PAGE MISSING
WITHIN THE
BOOK
ONLY

UNIVERSAL
LIBRARY

OU_156422

UNIVERSAL
LIBRARY

OUP—556—13-7-71—4,000.

OSMANIA UNIVERSITY LIBRARY

Call No. 658 · 54 Accession No. 16. 5793

Author E210 Edelson, R.T. & others.

Title Operational research in management

This book should be returned on or before the date last marked below.

OPERATIONAL RESEARCH IN MANAGEMENT

by

R. T. EDDISON, M.A.
Director of S.I.G.M.A. and Editor of the "Operational Research Quarterly"

K. PENNYCUICK, B.Sc., Ph.D.
Consultant Management Scientist

and

B. H. P. RIVETT, M.Sc.
Formerly Head of the Field Investigation Group, National Coal Board

JOHN WILEY & SONS, INC.

NEW YORK

First Edition 1962

Copyright © 1962

R. T. EDDISON
K. PENNYCUICK
B. H. P. RIVETT

PRINTED IN GREAT BRITAIN

GENERAL EDITOR'S FOREWORD

K. PENNYCUICK, B.Sc., Ph.D.

Consultant Management Scientist
Formerly Director of Operational Research
Overseas Food Corporation;
Formerly Senior Principal Scientific Officer,
The War Office

DURING the last decade, one of the features of the business and industrial scene has been an upsurge in the application of scientific methods to the solution of problems. The idea that scientific methods can help the business-man or the industrialist is not new ; it has a history extending back at least as far as the early days of the Industrial Revolution. There has recently, however, been a quite outstanding increase in the efforts devoted to the solving of the more difficult problems in the business-man's life. It seems to be true that the greatest interest has occurred in the United States and the United Kingdom, but it is by no means confined to these countries.

There seems to be a general tendency for industrial problems to become more complex. There also seems to be a general tendency for firms to grow larger : this may be due to amalgamation of small firms to form larger ones, to purchases or to problems of survival where the bigger firm always seems to stand in a more favourable position.

It is also true to say that over the years the accumulated bulk of knowledge on how to run a business is sufficiently well ingrained into the social structure of the countries concerned for the more elementary problems to be at least amenable to some form of treatment. The problems exposed are therefore the more recondite ones, and although there is a popular belief that all business problems can be solved by the judgment of the chief executive concerned, nevertheless it must be admitted that on occasions the chief executive may be mistaken—this mistake may be anything from a minor embarrassment to the

firm to a major calamity which will be followed by the dissolution of the firm.

This belief, that the voice of experience—or flair—will solve all business problems is no more justified than the opposite view that any business problem can be solved by scientific means. What is true, however, is that there are very few difficult business problems that cannot to some extent be lightened by scientific treatment. The scientific approach and the collection of data, its analysis, the deep understanding of the mechanisms of business that may come from such a study, calculation of chances all lead in the direction of making business decisions less uncertain, or at any rate removing much of the uncertainty from the mind of the man who takes the decision.

This series of text-books is being published in the belief that executives faced with difficult or novel problems may find in them some guides to action which will be useful. There are certain types of problem which can, of course, be solved with something very near completeness : some of these lie in the field of stock-holding and of transport and distribution. There are other classes of problems which are perhaps less certain, but, taken together, the books are intended to offer the business executive an opportunity of reading about those parts of scientific method which have been successfully applied to business problems and to learn about current researches which are directed to the application to business problems of scientific disciplines so far not applied. In the latter field much of the work is so new that satisfactory illustrations of successful applications are few and far between. This is no reason whatever for denying the business-man the opportunity of obtaining a fresh approach on his affairs.

These texts are intended to be suitable for consumption by a very wide audience. It is not necessary to be a graduate in Mathematics to read them or profit from them, although it may be necessary on occasion, for some applications, to call in a great deal more expert knowledge than is a normal requirement for an executive position in a business. It is hoped that anybody who has a professional qualification of use in the business world may be able to read these books. There may perhaps be the occasional mathematical expression which is rather difficult, but the object all the time is to explain in as simple

language as possible the application of scientific ideas to the business problem.

It has been said that what is measurable is science, and management science is largely concerned with placing decision making on a quantitative basis.

The parts of these books which can be found elsewhere are mostly found in advanced mathematical treatises or journals which are far from suitable for the business executive—some of them in fact are not very suitable for anybody except the author's use ! There are, however, some things in these books, for the sake of completeness, which may be found in other works, but this is mainly in order to give a link-up with previous work in the field.

K. P.

JOINT EDITORS' FOREWORD

THIS book was conceived some years ago with the prime intention of trying to illustrate the Operational Research outlook and way of thinking. To do so it has drawn on many authors ; the subject is exemplified by problems from many areas ; and it describes a variety of techniques. But the book does not pretend to introduce a new collection of case histories, many of those quoted having been published elsewhere ; still less does it pretend to be a text-book of techniques. The book is intended primarily for the potential user of Operational Research ; if he finds in places that the techniques are described with a fair amount of mathematical symbolism, he is asked to remember that this is there only to show how mathematics, as a way of thinking, can be useful in the right place ; there is no need to follow the mathematical development closely in order to understand the sort of problem to which this applies.

The applications and ideas should, it is hoped, be useful to anyone wishing to start Operational Research and anyone with a limited experience in the field. We hope that even the most experienced and advanced workers in the country and the firms who have the greatest use of Operational Research will still find something of interest, for the net has been widely cast. The book is an introduction both to the subject as a whole and to those methods and techniques that are referred to ; it is neither comprehensive nor exhaustive, but selected lists of useful references for further reading are given.

The editors are grateful to their many friends in the field of Operational Research who have contributed draft chapters in this book. Their names are given in the Contents. In the invitation to contribute drafts a hope was expressed that the writers would be willing to allow the editors considerable freedom with their manuscripts. This permission was freely and generously given. Revision of chapter drafts may have been done by the original drafter after discussion with an

editor. Most changes have been in the interests of homogeneity or brevity, but some have been additions for greater coverage of the subject. The effect in some cases has been to make the final version of a chapter almost unrecognisable to the original drafter. For these reasons, no blame should be ascribed to our contributors for any imperfections in their chapters ; this is the responsibility of the editors. It is, however, only fair to attribute to the original writers the merits that the chapters possess !

The editors, finally, wish to acknowledge with thanks the permission, freely given by Mr. R. L. Bellman of the Rand Corporation and the Princeton University Press, to make use of material from Mr. Bellman's book *Dynamic Programming.*

R. T. E.
K. P.
B. H. P. R.

CONTENTS

I	Introduction *R. T. Eddison*	13
II	The Measurement of Efficiency and Productivity *T. E. Easterfield*	30
III	Queueing and Congestion *E. D. van Rest & F. J. Toft*	44
IV	Stock Control *J. C. Beresford*	95
V	Simulation in Business Studies *W. P. Jolly*	118
VI	Control Systems *Stafford Beer*	128
VII	Linear Programming *A. Baines & K. B. Haley*	159
VIII	Distribution *E. Kay*	180
IX	Labour Wastage and Turnover *B. T. Houlden*	201
X	Military Applications *K. Pennycuick*	212
XI	Data Handling *A. P. M. Purdon & H. G. Jones*	222
XII	Dynamic Programming *M. G. Simpson*	239
XIII	Some Human Problems *Z. M. T. Tarkowski*	258
XIV	Presentation and Implementation of the Results of Operational Research *K. Pennycuick*	288
XV	The relation between Operational Research and other Management Techniques and Disciplines *K. Pennycuick*	299
XVI	Formulation of the Problem and Organisation *K. Pennycuick*	312
INDEX		327

Chapter I

INTRODUCTION

Background

THE purpose in presenting this book is to give the reasons for Operational Research, its importance, and an indication of what is involved in its practice. In this it is hoped to provide something of value to management who may wish to make use of Operational Research ; at the same time the book may serve as a useful introduction to the subject for young men coming into the field as well as providing more senior workers with some references to parts of the subject outside their own particular experience.

Operational Research has been developing with giant strides in the United States of America in recent years and has been making steady progress in the United Kingdom ; flourishing societies exist in both these and other countries. At the first International Conference on Operational Research, held at Oxford in September 1957, twenty-one countries were represented. The subject has long since graduated from the realms of military operations, from which it received its name and is now applied to many aspects of public affairs, industry, commerce and transport.

A second International Conference* was held at Aix-en-Provence in 1960.

Those who seek here for a precise definition of Operational Research will be disappointed. As its name implies it involves research into all or any aspects of conducting, or operating, a " business " and treating the business as a living organism in its proper environment. It is thus clearly distinct from laboratory research ; and the fact that it is directly aimed at the taking of decisions regarding the policy or organisation of the business also distinguishes it from academic research,

* The proceedings of both these conferences were published by English Universities Press, London, in 1957 and 1961 respectively.

although this too may be research into the operation of going concerns.

The methodical study of alternative courses of action has been in the past, is still—and no doubt always will be—called by various names. The name under which it is discussed in this book is " Operational Research ", but that does not imply that any organisation that does not use that specific title stands accused of lack of methodical study of its activities. For example, much of what is done under such titles as Cybernetics, Work Study, Systems Engineering, Consumer Research, or Planning shows great similarity to Operational Research. In some particular instances it could indeed be argued that these activities should be called Operational Research. Generally, however, they are of a specialised nature, dealing only with particular aspects of the whole concern and therefore forming only part of a wider study.

Civilisation is now of great, and growing, complexity. Organisations, processes, production methods, are all complicated, and technological change is both rapid and accelerating. Society as it exists today is the statistical outcome of a large number of small endeavours. In the past, each of the leaders of the small ventures was responsible to himself alone or to a small group. If he failed, the general course of the development of the community to which he belonged was little affected. The more successful communities were those in which the greatest number of individual decisions were right. In other words, progress was largely by trial and error.

Arising from the development of the greater complexities of modern life, three tendencies can be detected which are particularly relevant in the present context. In the first place all businesses and industries of any consequence—and consequence is not necessarily related to size—affect a modern interlocking economy ; furthermore, there is a growing sense of social responsibility, and management is more inclined to consider the impact of its decisions on all who are affected by them—participants, shareholders, employees and the community at large. The effect of this tendency, coupled with the sheer increase in size of so many business decisions, is greatly to increase the seriousness of error, and so to underline the need for conscious methodical study of all

possible courses of action before any of them are chosen and taken.

The second tendency is for modern organisations to pass from the stage of being single units to groups of differentiated, specialised parts with varying degrees of autonomy. The effect of this tendency is two-fold. Firstly it increases the risk of error (as distinct from the increased seriousness of error discussed in the previous paragraph) ; secondly, it makes it virtually impossible for a single human brain and experience to comprehend the whole and control it effectively. It thus becomes necessary to develop specialised techniques and organisations for undertaking the methodical study of alternative courses of action. The chairman of one of the world's largest companies was asked what it was like to be in charge of such an organisation. He replied—to his questioner's consternation : " It is like being in a fog which lifts only now and again to enable one to see a little way ahead ; one must react quickly to such glimpses to correct one's course." Operational Research is one of the management aids available to help lift the fog.

The third tendency is the increasing tempo of modern affairs. The art of management has sometimes been described as the ability to make time to put one's feet on the table and think. The speed of developments today makes this more and more difficult. All too often, the urgent displaces the important.

In all this the development of modern society has much in common with biological evolution. The first primitive creatures, consisting of a single cell capable of carrying out all the essential functions of life, have gone far to develop into modern animals with highly differentiated parts co-ordinated by a brain, central nervous system and other control devices. Modern organisations, whether they be industrial concerns, social structures or state activities, are at the present time going through the stage of developing their central controlling mechanisms ; this they must do if they are to remain viable. Electronic computers and other types of equipment as yet unborn may provide the machinery for this process ; cybernetics may provide the theory ; Operational Research may provide the techniques for understanding what goes on and why and what decisions and controls are needed.

Nature of Operational Research

Later chapters will deal with techniques. But sophisticated techniques do not make Operational Research, even if it sometimes provides an outlet for their application ; they come later and are often really the easiest part of the problem.

