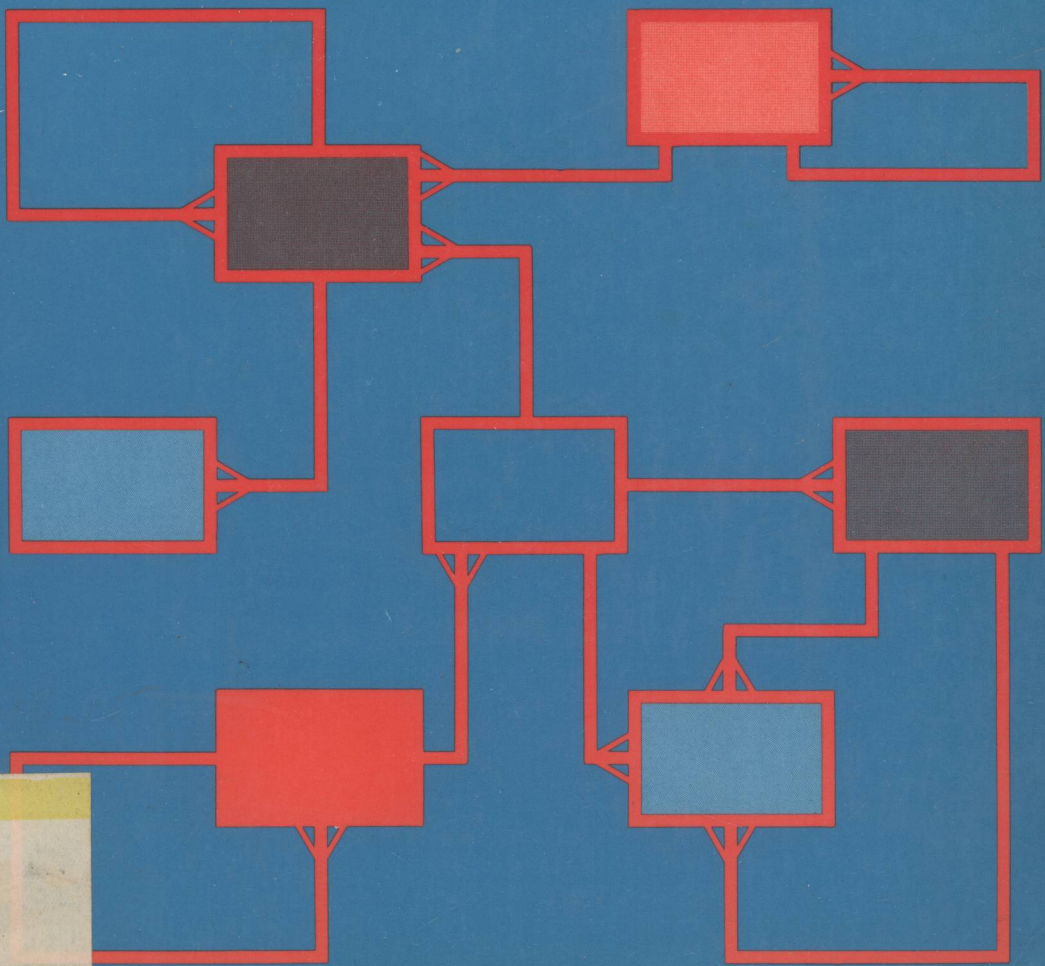
R. VERYARD

PRAGMATIC DATA ANALYSIS

Blackwell Scientific Publications



Pragmatic Data Analysis

R. VERYARD
MA, MSc
Data Logic Ltd,
London

BLACKWELL SCIENTIFIC PUBLICATIONS
OXFORD LONDON EDINBURGH
BOSTON PALO ALTO MELBOURNE

Blackwell Scientific Publications
Editorial offices:
Osney Mead, Oxford, OX2 0EL
8 John Street, London, WC1N 2ES
9 Forrest Road, Edinburgh, EH1 2QH
52 Beacon Street, Boston
Massachusetts 02108, USA
706 Cowper Street, Palo Alto
California 94301, USA
99 Barry Street, Carlton
Victoria 3053, Australia

First published 1984

Phototypesetting by
Parkway Illustrated Press, Abingdon.
Printed and bound in Great Britain

Distributed in North America by
Computer Science Press Inc.,
11 Taff Court,
Box 6030, Rockville,
Maryland 20850, USA

British Library
Cataloguing in Publication Data

Veryard, R.
Pragmatic data analysis
1. Mathematic statistics—Data processing
I. Title
519.4 QA276.4

ISBN 0-632-01311-7

PRAGMATIC DATA ANALYSIS

Preface

It is customary etiquette at this point for the author to modestly acknowledge the imperfections of the book (without being honest enough to admit what they are) and to express gratitude to a selection of friends and acquaintances who helped create the book. It is usually felt necessary to stress — although it should go without saying — that there is no connection between the faults and the friends.

