

MAIN AND INTERACTION EFFECTS IN SCHEDULER DESIGN STUDY

FACTOR	EFFECT
A−	4.25
A+	-4.25
B−	2.00
B+	-2.00
C−	-1.13
C+	1.13
D−	-4.63
D+	4.63
AB−−	-0.14
AB−+	0.14
AB+−	0.14
AB++	-0.14
AC−−	-3.60
AC−+	3.60
AC+−	3.60
AC++	-3.60
AD−−	0.37
AD−+	-0.37
AD+−	-0.37
AD++	0.37
BC−−	-0.18
BC−+	0.18
BC+−	0.18
BC++	-0.18
BD−−	-3.46
BD−+	3.46
BD+−	3.46
BD++	-3.46
CD−−	-0.04
CD−+	0.04
CD+−	0.04
CD++	-0.04
ABC−−−	-4.37
ABC−−+	4.37
ABC−+−	4.37
ABC−++	-4.37
ABC+−−	4.37
ABC+−+	-4.37
ABC++−	-4.37
ABC+++	4.37

(table con'd.)

FACTOR	EFFECT
ABD−−−	-0.65
ABD−−+	0.65
ABD−+−	0.65
ABD−++	-0.65
ABD+−−	0.65
ABD+−+	-0.65
ABD++−	-0.65
ABD+++	0.65
ACD−−−	1.46
ACD−−+	-1.46
ACD−+−	-1.46
ACD−++	1.46
ACD+−−	-1.46
ACD+−+	1.46
ACD++−	1.46
ACD+++	-1.46
BCD−−−	4.09
BCD−−+	-4.09
BCD−+−	-4.09
BCD−++	4.09
BCD+−−	-4.09
BCD+−+	4.09
BCD++−	4.09
BCD+++	-4.09
ABCD−−−−	-1.65
ABCD−−−+	1.65
ABCD−−+−	1.65
ABCD−−++	-1.65
ABCD−+−−	1.65
ABCD−+−+	-1.65
ABCD−++−	-1.65
ABCD−+++	1.65
ABCD+−−−	1.65
ABCD+−−+	-1.65
ABCD+−+−	-1.65
ABCD+−++	1.65
ABCD++−−	-1.65
ABCD++−+	1.65
ABCD+++−	1.65
ABCD++++	-1.65

VITA

Deky D. Gouw was born in Jakarta, Indonesia. He received his Ir. (Engineer, equivalent to B.Eng.) degree in Civil Engineering from Parahyangan Catholic University, Indonesia in 1987. He also obtained his M.S. degree in Systems Science from Louisiana State University, Baton Rouge, U.S.A. in 1990. He joined the Ph.D. program in Computer Science at Louisiana State University, Baton Rouge in the Spring of 1991, where he worked on problems of data analysis and system reconstruction. Deky Gouw has been an assistant to systems administrator maintaining the computer systems of LSU Computer Science Department and a programming class instructor teaching two 3-hour classes/semester for four semesters.

In addition to reconstructability analysis and data analysis, his research interests include numerical analysis methods. Deky Gouw has publications which are to appear in *International Journal of General Systems* and *Cybernetics and Systems: An International Journal*. He is a member of the Association for Computing Machinery, and the Institute for Operations Research and the Management Sciences (formerly Operations Research Society of America).

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Deky Dicky Joe Gouw

Major Field: Computer Science

Title of Dissertation: Data Analysis Using Reconstructability Theory

Approved:

Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

Date of Examination:

March 9, 1995