The essence of Operational Research lies first in the acceptance of the principle that logical thought backed by careful observation and methodical analysis is a necessary (though rarely a sufficient) basis for decision-making ; and secondly, in the determination, when undertaking the work, to produce the appropriate answer by the appropriate time. So far as techniques are concerned, it will often be found that methods of only quite childish simplicity are required—methods which certainly do not provide material for a book. Nevertheless the simple and obvious approach is frequently shunned for fear of appearing silly. This chapter therefore develops, without apology, simple ideas of what may be called the philosophy of the subject. If it lights the way with occasional blinding flashes of the obvious, so much the better.

As has already been said, a precise definition of Operational Research is not attempted. The reason for this is not just cowardice, nor can the difficulty of definition be regarded as a weakness in the subject. As with most subjects, it is easy enough to understand when one is at the centre of it, but it is extremely difficult to draw clear-cut boundaries where it shades off into other subjects. It is with margins that definitions are concerned if they are to be of any value. Few people would ask for a definition of Physics or Chemistry, yet how many could write a sentence which satisfactorily distinguished one from the other and the boundary between them ? Just so with Operational Research ; some activities clearly are Operational Research, others are clearly not ; others again are, or are not, according to the personal whims of the definer.

Most attempts to define Operational Research—and there are many—include a statement that it is " a scientific method ". This, however, begs the question, for where is there agreement on a definition of scientific method ? Scientific method, says one, is the method used by scientists in the course of their scientific investigations. James B. Conant,[1] on the other side

of the fence, prefers to think that there is no such thing as "the scientific method". After all, what is there in the methodical type of approach used by the scientist that is distinct from the approach of the scholar doing research into history or other humanities? In Conant's view the aspect of method which distinguishes the scientist is the use of experiment.

Use of Experiment

This is very pertinent for, as has already been said, Operational Research is basically the application of methodical analysis and logical thought to the study of alternative courses of action. If these cannot be regarded as the preserve of the scientist, is it right to call Operational Research a scientific method? Let it be said at once that historians, linguists and classicists—to name but a few—have proved extremely effective members of Operational Research teams, and they may have a unique contribution to make to many types of problem. Nevertheless, Operational Research must undoubtedly be considered to have a special link with science rather than with the other methodical disciplines for, although it obtains its data from actual on-going processes rather than from laboratory experiments, its outlook is generally experimental and its techniques are frequently those developed for the analysis of experiments.

The scientist conducts his experiments to find out what happens in order to determine how it happens and why it happens. In other words he is interested in the underlying laws of his science. The experiment is not an end in itself, although the ability to plan and conduct careful experiments neatly designed in relation to their objectives is the hallmark of the scientist.

A difficulty, however, is at once apparent. An experiment may establish what happens in the special circumstances in which the experiment is conducted; from these facts, it may be possible to deduce a law which could explain them; but it is never possible to say with complete certainty that the law deduced is the law that does explain them. For the possible laws are infinite and one can never test, or think of, them all. It is for this reason that the scientist generally tackles the problem the other way round, first propounding a principle

which forms an hypothesis regarding the law he wishes to discover, then working out what results he expects to flow from this hypothesis and finally carrying out experiments and observations to see whether the facts are consistent with his hypothesis. It is only after the hypothesis has been rigorously tested by extensive and well-planned experiments over many years that it becomes accepted as a law. "The principles are tested by the consequences derived from them." [2]

This is the basis of inductive reasoning. It does not, of course, overcome the difficulty of conclusively and uniquely establishing a law. Indeed, the history of science has many examples of accepted laws being subsequently proved wrong, or at least no more than part of the truth. The power of the method depends upon the ability to select reasonable, or likely, hypotheses in the first place ; this may be done either from initial exploratory experiments or from abstract reasoning based on experience. Subsequent experiments will be designed to test a small number of possible hypotheses.

The fundamental scientist may be satisfied when he has discovered the underlying laws that he is seeking. The applied scientist will want to see what use he can make of them. The Operational Research scientist is, by definition, only concerned with those laws which are of direct interest in the process of decision-making by executives. Management does not dwell in the past ; what has been, has been. The analysis of yesterday's events, however scientific, is of value only if it provides a guide to indicate what results may be expected tomorrow from various courses of action which may be adopted today ; it is with just this object in view that Operational Research workers attempt to discover the underlying laws governing operations.

Difficulties in this Method

The scientist who obtains his information about what happens from experiments is faced with three difficulties in relating his observations to the law which he seeks. The first is the problem of time : his observations are in the past ; strictly speaking they can, therefore, be related only to laws that operated in the past, as there is no evidence that the law will continue to operate in the future ; if the law is applied

to the future, one is in effect extrapolating in time. This is a point that has much bothered philosophers, but as a philosophical argument it can be ignored, for in practice one has no alternative but to work on the assumption that one is not approaching a discontinuity in time. Nevertheless, the Operational Research scientist, who is not dealing with the perpetual laws of nature, must guard with particular care against the risk that the laws he discovers for the past only operate in special conditions which will no longer apply in the future.

The other two difficulties are that in making his observation the scientist may affect the behaviour of what he is studying, and that, however comprehensive his experiments may be, the scientist's observations can never be more than a sample of the whole. These difficulties present special hazards to the Operational Research scientist. His objective is to find out what happens in a going concern. He has broadly two sources of information ; he can either make direct observations on the organisation while it is working or he can undertake an analysis of the records of its past working. The behaviour of an organisation depends on the activities of the people in it, and the mere fact that they are under observation will almost certainly affect their behaviour. This is sometimes referred to as the " Hawthorne " effect after the famous experiment at the Hawthorne plant where it was so clearly demonstrated.[3] Although this effect cannot be completely eliminated, there are various tricks for guarding against it, such as the snap-reading (or ratio-delay or activity ratio) method first developed by Tippett in the cotton industry.

The dangers arising from the fact that the observations only constitute a sample can be reduced by ensuring that enough observations are made under a sufficient variety of conditions. Sufficiency can here be estimated by the proper use of the statistical theory of errors. In this connection, planned experiments rather than random observations can be particularly advantageous. But planned experiments involve controlled conditions and these increase interference with normal routine.

The study of routine records of past operations has an advantage over direct observations in that it does not give rise to a Hawthorne effect. On the other hand, the accuracy of records is often dubious ; it is rarely that they provide

complete information on all the points required ; and all too often they fail entirely to mention changes in some important variable. For these reasons, research based solely on routine records without any direct observations is inadvisable. Where a mass of information is available either from observations or from records, it may be impossible to handle it all and some method of selection of the data becomes necessary ; it is then vital that the selection be made under proper statistical control to ensure that a suitable and unbiased sample is used.

These are but some of the pitfalls that beset the way of the Operational Research worker in the first, and what may seem to the uninitiated the easiest, stage of his work—the task of finding out what happens. The important part that statistical theory may play in helping him has been indicated, and the statistical techniques themselves will be referred to more fully in later chapters. Perhaps the greatest difficulty in Operational Research, however, is created by the time factor. The manager for whom the work is being done must make a decision one way or another, and a fairly good answer to his problem at the right time may be far more useful to him than the perfect answer too late. Furthermore, as with any other expenditure, cost must always be balanced against the probable value of the result. It can often happen that some simple application of Operational Research can produce a good answer quickly ; it may then be quite wrong to spend much money and effort on elaborate research to produce a slightly better result much later. For this reason, conclusions may be applied without tests of full scientific rigour. If the work is not impeccable it must be used with due caution, but it may provide the best information available.

Establishing an Hypothesis

After finding out what happens, the next step is to establish an hypothesis about how and why it happens, in order that this hypothesis may be used to predict the effect on future operations of making changes in the factors that can be varied. One way of doing this is by deducing the theory directly from the data that have been collected. Such an approach nearly always involves the application of some form of statistical analysis, such as regression analysis ; this is by no means

always possible, for standard statistical methods have been developed mainly for dealing with comprehensive information collected from experiments, and the information collected from the study of operations is often insufficient for this purpose. Furthermore, operating conditions are often very complex owing to the large number of variable factors involved and this makes statistical analysis correspondingly difficult. Where these difficulties arise merely from the volume of calculations required, not from inadequacy of the statistical theory, much can now be done on electronic computers whose advent has made possible calculations of a complexity that could not previously be contemplated.

An example of this type of approach, in which only fairly simple statistical analysis was required, is to be found in the work of Thornthwaite[4] into the growth of peas, from which it was possible to determine the effects of the time of planting and of certain key weather factors on the rate of growth. At the other extreme may be cited attempts which have been made to analyse the factors affecting the operation of blast furnaces used in smelting iron. Multiple regression analyses of major proportions have been successfully completed on electronic computers, but the work proved largely sterile owing to the difficulty of arriving at a logical interpretation of the results. This difficulty illustrates a basic danger that must always be guarded against, namely, that a purely statistical connection between two factors must never be interpreted as demonstrating a law of cause and effect unless it conforms in general to a causal connection which can be shown to be expected on other grounds.

When the information obtained from observations is not suitable for usual methods of analysis, a better approach may be to derive an expected law from other sources and then compare this expectation with the observed data to see whether they fit. This corresponds to the inductive method of science as already discussed. Such expected laws are sometimes described as " models " and they may be derived in a number of ways ; sometimes by purely theoretical reasoning, sometimes by a quasi-empirical approach, sometimes by the use of analogues.

An example of the use of a theoretical model can be found

in the work done by the British Iron and Steel Research Association on the discharging of iron-ore ships[5] referred to again in Chapter III. This involved a queueing problem. The ships arrived at the ports at irregular intervals; if, therefore, the discharging equipment provided at the ports were barely sufficient to discharge the total tonnage of ore during the year there would at times be very long delays to ships waiting for a berth to become vacant. At the other extreme, if sufficient discharging equipment were provided so that it could always discharge all ships expeditiously, however many arrived together within a short period of time, then there would inevitably be long periods when the discharging equipment would be idle because there were no ships to discharge. The problem was to determine how much equipment should be installed to obtain the greatest economy, having regard to the idle time of both ships and equipment. The need to balance capacities in this sort of way is typical of many industrial and social situations, and there must be few examples of Operational Research on record in which the striking of a balance in some form or another does not play an important part.

In order to determine how changes in the port installations would affect the delays to ships, the actual distributions of ship arrival times and of discharging times were first found from available records. The mathematical theory of queues was then developed to provide estimates of the expected delays, in relation to discharging capacities, for these distributions. The actual delays at particular ports with their existing discharging capacities were then compared with the theoretical expectation. As these were found to be in reasonable agreement it was possible to accept the theoretical expectation as a basis for predicting the effect of changes in the port installations.

Often the situation is too complex to determine any expected law by purely theoretical reasoning. In such cases it may be possible to adopt a quasi-theoretical approach. For example, the ship delays referred to above could have been derived by setting up a working model on a sheet of paper. Hypothetical arrival times could have been generated in accordance with the known characteristics of the actual arrival distribution (for example, by drawing numbers from a hat or by using a table

of random numbers, to conform to a random distribution). Similarly, a distribution of discharging intervals could have been obtained and the delays to be expected determined by working through the ships one by one. This approach is the basis of what has come to be known as the Monte Carlo method, which is described more fully in Chapter V. It simulates reality, but at a greatly increased speed. Work with paper and pencil, though sometimes sufficient, may become tedious. On a more ambitious scale, electronic computers can be used effectively, and hypothetical operations which would take years in reality can be compressed into a few seconds of time. A further refinement of this approach is to build a special analogue to simulate the real situation. A warning must be inserted, however. Simulation of this sort requires careful controls to ensure that it is not dissimulation, and to embark recklessly on large and expensive computer programmes without first demonstrating their validity can lead to disaster.

The processes of setting up laws or models form an important part of Operational Research, and examples will be found in later chapters in this book ; for example, in Chapters III and IV cases are quoted in which laws have been derived from the theory of queuing and stock-holding, and Chapter V illustrates the use of simulation.

The general methods of attack so far outlined all presuppose that the essence of the problem is expressed in the form of figures that are susceptible to mathematical or statistical treatment. It is true that a large number of Operational Research problems are of this type and statistical methods are the most common tool of the Operational Research worker. It is also true that Operational Research must establish measures and criteria of success. Nevertheless much—and by no means the least important—research requires no more than the logical breakdown of operations and the sorting of their components without any elaborate mathematical computation. It is not possible to classify the methods which may be applicable to such problems, but to limit the discussion only to those methods which can be readily classified would be to present a very biased picture of what may be involved in Operational Research.