As I see it, the main fault in the book is the inadequate division between description and prescription. I have neither given an objective and uninterpreted account of the methods employed in a typical DP department, nor have I started from first principles and the latest research in order to develop a brand new formulation of my own, which would be wholly logically sound and sophisticated. I have presented a mixture of prescription and description. It follows that the methods outlined in this book cannot be adopted uncritically, but the reader is obliged to think how best to adapt them to his own needs. He must work out himself how to bridge some of the gaps. I had not realized when I started writing just how many gaps there would have to be, which I could only fill by embarking on a completely different course than I had originally planned. It is probably impossible in this subject to be both comprehensible and comprehensive.

My gratitude goes to the following:

to anyone who may have had these ideas before me. I am not always able to remember where I read or heard these ideas, but I am certainly not clever enough to think it all up by myself — my apologies for any unwitting plagiarism;

to Ian Palmer, for whom I once worked, whose formation was a stimulus and starting point for my own thinking;

to those who have worked with me on data analysis exercises, especially Raimo Rikkilae, Francis Murphy, Angela Simonsson and Kevin Swindells — the best way of learning is to discuss with a colleague;

to those whom I have attempted to teach, whose incomprehension and questions have forced me to find new expressions and problems;

to my secretary, Sarah Starr, for managing the text;

to my employer, Data Logic, for encouraging and assisting me to go into print;

to the estate of the late Sonia Brownell Orwell and Martin Secker & Warburg Ltd for permission to quote from George Orwell's *Collected Essays, Journalism and Letters* (1931).

Contents

Chapter 1

Introduction

Introduction

This chapter introduces the notion of data analysis and explains its importance. The book as a whole rests on three premises, which are outlined below.

First, that information is an important asset or resource. For any information system, whether or not computerized, and particularly where data are shared between subsystems (i.e. almost always), a clear and accurate knowledge of the data structure is needed.

Second, that data analysis is a branch of systems analysis and therefore shares its principles. Of particular relevance are the separation of analysis from design, the clear statement of objectives, assumptions and priorities, the systematic top–down and iterative approaches to analysis, and the unambiguous documentation of results.

Third, that data analysis is not (at least not yet) a fully mechanical activity. The data analyst often has to make arbitrary decisions on a criterion of elegance, or according to what the users are likely to understand and agree with. This book gives guidelines to common sense; its procedures are no substitute for intelligence.

Data analysis

What is data analysis? As its name implies, it is a branch of systems analysis concerned with the structure of data within a system. Whereas the techniques of systems analysis were first worked out by military advisors for the design of weapons systems, data analysis is special to the data processing (DP) industry. Since the structure and flow of data is of central importance in any DP system, data analysis is central to any DP systems analysis. In the past, data analysis was often carried out implicitly, unrecognized. Ad hoc techniques, based on common sense and experience, were used to translate the users' data requirements into file types and record types. Nowadays, as DP systems grow more complex, and with the increasing use of large databases, these ad hoc methods are being replaced by more formal methodologies. The results

of the analysis are documented in a standard format, often using an automated data dictionary. Checks for logical consistency between subsystems can be carried out easily, and any file or database design proceeds smoothly, because the data requirements are unambiguously and clearly defined.

In particular, the following advantages of data analysis are generally recognized within commercial data processing departments.

1 The users' requirements of data structure tend to be much more stable than their functional requirements. This means that a computer systems design based on a proper analysis of the data structure is less likely to need extensive modification whenever the business needs change, than one based on current operations and functions. Maintenance of badly designed and inflexible computer systems is a serious problem in most large computer installations.

2 The results of the data analysis are in a form that can easily be understood by the users themselves. Discussion of their requirements, including negotiation between conflicting user departments, can be based on a clear statement of the data structure. This is obviously better than producing a computerese system specification, in which the user cannot hope to find the flaws. Systems developed from such documents rarely provide user satisfaction.

Data modelling

A data model of an organization is a systematic representation of the data requirements of the organization. Such a model is independent of any computerized or manually operated system; however, any system design should be based on such a model. The model documents the structure of and the interrelationships between the data. The differing requirements of the different departments and of the different functional areas of the organization are contained and reconciled in the model. The model is presented as a combination of simple diagrams and written definitions.

Data and information

It will come as no surprise to the reader of a data processing textbook to be told of the importance of data. In order to demonstrate this importance, let us compare a commercial organization with the human body.

Human anatomy includes a system to circulate oxygen and nutrients to all parts of the body, carried by the bloodstream and pumped by the heart. This corresponds to the circulation of goods, services and cash in a commercial organization. There is also a system to monitor the operations of the body, to issue instructions to the limbs and vital organs, and to gather information from the outside world via the sense organs. Messages are passed along special channels, i.e. nerves. This corresponds to the flow of data in a commercial organization.