Use of Electronic Computers

An example of this less statistical type of problem is the study of the possible uses of electronic computers as a tool for industry and commerce. Here is a new instrument which, by virtue of its speed and the complex operations it can perform, can entirely revolutionise ideas regarding the processing of data for the purposes of industrial and commercial control in the widest sense of the word. At the present time great, and increasing, amounts of labour are being devoted to the collection and handling of operating data for the preparation of management statistics and accounts. All this is done solely to enable management to understand what is going on and to control the business ; and this becomes daily more important as businesses increase in size and complexity. Yet, in spite of all the effort and expense that goes into it, in how many cases can the controls be regarded as effective ? All too often the control can only be described as too little and too late. Computers may change all that, as they make it possible to contemplate undertaking, as a matter of routine, calculations which are at present unthinkable because of their length and complexity ; furthermore, these may be done more quickly and more cheaply than is now possible. The injection of a true servo-mechanism into the controls of business is becoming a practicable possibility.

If the most effective use is to be made of computers in this way, it is quite clear that some radical re-thinking of the whole control process in each case must be undertaken. It will be a great tragedy if computers are introduced simply as a means of speeding up or cheapening the existing method. Already one often hears it said that computers enable a choice to be made between providing better statistics with the same staff or providing the same statistics with less staff. This is to take a much too simple view of the new opportunities. In most concerns, statistical controls have been built up over the years by adding one thing to another, with the result that many of these systems are top-heavy ; yet they provide the foundations for empires, so that pruning is difficult. The study of computers provides an opportunity to start again from scratch with fundamental questions about what statistics are needed and

how they can most effectively be obtained and processed. The final answer may well be that the controls can be made simpler as well as cheaper and quicker. Planning the use of these new tools offers a challenge to Operational Research. It is of interest to compare it with the first task to which the name Operational Research was applied—the way in which the then new military equipment, radar, could best be used.

Operational Research and Decision-Making

Whether the problem is basically numerical or non-numerical in form, however, it remains true that the method of the Operational Research worker is essentially scientific. But this is obviously not enough in itself to give a clear description of Operational Research. For then it is necessary to define in addition the context in which the scientific method is used. It is the essence of Operational Research that it is applied directly to assist in the making of decisions affecting the management of on-going businesses ; in other words, it forms a basis for executive decisions.

For the present purpose, three steps leading up to the making of a decision can be distinguished : the first is seeing where a decision is needed, the second is selecting the right objective, and the third determining the best way of attaining that objective. Operational Research may have a part to play in any or all of these three stages.

The first essential is to perceive where a decision is needed. Executives will probably object that they can usually see, without any assistance, too many decisions that need to be made, not too few, and it is better to leave well alone those processes where the drill has long been established and which do not thrust problems upon them. But it is often in just such cases that most can be gained by taking a fresh look : and, if this is to be done, someone must be given time to do it unhampered by urgent day-to-day decisions. All too often it will in fact be found that the underlying requirements have changed, but the old-established course of action has not been modified to keep pace with the changes. The introduction of a new outlook in such circumstances was probably one of the most important reasons why Operational Research proved so successful when applied to military problems during the war ;

and they are probably the circumstances in which Operational Research can achieve the most sensational results.

The well-known example of the study of depth-charge settings illustrates just this point.[6] The long-established criterion for the best settings no longer applied under modern conditions when the moment of attack by aircraft on submarines could follow so quickly after spotting the submarine on the surface. It cannot be expected that such big dividends can always, or even often, be obtained, but undoubtedly examples of similar importance will occur from time to time.

The clarification of the objectives for which the executive is working in making his decisions is a subject which often receives far too little attention. Much is heard, and rightly, about the importance of improving productivity, a subject dealt with in Chapter II. It is often possible to improve productivity of one resource simply by improved organisation and without increasing the consumption of some other resource. More often, however, it is a problem of balancing the use of one resource against another ; for example, to what extent is it right to increase the wastage of manufacturing material (in other words, to reduce its productivity) in order to increase the productivity of labour ? The right answer will depend on whether high labour productivity is the objective in itself or whether the overall economy is to be considered. To take another example, it may be possible to study the use made of locomotives in a steelworks, to find that they have quite a lot of idle time, and so to prove that the works should have fewer of them. But if that were done, what would be the effect on the service provided by the transport department for the production department ? A steelworks is there to make steel—some would say to make money—not just to run locomotives efficiently. Study of individual components in a business must be made in the context of the whole. This wholistic approach is an important part of Operational Research. The selection of the right objective is, of course, a matter for the executive, not for the scientist. Nevertheless the scientist can greatly help by setting out clearly the alternative objectives in any given situation.

It is in the third stage, that of deciding the best route to the objective, that Operational Research has its most obvious part

to play. And it is at this point that the special techniques for handling operational data have their greatest scope. Every decision made by every executive in every organisation is based to some extent on past operational data. Most of such decisions are based upon the unconscious collection and analysis of the data that have been thrown up in the course of the life of the individual (or group of individuals) who makes the decision. This process is called experience. But experience is often unreliable ; not only has it too narrow a base, but it tends to be biased in its selection of the data from the past. Scientific methods can remove that bias and ensure that the data are properly treated to give a true picture of the past ; they can determine by a rational process what rules appear to regulate the system being studied and hence show what is the probable outcome of any particular action ; furthermore they can provide a measure of the " imponderables " by calculating the probability of abnormal occurrences.

This is not to say, however, that Operational Research provides or can provide the answers to everything. It has many limitations. T. H. Huxley wrote " There is only one right but the possibilities of wrong are infinite."[7] In Operational Research, there is rarely one right answer ; it is a matter of choosing the best, or the one that has the most chance of being the best, out of a number of possible likely rules. The infinite number of wrong alternatives remain and it is quite impossible to test all rules against the facts. In practice, therefore, one must start by selecting the few most likely rules ; and the decision as to which are the most likely rules will be made as often from experience, intuition or inspired guess-work as from the analysis of data. In this sense the " practical " man is the one who has " good " theories, and that is usually the man with executive experience and judgment.

The function of scientific analysis is then to determine which of the theories thought worth testing fits the facts best, and how well it fits them. It can show in what ways the data are insufficient and suggest how further necessary data may be obtained. With all that, however, Operational Research is usually dealing only with probabilities. It does not say " this is the right decision to make ", but " this is probably the best decision ". A knowledge of probability is clearly very useful

when one is making a series of repetitive decisions, for the probability of success is high in the long run. But executive decisions are commonly seen as unique actions—one decides to do this or that and it proves to be right or wrong. In such circumstances it may be of little value to know that the probability of success is 19 out of 20 if in fact the one chance failure turns up. Decisions based on probability must be tempered by some consideration of the effects of failure, or of the need to succeed. In operating public transport, for example, the requirements of safety must be paramount. The smallest calculated risk cannot be allowed even if it were to make possible huge operating economies. This type of decision is not uncommon and, so long as a certain risk of failure is permissible, it is possible to make a rational approach to such problems, taking account of both the need and the expectation of success.

It is, however, rare that management decisions need to be regarded as unique actions. They nearly always involve a series of consecutive actions spread out over time. So, even after the initial decision has been taken, there is still much that can be done to ensure success. This is an important difference between normal business decisions and gambling decisions. In the latter, once the decision has been taken and the money staked, there is little that the individual can do to influence the result. It is most important that Operational Research should be organised to take advantage of the continuing nature of managerial decision-making. The absolute necessity of this is obvious when it is remembered—and it cannot be too often stressed—that the executive has to make decisions on the best information available to him at the time ; he cannot postpone them until the perfect answer is available, although one of his most frequent types of decision is whether to act now on the information available or wait in the hope of better. Operational Research must be geared to this reality.

REFERENCES

[1] CONANT, J. B. (1951). *Science and Common Sense* (Yale University Press).

[2] BRONOWSKI, J. (1952). "The Logic of Experiment", *Advanc. Sci.*, **9**, 289–296.

[3] ROETHLISBERGER, F. J. (1951). *Management and Morale* (Harvard University Press).

[4] THORNTHWAITE, C. W. (1953). "Operations Research in Agriculture", *J. Opns., Res. Soc. Am.*, **1**, 33–38; or, in a shortened form, in: (1954). *Operations Research for Management* (Johns Hopkins Press), Vol. 1, pp. 368–380, McCloskey, J. F., and Trefethen, F. N. (eds.).

[5] EDDISON, R. T., and OWEN, D. G. (1953). "Discharging Iron Ore", *Operat. Res. Quart.*, **4**, 39–50. (See also Chapter III.)

[6] BLACKETT, P. M. S. (1953). "Operational Research—Recollections of Problems Studied 1940–45", *Brassey's Annu.*, 88–106.

[7] HUXLEY, J. S. (1938). *Science and Education* (Macmillan & Co.).

Chapter II

THE MEASUREMENT OF EFFICIENCY AND PRODUCTIVITY

Object of Productivity Measurement

In many Operational Research studies, it will be found either that the effective working of the unit studied is being limited by a shortage of one or more resources (e.g. manpower, plant, working capital), or that other evidence suggests that even though such resources do not form a bottleneck it would be desirable to get more output for a given input of one or more of them. (By output is meant here any measure of the achievement of the unit : for example, physical production, profit, casualties inflicted on the enemy.) It is convenient to use the term "productivity" for any measure of physical output per unit of an element of input, or per unit of a weighted index of several elements of input. Much of this chapter will apply to measures of output without restriction as to their nature.

Defining a measure of output does not, of course, by itself lead to useful knowledge. Knowledge of the performance of one plant—or other operating unit—at one time is of little interest, except possibly as evidence (weak, but better than nothing) of the resources that would be required if the plant were to be expanded or a similar plant built. Productivity studies have been directed rather to finding the variations in productivity, relative to one or more elements of input, either between a number of comparable plants, or from time to time at the same plant, or between comparable bits—machines, working teams, etc.—of one or more plants. It is a matter of definition how many of these studies should be classed as Operational Research ; since, however, the methods and approach are likely to be of use to Operational Research workers, they are described here.

Such studies of variation of productivity have two main classes of use :

(i) They can throw light on factors affecting productivity : on how far different types of equipment, different systems of organisation, different types of supervision and so on, affect the output of a given productive unit. At the crudest level, the observation that, for example, some plants have outstandingly high or low productivity makes them worthy of closer consideration. This is not to say that the plant of high productivity is necessarily better ; but it is likely that the reasons why it stands out will be helpful in improving others. And though low productivity and low efficiency are not synonymous, it is likely that study of plants with low productivity will point to cases where efficiency could be raised.

(ii) By the analysis of figures for a number of plants or a number of periods, or both, target figures for requirements of labour, raw materials, plant, etc., may be deduced which will lead to better production planning and forecasting. In particular, some of the more sophisticated methods of analysis may lead to information about the rate of exchange between one input factor and another (for example, how much labour will be saved per unit increase of equipment of specified type). When such information is combined with cost data, the way is open for finding the optimum combination of resources for a given output.

Study Made by the Shirley Institute

The classic example of the study of a large number of comparable units is the work of the British Cotton Industry Research Association (commonly known as the Shirley Institute) on productivity in cotton spinning.[1] In 1945 the figures were collected at a number of spinning mills of the production (in pounds) of yarn of different counts.*

More recently, the European Productivity Agency (a branch

* The count of yarn is a measure of its line density ; it is the number of hanks that weigh a pound. The length of a hank varies with the type of yarn ; for cotton yarn it is 840 yards.

of the O.E.E.C. formed in 1953) has made comparisons on an international basis between cotton spinning mills.[2]

The research team of the Shirley Institute collected the number of man hours worked, together with many technical details of the types of machinery used, the staffing of the various sections of the mill and so on ; the figures of output were taken not only for the final spun yarn, but also for various intermediate stages for which records were obtainable, as the cotton progressed from the initial stage of the bale of raw cotton to the final yarn. It was convenient in these studies that although any mill would make yarn of a range of counts, that range was, as a rule, fairly small ; thus it was not unfair to take for each mill at each stage the total production as being at the mean count of the range produced. It was therefore possible, by taking the figures for production and for man hours, to get simple figures for output per man hour. In fact the Shirley Institute found it convenient to work in terms of labour cost, that is to say, operative hours per unit of output (in this case 100 lb. of yarn). This will be known hereafter as O.H.P.

The mere variation of O.H.P. from mill to mill is itself illuminating. It was clear, however, that not only for the overall production of yarn but also for production at each of the various intervening stages O.H.P. depended strongly on the count of yarn spun. When the figures for the various mills in the sample were plotted on a graph giving O.H.P. against count, the points were found to lie in a fairly coherent strip through which a mean curve could be drawn, but there was considerable variation about this, the highest O.H.P. at any count being about two or three times the lowest. It was therefore natural to seek for causes of this variation.

The striking thing that emerged from the attempt to find other factors that might account for the differences was that, for the mills taken as a whole, nothing that was suggested appeared to show a significant effect. It was to be expected, for example, that yarn made with poor quality, cheaper cotton, or yarn which had been more highly twisted, would take more labour to make it. The variations in O.H.P. due to these factors, however, were either zero, or so small that they were completely masked by other factors of organisation, personality, or whatever one may like to call it, so that only count appears

to be relevant. In the analysis of figures for particular processes, however, it was possible to isolate certain technical features of machine speeds and settings and to attribute effects to them also. Nevertheless there was still a large residual and unidentified variation which was attributed to a "management" factor. Studies in other places have suggested what sort of factors may be relevant in this connection. For example, studies in the shoe industry[3] have shown strong effects of such things as production planning and incentive payment systems. More detailed studies of an academic nature have shown the effects of certain human factors such as types of foremanship and degree of close supervision.

It may be of interest that the Shirley Institute have recently made a repeat survey of a number of firms and have used the information from the earlier survey to remove the effect of changes of count between the two surveys. The figures show on the average some improvement between the surveys in productivity in ring spinning (the commoner process). However, in those parts of the spinning process for which the effect of technical factors can be allowed, it has become clear that almost all the improvement lies in better technical settings. There is on the average no improvement on the residual side. The other thing that is striking is that firms have considerably altered their order of relative productivity. There is evidence from a number of places that firms vary in their relative productivity, from time to time, and indeed that the fluctuations may be comparatively rapid.

A second example of productivity measurement and analysis is given by the work of Stafford Beer at Samuel Fox & Co.[4] This work, which was intended to improve production planning and control, is mentioned in some detail in Chapter VI for its applicability to control problems. In this example in a single firm two things were achieved. First, figures were obtained which provided a basis for gauging machine capacities and for estimating the dates of fulfilment of orders with a good deal of confidence ; it was also possible to plan out what targets should be set as an indicator of efficiency. Secondly, there was a drill for a rapid calculation of these dates and for the regular scrutiny of the production records to see if efficiency was rising or falling.

These two cases between them exemplify most of the uses to which productivity measurement may be put, and most of the problems which arise in the course of trying to carry it out. These may be classified under three headings : reducing output to some sort of comparable measure, defining input, and the search for the factors which may cause productivity to vary from time to time or place to place. Of these the measurement of output has probably given the most trouble and the most references in the literature. The trouble is, of course, that output is very seldom any simple thing that can be measured in a constant unit from point to point, although such cases do arise : for example, refined sugar, coal, and flying hours have turned up in productivity studies and have been measured in tons of sugar, tons of coal, and hours, without any problems arising of comparing one sort with another. (Coal, of course, does vary considerably in type and quality, but many studies have been made on the assumption that this variation can be ignored.)

The Need for a Comparable Measure of Output

In most cases a variety of goods or activities will be produced which have in some way to be made commensurable. In some simple cases one or two qualities may be sufficient to group the products so that to each group a unit of output can be put in a fair relationship to other units. For example, in the Cotton Industry study described overleaf, the only thing that appeared to have a recognisable effect on productivity in spinning was the count of the yarn spun. It was therefore reasonable to draw a smooth curve to lie evenly among the points representing the average man-hour requirements per unit output for mills spinning particular counts of yarn. Any individual mill could then be compared with a notional mill spinning the same count and with a requirement lying on this curve. Various suggestions have been made for getting such factors for particular products, some of more validity than others. In the steel rolling case mentioned above, it was possible by detailed statistical study to get target figures for machine hours per ton rolled for various types of steel and these figures could be given with some knowledge of their limits or error. It has been suggested, from time to time, that

for many engineering operations the figures produced by time study could be of value. On the other hand this view has been strongly attacked. Whether such figures are usable almost certainly depends on the type of comparison that is to be made. For comparisons from time to time at one factory or other unit where the market basket of goods produced will vary only slightly, and where the same men are making the time studies (so that one of the biggest sources of variation is eliminated), it is probable that time study figures are quite good enough for comparisons. Errors in the weights when the market basket is varying only slightly will have a trivial effect. On the other hand, for comparison between productive units, figures produced by time study are likely to be doubtful in view of the known variability of the rating technique, even if they are produced by the same observer, and if produced by different observers they are almost certainly valueless.

The conclusion of the special committee of the I.C.W.A. and I.P.E.[5] was that in the present state of the art of time study, the variations in standards and consistency of time study officers made standard times set by these means no basis for inter-firm comparison of productivity. Tests by Winston Rodgers, Kinniburgh and Vallance[5] and other workers have shown a very low consistency in rating. These tests have been attacked on the grounds that the officers tested were not chosen to be highly skilled in the art; certain firms claim that by internal training and testing they can get a high degree of consistency; unfortunately no results on which to check this claim have yet been published, nor have any tests been made to show whether time study officers showing consistency within firms would also show consistency between firms in rating the same tasks. The I.C.W.A./I.P.E. condemnation of the use of time study for interfirm comparison therefore appears still to be justified. Within a single plant, for reasons given later, average times either as measured or corrected by rating are probably a usable basis for inter-temporal comparisons. For comparisons between branches of one company all making the same sort of output, it is a matter of internal testing whether raw times, times corrected by rating, or other measures can be used for comparisons.

Another attack on this problem has been made by the

British Boot, Shoe and Allied Trades Research Association (S.A.T.R.A.)[3] by reducing the products to one standard product. It is clear that this method is applicable only when the goods vary but little among themselves. In the S.A.T.R.A. studies a number of factories all making the same general type of shoe were considered—for example, twelve factories making "Goodyear" welted men's shoes. Such shoes were reduced to comparability by translating the factory figures to those for a "standard" shoe, made by the basic operations required to make a pair of shoes of the particular type. If the factory used extra operations they would be ignored, their times not counted in. If operations were done by a non-standard method an attempt would be made to translate these figures into those for the standard method by any means available, but usually on the principle that the piece rates in the factory for the standard and non-standard methods would be proportional to the basic amount of work involved. Thus S.A.T.R.A. was able to produce for each factory, and for each operation, a number of man minutes per dozen standard pairs involved to show the range of variation between firms and to go a long way to identifying factors that were usually associated with high or low productivity.

The method used by the Bureau of Labor Statistics is a variant of that for which standard weights are applied to the products.[6] In the development of its methods the Bureau was undoubtedly influenced both by the fact that initially its aim was to get figures which could be combined into a global index for the United States economy and also by the need for a method which could largely be applied by the accountants at the firms studied. In its approach to an industry it would draw up not only a sample of firms in the industry but also, after discussion with the firms, a list of standard products. In the base year it would find the production of these products and the number of man hours required to make each of them. In subsequent years it would take the quantities of these products made and the number of man hours required to make the total output of them ; one could then compare the number of man hours required for these with the man hours that would have been required at the productivities for the individual products of the base year. The B.L.S. then used

these figures to show scatter within the industry from year to year and trends for the whole industry.

Defining Input

To turn now to the question of inputs, these are in general more easy to identify, and the problem is rather that of deciding what to do about the relationship of the productivity of one factor to that of another. In some cases one factor is so clearly the important one that it is sufficient to take productivity in terms of that one factor alone. For example, during the war, in studies of the relationship of flying to the bottlenecks limiting it, it was found that as a rule maintenance manpower was the sole limiting factor except for certain particular operations and types of aircraft. Flying hours per maintenance man were therefore the thing that was particularly studied in the development of planned flying—though it was realised at the time that in other circumstances different limiting factors might be important. Most American industrial studies also concentrate on output per man hour almost exclusively, probably because of the tradition in the States of a relative labour shortage arising from an open frontier to which labour could always go.

However, the emphasis on labour to the exclusion of everything else has come under strong attack in Europe. Clearly raw materials, fuel and capital may all be important. Two approaches are then possible. One is to measure the productivity of each of these factors (output per man hour, output per kilowatt hour, output per ton of coal, etc.), look at the fluctuations of these from place to place and from time to time, and see if there are any correlations between them. The other possibility is to try to get some sort of a joint production function by statistical or other means. These two, of course, merge into one another. Methods for dealing with them probably have to be developed *ad hoc*, but it is worth drawing attention to two particular cases. First, in the study of production functions for farms, J. O. Jones[7] has made a number of statistical studies which are a somewhat simplified approach to multiple correlation analysis, in which the type of analysis is designed to show up differences in type of relationship from one end of the scale to the other and changes from straight line to curvilinear regression if these should appear.

Roughly speaking, his method has been to take the output of the farms and to relate it to a number of factors such as acreage, man hours put in over the year, tenant's capital, expenditure on fertilisers. These figures were used by classifying the farms into groups on two of the factors and looking for trends in each *group* for output per unit of a third factor. By this means Jones was able to develop production functions with a high degree of confidence. One of his results was distinctly unexpected, namely, the lack of apparent effect of tenant's capital on farm output. This result seemed inescapable from the figures he used, surprising as it may seem ; it gives room for a good deal of conjecture as to its possible cause.

A number of writers have attempted to develop productivity indices that will somehow take all factors into account. For example, Sir Ewart Smith and Dr. Beeching[8] have suggested the estimation of all the man hours that have gone into a product (including those required in making all raw material ; bought out parts, fuel and power used in the manufacture, and those which are notionally consumed by the depreciation of the capital used in its manufacture), and comparison of output with this overall figure. The estimation of such a figure presents formidable problems and little has been published to show how it could be done in a practical case. One attempt at such a calculation has been given by W. Langenburg.[9] He remains, however, slightly sceptical about the value of such an overall figure when it has been obtained. It has already been pointed out that the usefulness of a productivity figure is not its indication that a firm should be praised or blamed, but its spotlighting of the parts of the firm, or the particular firms, that should be looked at, or the light it throws on the causes of variation from place to place. The trouble about such an all-embracing figure, even if theoretically impeccable, is that it is so very difficult to interpret. For example, the figures which show how output per man hour, output per unit of fuel, output per unit of raw material and so on all vary, may throw a good deal of light on the exchange rate between one of these inputs and another, but the overall index which manages to bring all the inputs nearer into one denominator will hide this and it is very difficult to see what one can learn from the figures. From the businessman's

point of view, of course, there is one way of lumping together all inputs, namely, by the prices at which they may be bought. Clearly, it is of interest to get as high an output as possible, preferably in terms of selling price per £1 worth of input. The inputs are put together on the basis of their cost; but cost accountancy has not usually been considered a branch of productivity measurement. This is perhaps a pity; the combination of the study of the real causes of variations of output per unit of labour or raw materials and of costs would probably be most valuable to production engineer and cost accountant alike.

Identifying the Factors which cause Variations in Productivity

To turn to the third group of problems, that of identifying the factors which cause productivity to vary, one can only say here that a wide variety of types of factor have been studied. Which factors are looked at will depend partly upon the working hypothesis of the Operational Research worker, but partly also upon which factors show possibilities of being changed. Working conditions, capital equipment, technical process, training of the worker, organisation, state of trade, length of run, human relations in the firm, design of product, all these factors may affect the productivity of the firm. Theoretically it is possible to try to cover all factors, and this is in fact being done in a study at present being carried out on three firms by a team at the Oxford Institute of Statistics[10] but to attempt to cover all factors at once requires extremely detailed first-hand study for a long time, and even then it may be very difficult to see the wood for the trees. It is certainly not something to be attempted in every case where it is thought that efficiency may be varying from time to time or from place to place.