Decisions are made in the brain on the basis of available information. This information is provided by the nervous system and either used immediately or stored for future use. This corresponds to the supply of information to the managers of a commercial organization for the purposes of decision-making.

Finally, let us consider what happens to a part of the human body when the nerve connections are broken. This usually results in paralysis, and if the paralysed part is a vital organ then the body will soon die. This corresponds to a major loss of data within an organization, such as might be caused by the computer catching fire. It may surprise the reader to learn that a serious fire in the computer room is more likely to cause bankruptcy than a serious fire in the warehouse. In other words, the loss of data may be more disastrous than the loss of goods. Therefore data may be more valuable than goods.

The value of data depends on six aspects.

1 *Accurate.* The data must enable correct decisions to be taken. However, unnecessary precision should be avoided. Detailed accounts may be accurate to the nearest penny, but the national budget may be accurate only to the nearest million pounds; further accuracy would be pointless.

2 *Prompt.* The data must be sufficiently up-to-date to allow prompt action to be taken. Often it is the first person to obtain information who gains the commercial benefit. The wealth of the Rothschild family is at least partly due to the intelligence of Nathan Rothschild, who was the first person in London to learn the result of the battle of Waterloo. He had set up a network of spies to bring him the news, and he had time to make many successful speculative deals before the news became public.

3 *Well directed.* The data must be available to the appropriate person within the organization. This has a positive aspect (the right person gets the information he needs) and a negative aspect (the wrong person does not have access to confidential or private data).

4 *Brief.* The important data should not be submerged in a mass of non-essential details, which may lead vital items to be overlooked.

5 *Rare.* The value of any piece of information depends on its unlikeliness or unexpectedness.

6 *Complete.* A list of items will usually be assumed complete, unless the contrary is made clear. An item being incorrectly omitted may be as dangerous as an item being incorrectly included. In some circumstances, however, selective lists are acceptable.

Finally in this section we should state the relationship between data and information. There is no formal distinction between the two; for the purposes of this book, *information* is that which is used for decision-making, *data* are the messages that are passed and stored within an organization. In other words, data are transformed into information by being interpreted and used by a decision-maker. Data plus interpretation equals information.

Shared data

The importance of data stems from their use in communication. Data are communicated in two dimensions: in time, i.e. from the past to the future, and in space, i.e. from one part of the organization to another. Both of these aspects of data must be understood by the analyst. Communication in time requires the concept of data storage; communication in space requires the concept of data sharing.

Why should data be shared at all? If two areas require the same or similar information, the alternative to sharing is copying; but if several copies of the information are maintained independently, this leads to inefficiency and inconsistency. Sharing data means that all user areas are provided with the same, equally up-to-date information.

The data analyst is therefore obliged to examine each item of data that is common to several user areas and ensure that the assumptions made about these data in these different areas do not conflict.

Systems and system objectives

Systems are not restricted to computer software, and the analysis of a clerical system with the aim of computerizing some of it should not limit itself to analysing the parts that are destined for automation. The scope of an analysis should always be broader than that of any subsequent design. A system exists to carry out some *function* of the organization in which it resides; the carrying-out of the function does not end when the computer displays or prints the result of some calculation, it does not end until some human being has taken the result and used it to some

purpose. A good system involves the co-operation of man and computer; a good analyst must understand the human side of the system as well as the machine side.

A decision may be made by a human expert on the basis of intuitive judgement and 'feel'. It might be impossible or prohibitively expensive (or politically inconvenient or morally unacceptable) to program a computer to make that decision. But the system as a whole needs that decision made in order to carry out the function for which it is designed. A description of the system would be incomplete without mentioning the need for human intervention.

The objectives of a system should be expressed in terms of the function of the system for the organization and not in terms of the role of the computer specificially. Words like 'facilitate' usually imply the latter.

The objective of a stock control system is 'to avoid excess stock building up, without running out of anything'. It is *not* 'to provide regular management reports facilitating prompt decisions on stock levels'. The computer may do the latter, the system does the former. In evaluating the system, the contribution of any computer should be considered, but always in terms of the objectives of the whole system.

No methodology solves all problems

A few warning words about methodologies in general, and data analysis in particular. There is an ancient Chinese proverb: 'A legless man cannot walk on stilts.' No methodology can be a substitute for common-sense and experience, it can only supplement these. If a methodology gives you checklists or questionnaires, use them — but use your brain too! Data analysts are often given forms to fill in, to aid the documentation of each entity, relationship and attribute. It is a mistake to suppose that filling in such a form always exhausts the information that should be recorded. No standard form, no methodology can be that flexible.