It is impossible to go here into all the various things that might be studied, but a few cases may be mentioned as being suggestive. The cotton spinning studies referred to above produced a good deal of information about the effect of machine speeds and settings on labour productivity. The S.A.T.R.A. studies also took into account the effect of quality, factory organisation, and systems of payment. The Building Research Station's analysis of the costs of building on sites showed up

effects of number of houses, size of house, features of design, and degree of subcontracting.[11] Using the number of houses on the site as a parameter, their results tally with those found for a number of other industries, in which productivity continues to rise as a small power of the total production of a particular product, for an astonishingly long time.[12] J. A. Mark[13] has made studies of the effect of age on labour productivity, and C. D. Harbury[14] has compared urban with rural workers.

It may be possible to study the effect of such imponderable things as human relations even where no numerical measure of productivity is available. A good deal of work at the University of Michigan[15] and at Oxford has been devoted to studying the effects on productivity in a firm of the relationship between the foreman and his team and the type of organisation in the firm. Their method was to get pairs of teams such that both teams of each pair were doing work generally agreed to be comparable and such also that there was general agreement which of the teams was of the higher productivity. Then it was possible to look for measures of statistical association between high productivity and such characteristics as authoritarian or democratic relationship between foreman and men, degree of training given by the foreman, the extent to which a foreman kept the men in his confidence, and so on. These results, though containing no numerical measure of productivity, are probably perfectly valid.

Many warnings must be issued: since the time of the Hawthorne experiments at the General Electric Company it has been known that worker performance may depend quite largely on the attitude of the workers, notwithstanding changes in conditions. In the Hawthorne experiments the conditions of a group of workers were changed and it was found that every change led to a rise in productivity, even though some of the changes were only reversals of previous ones, and it appeared that the operative factor was the feeling on the part of the team studied that they were worthy of individual attention and hence people who had some special contribution to make. This effect on their attitude to work was far more important than the changes in physical conditions.

There is not as a rule room for elaborate statistical methods

in studying the factors that may affect productivity; far more important are first-hand observations that will lead to useful hypotheses about likely factors and the collection of sufficient material for these factors to have a chance to show their effect.

One plausible statistical approach, that of finding the labour (or other input) used day by day and the corresponding outputs of a list of known products, and using regression analysis to find the input required per unit of each product, has proved to lead to nonsensical results in several studies (all, perhaps understandably, unpublished communications). It is likely that the variability in productivity from day to day on isolated jobs, and the tendency for records to pick up errors of their own, introduce too much randomness into the figures for them to stand such detailed treatment. A good account of the difficulties produced by errors, arbitrariness, omissions and other peculiarities of works records is given by Luttrell ;[16] it cannot be over-emphasised that the origin and meaning of any figures found in firms should be scrutinised very carefully before they are used for purposes other than those for which they were intended.

It must always be remembered moreover that productivity is liable to vary from time to time, and there is probably room for investigation of the causes of this fluctuation. Several studies have shown that if firms are ranged in order of productivity not only will the range from highest to lowest, for any particular input factor, be quite large, but also that within this range the particular firms are liable to change their order from time to time in quite short periods. There is room for study of the reasons why these variations, from time to time, in the order occur, and why also some firms appear to be able to survive in competition with others which are apparently a great deal more efficient technically. It may be that those which seem to be most efficient fall back after a time and some of the less efficient climb up in response to the goad of poor results or for other reasons, or it may be that changes in trade conditions favour first one firm, then another. There is room for investigation here, though this should probably be regarded rather as industrial economics than as Operational Research, except possibly if it is carried out within a multi-plant firm which requires the information for immediate business decisions

and is concerned rather with the operations of the components of one firm than with a general study of the variations in efficiency within industry as a whole.

REFERENCES

[1] TIPPETT, L. H. C., and VINCENT, P. D. (1953). "Statistical Investigations of Labour Productivity in Cotton Spinning", *J. R. Statist. Soc.*, **A 116**, 256–67.

[2] VANDEN ABEELE, A. M. (1957). "A Method for Analysis and Comparison of Labour Productivity in Cotton Spinning Mills", *Productivity Measurement Review, Special Number.*

[3] COOPER and MARTIN (1950). "The Productivity of Twelve Goodyear Welted Shoe Factories", *J. Brit. Boot and Shoe Institution*, **4**, 364, and a number of subsequent papers in the same journal.

[4] BEER, S. (1955). "The Productivity Index in Active Use", *Applied Statistics*, **4**, 1–14.

[5] *Measurement of Productivity*—Interim Report of the Joint Committee of the Institute of Cost and Works Accountants and the Institution of Production Engineering (1949). (London, Macdonald and Evans).

KINNIBURGH, W., and VALLANCE, L. S. (1951). "The Accuracy of Time Studies in Bricklaying", *Engineering, London*, **171**, 353.

RODGERS, W. M., and HAMMERSLEY, J. M. (1954). "The Consistency of Stop-Watch Time Study Practitioners", *Occup. Psychol.*, **28**, 61.

FORNALLAZ, P. F. (1950). "Neue Untersuchungen auf dem Gebiete der Schätzung des Leistungsgrades", *Industrielle Organisation (Zürich)*, **19**, 541.

[6] *Measurement of Productivity* (Methods used by the Bureau of Labor Statistics in the U.S.A.—Report of a Group of European Experts) (1952). (Paris : O.E.E.C.)

[7] JONES, J. O. (1955, 1956). "Productivity of Major Factors in British Farming", *Fm Econ. Oxf.*, **8.4**, 1, and **8.5**, 11.

[8] SMITH, SIR EWART, and BEECHING, R. "Measurement of the Effectiveness of the Productive Unit", *B.I.M. Winter Proceedings 1948–1949*, No. 4.

[9] LANGENBURG, W. (1952, Jan.). "An Experiment in Productivity Accounting", *N.A.C.A. Bulletin*, 584.

[10] COLE, H. J. D., HOLLAND, D. G., and POSNER, M. V. (1960). "Factory Productivity and Efficiency", Part I. *Bull. Oxford Institute of Statistics*, **22**, 151, Parts II–IV (*in press*).

[11] *Productivity in Housebuilding* (1950). H.M.S.O. National Building Studies, Special Report 18 ;

Productivity in Housebuilding (1953). Second Report. H.M.S.O. National Building Studies, Special Report 21.

A brief account is given in :

STONE, F. A., and REINERS, W. J. (1954). " Organisation and Efficiency of the Housebuilding Industry in England and Wales ", *J. indust. Econ.*, **2**, 118.

[12] BERGMAN, H. (1951). " Productivity in the Airframe Industry ", *Econometrica*, **19**, 329.

ANDRESS, F. J. (1954). " The Learning Curve as a Production Tool ", *Harv. Bus. Rev.*, **32**, 87–97.

HIRSCH, W. Z. (1956). " Firm Progress Ratios ", *Econometrica*, **24**, 136–43.

[13] MARK, J. A. (1956). " Job Performance and Old Age ", *Washington, D.C. Bureau of Labour Statistics Bulletin 1203* summarised in *Monthly Labour Review* (1956), **79**, 1410.

[14] HARBURY, C. D. (1958). " The Industrial Efficiency of Rural Labour ", Cardiff : University of Wales Press.

[15] KAHN, R. L., and KATZ, D. (1953). " Leadership Practices in Relation to Productivity and Morale ". This is Chapter 41 of *Group Dynamics* (Cartwright and Zander. Evanston Ill.: Row, Peterson & Co.). It gives references to the original papers turned out by the Michigan School on this subject.

ARGYLE, M., GARDNER, G., and CIOFFI, F. (1957, 1958). " The Measurement of Supervisory Methods ", *Human Relations*, **10**, 295. " Supervisory Methods Related to Productivity, Absenteeism and Labour Turnover ", *Human Relations*, **11**, 23.

[16] LUTTRELL, W. F. (1954). " Some Problems of Interplant Cost Comparisons ", *Accounting Research*, **5**, 107–20.

RUDD, E. (1955). " The Accuracy of Data Collected from Firms ", *J. Indust. Econ.*, **3**, 72.

GENERAL REFERENCES

EASTERFIELD, T. E. (1959). *Productivity Measurement in Great Britain.* (London : Department of Scientific & Industrial Research.)

Manual of Productivity Measurement, Vol. 2 : *Methods and Results* (1956). (European Productivity Agency, Paris.)

Chapter III

QUEUEING AND CONGESTION

Introduction

AFTER he has selected the goods he wants, a shopper in a self-service store must take them to a desk to be checked and paid for. He may be served immediately or may have to wait for other customers to be served first. The customer would like always to have immediate service ; the store manager would like to see the cashier busy most of the time. These requirements cannot both be met, as customers do not arrive regularly and some take longer to serve than others. If the cashier is usually busy, a queue will develop when customers arrive more frequently than usual or take longer than usual. If customers find that they often have to wait, they may decide to shop elsewhere in future ; if the queue gets too long it will begin to block the shop. For these reasons, the manager might consider employing another cashier. He knows that it will double the cost of the checking service, but what effect will it have on average waiting time or queue length ? Will the extra cost be justified by avoiding loss of customers or by the extra shop space made available ?

This is the kind of problem to be studied in this chapter. The basic feature is that an input of some kind of unit is provided with a service by a service channel. The problem arises because of the irregularity in the rate of input, or service, or both, so that they cannot be matched exactly and the best compromise is sought. This queue situation can occur in a wide range of operations and its existence is not always obvious. One of the earliest examples to be studied was the automatic telephone exchange where the units are callers and the service channels are switches at the exchange ; in this case, if the channels are all busy, unsuccessful callers do not form a queue. In a factory, a product may be processed by a number of machines in turn ; if the processing time is not constant a queue problem arises, and the input to one service channel is

the output from the previous one. Finished products may accumulate in the despatch bay and be sent away in batches by lorry ; in this example, the service channel (lorry) serves several units simultaneously—a process referred to as bulk service.

Before a queueing problem can be analysed, information is needed about the input, the service and the behaviour of the units. The rate of arrival of units (or the intervals between arrivals) and the service time may be specified by frequency distributions. The arrival and service rates may be affected by the length of the queue, or may change with time, and the service rate may depend on the type of unit being served. In many problems, mathematical expressions can be found which give a good enough approximation to the actual distributions, but sometimes the actual distributions—found by observation—must be used. The behaviour of the units is described by the " queue discipline ". When a queue forms, the units may be served in the order in which they arrive (as in the shop example) or in arbitrary order (as may happen with goods in a factory) or some types of unit may have priority. Provided that the order of service is not related to service time, it does not affect the queue length or average waiting time, but it does affect the time an individual unit has to wait. Other types of queue discipline are for units to leave a queue after waiting a certain time (they may be, for example, impatient customers or perishable materials) or for them not to join a queue which is greater than a certain length.

If the average rate of arrival is less than the average rate of service, and both are constant, the system will eventually settle down into a steady state ; the probability of finding a particular length of queue will then be the same at any time. If the rates are not constant, the system will not reach a steady state, but it could remain stable. If the rate of arrival is not less than the rate of service, the system is unstable and the probability of a long queue steadily increases ; this happens even if the two rates are equal on average, because the time lost when the service channel is idle can never be made up. Imposing a limit on the maximum length of queue (so that further arrivals are not accepted) automatically ensures stability. Queueing situations which are unstable for a limited time are common in practice—rush-hour traffic is an example.

The Use of Models

In common with other O.R. problems, queueing problems require the choice of a model which shall represent the basic elements of the system, often in a simplified way. It is necessary, then, to observe the system to determine suitable values for the parameters of this model, to ascertain the effects on the model of changes in these parameters and to translate these changes back into effects on the actual system studied. As with other problems, the value of the model lies not only in its ease of manipulation as compared with the actual system, but in its power to describe the system shorn of the many complications in the shape of specialised circumstances which are always to be found in any real system. The advantages of thus seeing the skeleton, as it were, of the system are that certain general principles can be applied straight away as soon as the legitimacy of the model is proved ; and secondly, of course, that the simplified model is much more easily explored for the effects of changes of parameters.

Thus the techniques involved in an attack on a congestion system are concerned first with choosing and fitting a model ; second with manipulation of the model, and third with the interpretation of the results obtained from manipulation of the model in terms of the real system.

Although the study of processes involving random demand and congestion does not extend far back in time, there are available a number of models whose behaviour has been studied. The choice of an appropriate one is not, however, always simple ; for a reason which will be made apparent later, the same system may sometimes be described by more than one model and one may have advantages over the others. A certain flexibility of mind and power to generalise are needed to choose the model ; not only to choose the one that best represents the system, but the one likely to be most fruitful in interpretation of changes.

An example both of the need for flexibility in interpretation and of possible alternative models is provided by a special stock—or inventory—problem. Stock, in discrete quantities, is never more than c (packets) ; at the time of each sale a re-order is made ; times for delivery are distributed (as χ

squared). Demands when stock is zero do not result in a re-order. Interest lies in the proportion of lost sales. At first sight, it might seem to be necessary to attack this by compounding probabilities, first of arrival of a demand in random fashion at time t, and second the replacement of the item of stock at a later time t_2. However, by considering the purchases as demands on a re-order service which has c servers and distributed service times, stock is zero and a sale is lost when all servers are occupied, i.e. c undelivered orders in being. This is exactly the situation for which Erlang's loss coefficient in telephone theory was calculated.[1] Thus the transformation of meaning of service from the sale to the re-ordering process makes the system describable by a well-known model. The same system may alternatively be described in a cyclic system[2] with two phases and queues in each phase. One phase is the undelivered order, the other the unsold stock on the shelf. The solutions for this model give the probabilities of various partitions of the numbers c between the phases, thus giving directly the probability that none is in stock, i.e. that a sale is lost.

The fitting of the model involves observation of the system, chiefly of the form and parameters of the distributions of the various quantities involved, usually arrival intervals and service times. Such observation involves sampling and the interpretation of samples. Where the system is complicated and a model simpler than the system is to be used, the observations must be in sufficient detail to enable some reconstruction or "correction" of the data to the simpler conditions. An example of this is the construction of loading times under model one-wharf conditions from observations made on ships partly unloaded at adjoining wharves, in the industrial example considered later in this chapter. Unless the observations are made in great detail such correction is not often possible.

The techniques of taking samples can be the subject of discussion. Where records are ordinarily kept for other purposes, it may be possible to use these, wholly or in part; where a part only of the records is to be used, this part should be chosen in such a way as to avoid bias, to explore trends and to give low variance estimates of the quantities observed. While random choice—either of time of observation or of ordinal

number of event—is a universal way of achieving the first of these requirements, the method of choice may need to be modified to achieve the others. A lower variance can sometimes be achieved by taking observations not by the "snap" method usual, for example, in Work Study, but by observing what sets of the different events occur in a number of short periods whose beginnings are chosen at random.[3]

The manipulation of the model in order to ascertain the effects of changing parameters will usually result from a mathematical analysis, though, if the system proves intractable, it may be necessary to resort to experiment under the simplified conditions of the model.

Interpretation of the results of manipulating the model brings the realisation that, because of the statistical nature of the system, the answers will be statistical. That is to say, the information desired—the number in the queue, the number turned away, the waiting time of each—will be given in the form of frequency distributions, statements of how often each of the possible values will occur. The merit of a system from the planner's point of view will not be judged solely—if at all—on the mean values of such quantities. It is more likely that the planner will be interested to know such characteristics as frequency of no wait, frequency of wait longer than one, two or other numbers of average service periods, frequency of encountering a queue greater than a given size, length of continuously busy period for the service not often exceeded, and so on; these are quantities which need to be planned, since they govern the provision of service, of service aids, and of accommodation for the queue.

If the model has been chosen to represent only the basic, simple part of the system, shorn of its special occurrences, it may be necessary to effect a reconstruction from the model results by replacement of the special occurrences. Unless these are well understood, it is better to assess their effect separately; that is, to consider that the investigation has aimed at improvement of only the basic part of the system. An example of successful division of the problem, its separate investigation and subsequent reconstruction is the work of Benson,[4] on ancillary work done by operators attending textile machines.

Experimental or Monte Carlo methods are, of course, frequently employed in queueing as in other O.R. problems. Their value, in this class of problem at least, lies chiefly in the stage of manipulating the model where this proves to be intractable mathematically. It is usually easier and more profitable to experiment on the model than on the actual system ; for one thing, the effects of changing parameters are more readily interpreted and interpolation or extrapolation carried out.

Monte Carlo methods have their own problems which, being common to all Operational Research, are dealt with separately in Chapter V.

An extensive literature on congestion problems now exists.[5] Many models have been analysed mathematically by a variety of methods, and several special types of congestion situation have received particular attention. Some of the methods and special situations are described later in this chapter.

An Industrial Example

A congestion problem of some importance is that of ships arriving at a port for unloading. The cost of delays while a ship and its crew wait—perhaps outside a port for a favourable tide, again inside for a berth to be available, and again for unloading facilities—can be considerable. Ships cannot keep exactly to schedule, nor is it always possible to allot berths in advance. The picture is the same as that of the service counter on a much larger—more costly—scale.

There is the complication that the provision of port facilities is under the control of numerous authorities, each with its own problems of finance and planning and dealing with trades in many different materials. There is the danger therefore that any expansion of the port facilities will lag considerably behind any projected expansion of the processing capacity for an imported raw material, so that no such plan can be considered complete which does not either include means of handling the additional material or take account of the congestion and consequent additional costs which may arise at the ports involved.

As a contribution to the general planning problem of bringing iron ore to this country, the British Iron and Steel Research Association undertook an investigation[6] to find to what extent

the usual congestion models apply to this problem and to what extent delays due to congestion could be predicted from a knowledge of the volume of traffic and of the means of handling it.

A considerable volume of data about the movement and unloading of ore ships over a period of eighteen months was available to the Association. From these the data relating to six ports were selected as being more amenable to interpretation in terms of simple models than the others. Even at these six, however, there were many peculiarities of operation.

It might be expected that all transport would be so planned that ship arrivals would occur to a known schedule, or with only slight deviation therefrom. Although all supplies were planned, and ordered, and ships chartered, by a central organisation, the number of ships engaged was large—though not unlimited. The journey times were extremely variable, since ships of a variety of shapes and sizes were involved and all were subject, in varying degrees, to the weather. Many of the ships were required to carry return cargoes perhaps from ports other than the one used for unloading. There was also a seasonal fluctuation. The problem was complicated by the practice of diverting traffic—when severe congestion was foreseen, or actually imminent—to the neighbouring ports. The actual times of berthing, which are the ones to be considered when measuring congestion, were affected by further delays at those ports to which entry could be obtained only at certain states of the tide. Some ports were served by ships of such a size that they could enter only on spring tides. In consequence, a number of ships were liable to enter the pool at the same time with resulting congestion inside the port, because of the few berths available. Different types of ships required different types of unloading equipment, and there might therefore be a restriction on the type of berth used. Sometimes a ship would berth and partly unload at one berth, subsequently transferring to a more suitable berth to finish unloading. The time of berthing was also affected by the current use of the port by other types of ships not concerned with ore, a use which, not being subject to any unified plan, was bound to introduce an element of randomness into the availability of berths.

With all these complications, it might be thought that any theory would need to be very complex to achieve predictions with even a fair agreement with reality. In fact, however, as is generally true with problems of this complexity, it is possible to make predictions with fair accuracy, by shearing the data from the individual ports of their individual peculiarities and by requiring the model to deal only with the remaining common element of near-random arrival and distributed unloading times. To make predictions in any particular case therefore involves a knowledge of what these other conditions, forming the part of the problem that has been shorn away, are. Since, however, the predictions are usually required to foretell the effect of a change in the subject under study (in this case, ore ships) whilst other conditions may be supposed to remain constant, this is no real drawback.

The main parameters of a system serving random demand are those of the distributions of arrival and of service times. Examining the data in order to discover these, the following peculiarities at the six ports were noted :

Neither at port A nor at B were arrival times affected by tides, but at B delays were sometimes averted by diversions to another port. The distributions for the two ports combined were therefore examined, as well as for B alone.

At ports C and D, berthing occurred after passing through a lock which ships entered almost always in daylight and within one hour of high tide. It is therefore assumed that only alternate tides are used.

At port E, there were a large number of semi-independent wharves and the times of arrival noted did not uniformly refer to the same point, with the result that clearly some of the observations included a tidal restriction, while others did not.

At port F, the arrival times observed were those after passing through a lock, which could take place during night or day, but must be within one and a half hours of high tide. A number of ships unloaded at nearby wharves were omitted.

At ports C, D and F, ships sometimes waited outside the ports for reasons other than the tides.

The simplest assumption about the arrival times would be that the arrival is equally likely at any time, how likely being governed by the average rate of arrival characteristic of each

port. Although this may seem a drastically simple assumption to make, it does turn out that, shorn of the effects of some of the major factors which can be allowed for, the arrivals were in fact in fair agreement with such an assumption. This is not so remarkable when one remembers that so-called randomness is always produced by the combined action of a large number of causes of deviation from regularity.

TABLE 3.1.1. *Distributions of arrival intervals compared with those expected from random arrival*

Interval (days)	*Frequencies*					
	Port A		*Port B*		*Port B**	
	act.	*exp.*	*act.*	*exp.*	*act.*	*exp.*
0–1	9	10	55	59	84	85
1–5	7	8	44	43	59	57
2–2	8	7	29	29	31	36
3–4	8	6	20	20	20	22
4–5	4	5	17	14	17	16
5–6	2	4	15	10	12	9
6–7	5	4	7	7	6	6
7–8	2	3	6	5	6	4
8–9	4	2	4	3	4	2
9–10	4	2	2	2	2	2
10–20	9	9	1	5	1	3
over 20	0	2	1	—	0	—

* Including diverted ships.

The extent of the agreement was tested by comparing observed and expected distribution of lengths of intervals and of observed and expected numbers of arrivals in periods, first of one week and then of four weeks. Although these are alternative tests, they would each reveal a rather different kind of departure from the assumption. The comparison of the observed and expected distributions of intervals is shown in Tables 3.1.1, 3.1.2.

From these, and from the other two comparisons, it can be concluded that ore ships' arrivals are very nearly at random ; but that at most ports—as might be expected—the extreme intervals do not occur.

At port F, more control appears to have been exercised than at the others ; at port E, there is evidence of much less control.

At those ports entered through a lock, or where there were tidal restrictions on movement, the difference between the observed distribution and one derived from the randomness assumption would not be large enough to affect prediction of congestion delays. On the whole, if randomness is assumed,

TABLE 3.1.2. *Distributions of arrival intervals where tide restrictions exist compared with those expected from random arrival corrected for the restrictions*

	Frequencies									
Interval (tides)	*Port C**		*Port D**		*Port E†*		*Port E‡*		*Port F*	
	act.	*exp.*	*act.*	*exp.*	*act.*	*exp.*	*act.*	*exp.*	*act.*	*exp.*
0	31	31	22	18	6	10	3	6	16	20
1					10	17	5	10½	25	35
2	36	49	32	31	12	15	9	9	26	29
3					14	13	12	8½	29	24
4	39	34	14	23	17	11	10	8	24	19
5					9	10	7	7	18	16
6	36	24	17	18	8	9	4	6	14	13
7					8	8	3	5½	13	10
8	17	17	13	13	10	7	8	5	9	8
9					12	6	9	4½	11	7
10	10	11	13	10	6	5	9	4	6	6
11–15	20	21	27	24	20	17	13	15	18	17
16–20					7	8	8	9	2	6
21–25					4	4	3	5	2	2
26–30	2	3	7	6	1	2	3	3	0	1
over 30	0	1	0	2	0	2	4	4	0	—

* Small frequencies for odd numbered tides shared between neighbouring even numbered tides.

† One wharf only, all arrivals including part unloaded elsewhere.

‡ Same one wharf, primary arrivals only.

the resulting estimates of delays may be expected to be in general a little too high. At port E the use of berths for part unloading and other forms of co-operation between wharves results in a comparatively steady stream of arrivals at berths, and the random hypothesis is not necessary. At port F, the arrivals were restricted by the number of vessels which can pass through the lock on one tide and by the fact

that vessels may wait several days for a sufficiently high tide ; for these reasons, agreement here is poor.