This book is compatible with most commonly used methodologies, in that they all call for some kind of data analysis, and for the production of a data model. The format of the data model itself may vary, but the concepts are usually the same. When working within a particular methodology, there may be additional procedures to follow, documentation to be produced in a different way, or even a different notation to learn. However, if the reader has grasped the essential concepts introduced in this book, their adaption to new problem areas,

within the guidelines of a particular methodology or a set of formal standards, should not cause any difficulties.

Furthermore, a good analyst does not restrict himself to the form of data modelling he has been taught, but is prepared to use different constructs and techniques — of his own invention where necessary — if the nature of the problem requires it. The models described in this book have been proved adequate over a very wide range of problems, but are by no means universally applicable.

Chapter 2

Data Modelling Concepts

Introduction

The objective of analysing the data structure of a particular information system is to express this structure in the form of a *data model.* In a data model, the information is represented by a small number of different constructs. In this chapter, the most widely used form of data model will be described.

There are two philosophical attitudes towards data modelling, known respectively as *semantic relativism* and *semantic absolutism.* According to the absolutist way of thinking, there is only one correct or ideal way of modelling anything; each object in the real world must be represented by a particular construct. Semantic relativists, on the other hand, believe that most things in the real world can be modelled in many different ways, using any of the basic constructs. Marriage, for example may be represented in the model by an entity, a relationship, an attribute, a function, a domain, a constraint, a role, or in a number of other ways. Depending on the circumstances, some of these ways may be more useful, or may result in a more elegant model, but none is uniquely correct. This book adopts a relatavist standpoint.

The philosophical difference does not have to cause any practical problems when semantic relativists work alongside semantic absolutists. The absolutist's quest for the ideal model need not conflict with the relativist's quest for the best model.

Many different modelling schemes have been proposed, offering different sets of constructs from which the models can be built. The basic constructs may include any of those mentioned in the context of the marriage example above. For the purpose of this book, the constructs that will be used to build the data models are *entities, relationships* between entities, and *attributes* of entities.

Having decided how to represent the information in the real world with a set of entities, relationships and attributes, each entity, each relationship and each attribute must be named and precisely defined. Additional information (for example, statistical data) may be recorded where available. Sometimes a picture of the model, known as an

entity-relationship (E-R) *diagram*, is drawn, showing the entities as boxes and the relationships as lines connecting the boxes; such a diagram gives a useful overview but cannot replace the definitions.

What is an entity?

An *entity* is any object of interest to the organization under investigation, any part of the system, any object about which data can be collected and stored. An entity can be real or conceptual; an activity, a passing state or a grouping can be an entity.

A data analysis for an insurance company might describe the business in terms of the following entities:

policyholders
brokers
accidents
policies
claims
payments
vehicles
risks

A vehicle is a real object, a broker is a person, an accident is an event, a policy is a written agreement, and a risk is an abstract classification. All of these may be entities.

In this chapter, I shall use as my main example of an organization a sports authority. Cricket fans may like to think of the Test and County Cricket Board; soccer supporters could think of the F.A.; American readers may prefer the National Football League. The entities of interest to the authority could well include the following:

players
clubs
teams
managers
matches
results of matches
trophies
playing grounds

Exercise 1
List some probable entities for a clearing bank, for an airline, for a local education authority, for the Ministry of Defence.

Entity occurrences and entity types

It is sometimes necessary to distinguish between entities as individuals on the one hand and entities as classifications of individuals on the other. Where confusion would otherwise arise, it is common to speak of *occurrences* for the former entities and *types* for the latter. From our sports example: Old Trafford is an *entity occurrence* which belongs to the *entity type* GROUND.

Exercise 2
Name some occurrences of the entity types PLAYER, CLUB. To which types do the following entities (i.e. entity occurrences) belong: Middlesex, Ian Botham, Washington Redskins, Lords, Kevin Keegan, Super Bowl, Spurs, Cup Final, Roger Staubach.

Relationships

Although the basic component of the data model is the entity, the data model is not merely a collection of entities. The model must be given a structure, which we define in terms of the relationships between entities. Loosely speaking, two entity occurrences are related if the removal of one makes a significant difference to the other. For example, when a married man dies, his wife becomes a widow. If such a connection between two entity occurrences can be generalized and named, so that it can be applied to several similar situations, that is then a *relationship*. For example, the relationship between married women and their husbands is called *marriage*.

Relationships may be permanent and unchangeable, or they may be temporary or transient, only lasting for a short while. For example, the relationship between a buyer and a seller only lasts as long as the sale itself. There are also semi-permanent relationships, such as that between an employer and an employee.

Some of the classes of relationship that can be included in the data model are

logical & causal	e.g.	*is necessary for* *is controller of*
inclusion & membership	e.g.	*is part of* *belongs to*
personal & contractual	e.g.	*is supplier of* *is manager of*

Just as the names of entities tend to be nouns, relationships tend to be