The other parameter of a congestion system is service (or unloading) time. This will usually be a distributed quantity and it is necessary to know the form and parameters of the distribution. First there is the matter of definition. The most useful definition in connection with congestion is clearly that which includes in the service time the whole of the time that one ship denies the berth to others. Thus the time taken in manœuvring, tying up, waiting for labour and machinery will all be included in the service time, which will extend from the moment a ship begins to approach a berth with the intention of tying up until it has vacated the berth and moved far enough away for another ship to begin to approach. Only so can a period of occupation of a berth by a series of ships be taken as continuous.

Since the data used were not collected for the purpose of this study, it was not always possible to arrive at the service time so defined. In two ports, only the actual unloading times were available. As a matter of interest the distribution of these times was studied as well as that of the times deduced by the addition of allowances for manœuvring, etc. In most of the ports, there was the added complication of ships other than ore ships interfering with movement and sometimes with the use of berths.

At only one berth were observations of actual service as thus defined available. At the others additions—sometimes also based on observation—had to be made. At one port, where some partial unloadings were carried out and where two small ships would sometimes unload at one berth, it was possible from the data available to reconstruct the delays which would have occurred under straightforward working and thus test the applicability of a theory applicable to the general case.

The distributions of the estimated service times (i.e. with additions as described above) are reproduced in Table 3.2. The differences in mean values are to be explained by the differences in the equipment available for unloading and by differences in the hours and shifts worked. The alternate large and small frequencies appearing in the distribution for port B were caused by the short hours of working which, with a limited

TABLE 3.2. *Distributions of unloading or service times*

Time (days)	*Port*					
	A†	*B*†	*C*	*D*	*E*	*F*
0–½	1	1			3	
½–1	25	2		1	6	2
1–1½	10	8	3	2	16	41
1½–2	11	3	5	2	24	62
2–2½	4	14	13	14	18	35
2½–3	7	1	10	2	15	25
3–3½	3	15	29	11	11	14
3½–4	0	5	24	12	10	12
4–4½	1	19	21	12	13	7
4½–5	1	5	20	11	8	9
5–5½		14	18	13	2	5
5½–6		3	4	6	4	1
6–6½		10	9	18	5	1
6½–7		5	7	3	3	
7–7½		14	10	11	0	
7½–8		5	4	2	3	
8–8½		11	5	6	1	
8½–9		4	3	3	1	
9–9½		17	1	6	0	
9½–10		4	2	3	2	
10–10½		7	1	2		
10½–11		2	0	2		
11–11½		10	1	1		
11½–12		1	0	0		
12–12½		1	1	1		
12½–13		0	1	1		
13–13½		4		1		
13½–14		0				
over 14		13				

* See text for definition of service time.
† Only unloading times were available.

amount of overtime, ensured that ships—which almost always began unloading in a morning—finished unloading also during a morning, so that times consisting of an odd number of half days are less frequent than those of an even number. The main point of interest for this study lies, of course, in the shape of the distribution. In Fig. 3.1 the distributions for the

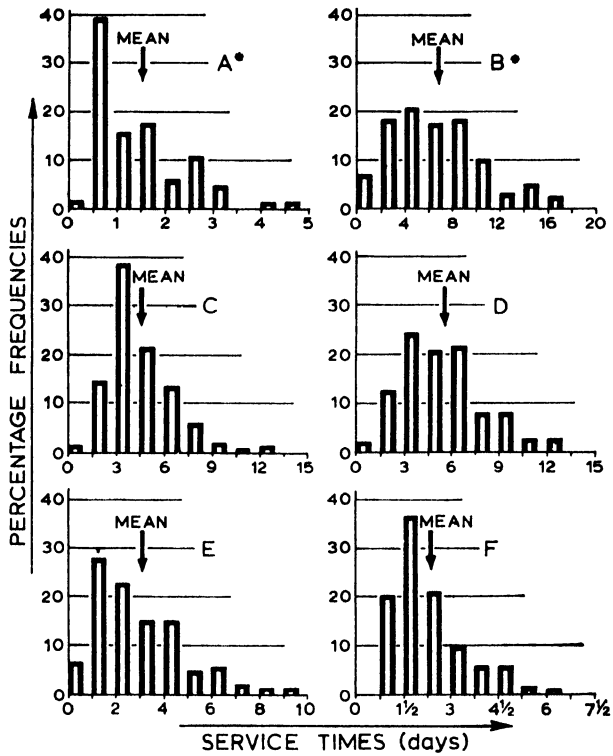


FIG. 3.1

different ports are plotted, with the times scaled according to the mean time of each, the frequencies being regrouped to preserve approximately the number of classes. The similarity of shapes is obvious, the spread of each—as scaled down—is approximately the same. The actual values of the co-efficient of variation (standard deviation divided by the mean) are :

A	B	C	D	E	F
0·57	0·56	0·45	0·46	0·60	0·50

The minor differences in shape are readily explained. At port A, there was a proportion of small ships, quickly unloaded, the remainder being considerably larger. At E, part cargoes gave the same effect ; at B, wagon shortages also contributed to the variability of service times. At C, two shifts a day and sometimes Sunday working were common.

As might be expected, the service time distributions are not of the simple negative exponential form that can safely be assumed in the case, for example, of a telephone call. The negative exponential distribution is generated by a random beginning and ending to a call, using the word " random " in the sense that it was used in describing the generation of arrival times—that is, the event is equally likely at any moment. Here, where the service times are known to be compounded of several parts, it is reasonable to expect a gamma distribution—a form which would result from the addition of times from several such exponential distributions. It is convenient, for further application of theory, that the coefficient of variation for such distributions is intermediate between the value zero it assumes for constant service times and that of unity for the simple negative exponential form.

It is concluded from this study that the application of the ordinary formulae of congestion theory using distributed service times is justified, discrepancies being recognised as due to such causes as diversion and the speeding up of unloading during busy periods and to the special characteristics of the ports themselves.

It is therefore possible not only to make predictions of likely delays under given conditions, but to formulate general, economical conditions. The great advantage to be obtained from constancy of service time which becomes apparent in the solutions of almost all congestion problems shows itself here as an advantage of the uniformly sized fleet over the fleet of mixed sizes of vessels. The intensity of traffic at a port is measured in this study by *berth occupancy*, a quantity which is identical with the proportion of busy time of the server in the general congestion problem. By plotting the curves relating average delay to berth occupancy, the economics of any particular plan of port management can be evaluated. The relationship of the average delay to the berth occupancy follows

the usual shape characteristic of most congestion problems, with the rapid rise in average delay for small increases in traffic intensity at the high intensities.

Mathematical Methods

Solution by Differential Difference Equations

The earliest method of analysis applied to queueing problems was the use of differential difference equations. This method is described first and applied to the simplest queueing situation. More recently, integral equations and Markov chain analysis have been developed to solve queueing problems and these are mentioned later.

The method depends on deriving the rate of change of the probability of a given state of the system. This is done by considering the probabilities of transitions (e.g. by the arrival of a unit or the completion of service) in an infinitesimal interval. In order that a problem can be solved explicitly by this method, it is necessary that these probabilities should be fairly simple functions. This means in practice that the method is limited to a few types of input and service time distributions.

The simplest queueing situation has an input of identical units, an unlimited queue served in order of arrival and a single service channel, the probabilities of an arrival and of a completion of service (of a unit already in the service channel) being constant, independent of time or of the state of the system. The constant probability of occurrence implies a Poisson distribution of the events and an exponential distribution of intervals between events. Thus, if the average rate of arrival of units is λ in unit time,

Probability of one arrival between time t and $t + dt$ is $p\,dt$

Probability of n arrivals in time t is $\frac{1}{n!}(\lambda t)^n e^{-\lambda t}$

Probability of an interval between t and $t + dt$ before the next arrival is $\lambda e^{-\lambda t}\,dt$

Probability of an interval greater than t is $e^{-\lambda t}$

Average interval between arrivals is $1/\lambda$

If the average service time is $1/\mu$, similar expressions, with μ in place of λ, apply to completions of service, provided that the service channel is busy ; if the channel is unoccupied the probability of a completion of service is zero.

In this simple situation, the state of the system is completely described by the number of units in the system (including any being served). Let $P(n, t)$ be the probability of the system being in the state n at time t (i.e. that there are n units in the system). Consider the ways in which the state $n - 1$ may be reached at time $t + dt$; they can be tabulated as follows :

	State at time t		*Arrivals in dt*		*Completions of service in dt*	
Way	*No. in System*	*Probability*	*Number*	*Probability*	*Number*	*Probability*
1	$n-1$	$P(n-1, t)$	0	$1-\lambda\, dt$	0	$1-\mu\, dt$
2	$n-1$	$P(n-1, t)$	1	$\lambda\, dt$	1	$\mu\, dt$
3	n	$P(n)$	0	$1-\lambda\, dt$	1	$\mu\, dt$
4	$n-2$	$P(n-2, t)$	1	$\lambda\, dt$	0	$1-\mu\, dt$

All other ways (and the second one in the table) involve two or more events in the short time dt so that their probability is of the order of dt^2. The probability of the state $n - 1$ at $t + dt$ is the sum of the probabilities of the possible ways of reaching the state. If terms in dt^2 are neglected it can be written :

$$P(n-1, t+dt) = P(n-1, t)(1-\lambda\, dt-\mu\, dt) + P(n, t)\mu\, dt + P(n-2, t)\lambda\, dt \text{ for } n > 1$$

If $n = 1$ the state $n - 2$ cannot occur and there is no possibility of a service finishing in the state $n - 1$, so the equation is slightly altered and becomes

$$P(0, t+dt) = P(0, t)(1-\lambda\, dt) + P(1, t)\mu\, dt$$

When dt tends to zero, the fraction

$$[P(n-1, t+dt) - P(n-1, t)]/dt$$

becomes the rate of change of $P(n-1, t)$ and by dividing the equations by dt then

$$\frac{dP(n-1,t)}{dt} = -(\lambda+\mu)P(n-1,t) + \mu P(n,t) + \lambda P(n-2,t)$$

$$\frac{dP(0,t)}{dt} = -\lambda P(0,t) + \mu P(1,t)$$

These equations can be solved to give explicitly the probability of n in the system at any time. It can be shown that, provided μ is greater than λ, the system tends to a steady state in which the probabilities are independent of time. In this section, only this steady state will be considered ; the rates of change of P are zero so that :

$$\mu P(n) = (\lambda+\mu)P(n-1) - \lambda P(n-2)$$

$$\mu P(1) = \lambda P(0)$$

where $P(n)$ is the steady-state probability of n units in the system.

Hence $P(1) = \lambda P(0)/\mu = \rho P(0)$ putting $\lambda/\mu = \rho$

and $P(n) = (1+\rho)P(n-1) - \rho P(n-2)$ for $n > 1$

so that $P(2) = (1+\rho)P(1) - \rho P(0)$

$$= \rho(1+\rho)P(0) - \rho P(0)$$

$$= \rho^2 P(0)$$

Similarly n can be made 3, 4, 5 in turn, giving

$$P(n) = \rho^n P(0) \qquad (\text{for } n \geqslant 0)$$

The sum of the probabilities from $n = 0$ to ∞ must be 1.

$$1 - \sum_0^\infty P(n,t) = P(0,t)\sum_0^\infty \rho^n$$

$$= P(0)/(1-\rho)$$

Hence $P(0) = 1-\rho$

and $P(n) = \rho^n(1-\rho)$

The average number of units in the system is

$$N = \sum_0^\infty nP(n)$$

$$= (1-\rho)\sum_0^\infty n\rho^n$$

$$= \sum_0^\infty n\rho^n - \sum_0^\infty n\rho^{n+1}$$
$$= \sum_0^\infty n\rho^n - \sum_0^\infty (n-1)\rho^n$$
$$= \rho/(1-\rho)$$

The probability that the service channel is idle is $P(0)$. The average utilisation of the channel is $1 - P(0) = \rho$. This may be regarded as the proportion of time the channel is busy or the average number of units in the channel. The ratio of input rate to service rate, ρ, is often known as the traffic intensity, which in this case is equal to the utilisation.

The average number of units in the queue (excluding any in the channel) is then $\mathcal{N} - \rho = \rho^2/(1-\rho)$.

The probability that a queue exists is the probability that n exceeds 1, which is $1 - P(0) - P(1) = \rho^2$.

More generally, the probability that the number of units in the system exceeds n is

$$P(>n) = 1 - \sum_0^n P(r)$$
$$= 1 - \sum_0^n \rho^r(1-\rho)$$
$$= \rho^{n+1} \quad \text{(by a calculation like that above for } \mathcal{N})$$

The waiting time of a unit, from its arrival to the time it starts to be served, can be considered in two parts. There is a finite probability that the waiting time will be zero; this is the probability that there are no units already in the system, or $P(0) = 1 - \rho$. There is also the probability that the waiting time will be from t to $t + dt$; let this be $w(t)\,dt$. This waiting time occurs only if the last of the units already in the system finishes its service in the interval t to $t + dt$, all the others having been served during the period 0 to t. During this time the service channel is continuously busy so that the random completions of service give a Poisson distribution of completions in a given time. That is, the probability of $n-1$ completions in the time t is $(\mu t)^{n-1} e^{-\mu t}/(n-1)!$. The probability of one completion from t to $t + dt$ is $\mu\,dt$. The probability of n units being in the system when one unit arrives is $P(n) = \rho^n(1-\rho)$. Thus the probability that this unit will have to wait a time t to $t + dt$ as a result of waiting for n

previous units to be served is the product of these three probabilities (i.e. the joint probability of all three events occurring). Finally, the total probability of this waiting time arising in any way is the sum of all these products for all values of n from 1 to ∞.

$$w(t)\,dt = \sum_{n=1}^{\infty} \rho^n(1-\rho)(\mu t)^{n-1}\,e^{-\mu t}\mu\,dt/(n-1)!$$

$$= \mu\rho(1-\rho)\,e^{-\mu t}\,dt \sum_{r=0}^{\infty} (\rho\mu t)^r/r! \quad \text{(putting } r \text{ for } n-1)$$

$$= \mu\rho(1-\rho)\,e^{-\mu t(1-\rho)}\,dt$$

or $w(t) = \rho(\mu-\lambda)\,e^{-t(\mu-\lambda)}$ as $\rho = \lambda/\mu$

The average waiting time of a unit is

$$W = \int_0^{\infty} t w(t)\,dt$$

$$= \rho \int_0^{\infty} t(\mu-\lambda)\,e^{-t(\mu-\lambda)}\,dt$$

$$= -\rho[t\,e^{-t(\mu-\lambda)}]_0^{\infty} + \rho \int_0^{\infty} e^{-t(\mu-\lambda)}\,dt$$

$$= \rho[-e^{-t(\mu-\lambda)}/(\mu-\lambda)]_0^{\infty}$$

$$= \frac{\rho}{\mu-\lambda} = \frac{\rho}{\mu(1-\rho)}$$

The probability of a waiting time greater than t is

$$w(>t) = \int_t^{\infty} w(x)\,dx = \int_t^{\infty} \rho(\mu-\lambda)\,e^{-x(\mu-\lambda)}\,dx$$

$$= [-\rho\,e^{-x(\mu-\lambda)}]_t^{\infty}$$

$$= \rho\,e^{-t(\mu-\lambda)}$$

It should be noted that this last result includes the case where $t = 0$.

The total time spent by a unit in the system, including being served, can be obtained in a similar manner. In this case the probability of zero time is zero, and the probability $s(t)\,dt$ of a time from t to $t + dt$ is given in the same way as $w(t)\,dt$ except that there are now n completions in the interval 0 to t instead of $n-1$. Following through the arithmetic it will be seen that:

$$s(t) = (\mu-\lambda)\,e^{-t(\mu-\lambda)}$$

Average time spent $S = 1(\mu - \lambda) = W + 1/\mu$

$$s(>t) = e^{-t(\mu-\lambda)}$$

All the probabilities and averages which have been derived so far refer to the whole of the time considered. Similar results referring only to the time when queue length or waiting time are greater than zero can be obtained by dividing by the probabilities of these conditions (ρ^2 or ρ respectively).

Example

The results which have been derived can be used to examine the example mentioned in the introduction—that of a self-service store with one cashier. In this case, it is not unreasonable to assume random arrival and service times. (The average rate of arrival may depend on time of day, in which case the modifications mentioned in transient solutions would be needed.) Suppose that 9 customers arrive on the average every 5 minutes and the cashier can serve 10 in 5 minutes. That is,

$\lambda = 1{\cdot}8, \quad \mu = 2{\cdot}0, \quad \rho = 0{\cdot}9$ (measuring time in minutes)

The features which may be important to the manager are :

(*a*) The average number of customers waiting for service (which may decide whether a shopper comes into the store) which is

$$Q = \rho^2/(1 - \rho) = 0{\cdot}81/0{\cdot}1 = 8{\cdot}1$$

(*b*) The chance of having more people waiting than there is convenient room for—say, 10 altogether—

$$P(>10) = \rho^{n+1} = 0{\cdot}9^{11} = 0{\cdot}31$$

(*c*) The chance of a customer having to wait long for service, say, more than two minutes

$$w(>2) = \rho\, e^{-t(\mu-\lambda)} = 0{\cdot}9e^{-2\times 0\,2} = 0{\cdot}60$$

If service can be speeded up to 12 in five minutes by using a different cash register, the quantities then become

$$\lambda = 1{\cdot}8, \quad \mu = 2{\cdot}4, \quad \rho = 0{\cdot}75$$

$$Q = 0{\cdot}5625/0{\cdot}25 = 2{\cdot}25$$

$$P(>10) = 0{\cdot}75^{11} = 0{\cdot}042$$

$$w(>2) = 0{\cdot}75\, e^{-1{\cdot}2} = 0{\cdot}23$$

It can be seen that this slight change in speed of service makes a great difference to the operation of the store.

Other Applications of Differential Difference Equations

The queue situations to which this method is applicable can be grouped in four classes according to the type of distributions of arrival intervals and service times.

(*a*) *Exponential.* The simplest case has been dealt with; other cases may have more complex queue discipline or more than one service channel.
(*b*) *Erlang.* Erlang distributions are derived by combining a number of identical exponential distributions. A constant rate can be considered as a special case of an Erlang distribution.
(*c*) *Exponential with varying parameter.* The parameter λ or μ may be a function of time or of the number of units in the system.
(*d*) *Hyper-exponential.* The variance is greater than for an exponential distribution. These distributions appear to have little practical application and are not further considered here.

In this section, special features of the method of solution for types (*a*) to (*c*) are discussed. The main results are included in the list at the end of the chapter.

Exponential Distributions—Single Channel and Limited Queue Length

Suppose that the maximum number of units in the system is limited to L. Any units arriving when there are l already there do not join the queue. The situation is otherwise taken to be identical with that considered in the section on Solution by Differential Difference Equations.

The basic transition equations given in that section still apply when n is less than $L + 1$, but in addition

$$P(L, t + dt) = (1 - \mu\, dt)P(L, t) + \lambda\, dt\, P(L - 1, t)$$

as there is no possibility of a transition to or from the state $L + 1$. As before, the solution for the steady state can be obtained by putting the rates of change equal to zero. In this case, as the number in the system is limited, a steady state

is possible for any value of ρ and not only when the input rate is less than the service rate. However, if L is large and ρ is greater than one, the steady state may be reached very slowly and in practice may not be achieved before the operating conditions change. In using the results it is necessary to check that the steady state is reached.

In this state, as before

$$P(1) = \rho P(0)$$

$$P(n) = (1 + \rho)P(n - 1) - \rho P(n - 2) \qquad (1 < n < L)$$

and from the equation above, $P(L) = \rho P(L - 1, t)$

Hence $P(n) = \rho^n P(0) \qquad (n \leqslant L)$

The total probability of all state $= 1 = \sum_0^L \rho^n P(0)$

$$= P(0)(1 - \rho^{L+1})/(1 - \rho)$$

so that $P(0) = (1 - \rho)/(1 - \rho^{L+1})$

If $\rho = 1$, $\sum_0^L \rho^n = L + 1$ so that $P(0) = 1/(L + 1)$

The various measures of the state of the system can be obtained by calculations similar to those of the section before. An important quantity in this case is the proportion of units lost to the system because they arrive when the system is full; this proportion is $P(L)$. Because some units are lost, the utilisation of the service channel is no longer equal to ρ. It is

$$1 - P(0) = \rho(1 - \rho^L)/(1 - \rho^{L+1})$$

Multiple Channels

It is again supposed that there is an input of identical units with mean rate λ, but that there are c service channels in parallel, each with service rate μ. The units form a single queue and go to any channel which becomes free. The basic transition equations are

$$P(0, t + dt) = P(0, t)(1 - \lambda\, dt) + P(1, t)\mu\, dt$$

$$P(n - 1, t + dt) = P(n - 1, t)[1 - \lambda\, dt - (n - 1)\mu\, dt] + P(n - 2, t)\lambda\, dt + P(n, t)\, n\, \mu dt \qquad (\text{for } 1 < n \leqslant c)$$

$$P(n - 1, t + dt) = P(n - 1, t)(1 - \lambda\, dt - c\mu\, dt) + P(n + 2, t)\lambda\, dt + P(n, t)c\mu\, dt \qquad (\text{for } n > c)$$

as the probability of a service finishing in the interval dt is $r\mu\, dt$ where r is the number of channels occupied. From these equations the steady-state probability $P(n)$ can be derived. For a steady state to be possible λ must be less than $c\mu$.

In this state

$$P(n) = \rho^n P(0)/n! \qquad (\text{for } 0 \leqslant n \leqslant c) \text{ where } \rho = \lambda/\mu$$
$$= c^{c-n}\rho^n P(0)/c! \qquad (\text{for } n \geqslant c)$$

With a multiple channel system there may again be an upper limit L to the number in the system, leading to an additional transition equation,

$$P(L, t + dt) = P(L, t)(1 - c\mu\, dt) + P(L - 1, t)\lambda\, dt$$

the previous equation applying only for $c < n < L$.

Hence $$cP(L) = \rho P(L - 1)$$

so that the expressions given above for $P(n)$ apply up to $n = L$, but with a different value of $P(0)$.

An important particular case is that where no queue is allowed; that is, $L = c$. The steady-state results are then considerably simplified and

$$P(n) = \frac{(c\rho)^n/n!}{\sum_{r=0}^{c} (c\rho)^r r!} \qquad (\text{for } n \leqslant c \text{ and } c = L)$$

As for a single channel, an important quantity is the proportion of units lost to the system, which is $P(L)$, or $P(c)$ when no queue is allowed.

Example

The example of the self-service store can now be completed by considering the effect of an additional cashier. Suppose as before that $\mu = 2{\cdot}0$ for each cashier and $\lambda = 1{\cdot}8$. Then $\rho = 0{\cdot}9$ and $c = 2$.

The results needed are listed later.

$$P(0) = 1 \Big/ \left\{ \sum_{i=0}^{c-1} \rho^i/i! + \rho^c/(c - \rho)(c - 1)! \right\}$$
$$= 1/(1 + 0{\cdot}9 + 0{\cdot}9^2/1{\cdot}1 \times 1)$$
$$= 0{\cdot}38$$

The probability of a customer's having to wait at all is

$$P(>1) = c^{c-n}\rho^{n+1}P(0)/(c-\rho)c! = 2^1 \times 0{\cdot}9^2 \times 0{\cdot}38/1{\cdot}1 \times 2 = 0{\cdot}28$$

The measures used previously can now be calculated.

The average number in the queue is

$$Q = \rho^{c+1}P(0)/(c-\rho)^2(c-1)! = 0{\cdot}9^3 \times 0{\cdot}38/1{\cdot}1^2 \times 1 = 0{\cdot}23$$

The probability of having more than 10 customers in the system is

$$P(>10) = 2^{-8} \times 0{\cdot}9^{11} \times 0{\cdot}38/1{\cdot}1 \times 2 = 0{\cdot}00021$$

The probability of having to wait more than 2 minutes for service is

$$w(>2) = P(>1)\, e^{-\mu t(c-\rho)} = 0{\cdot}28 \times e^{-2\times 2\times 1{\cdot}1} = 0{\cdot}0034$$

The introduction of an extra cashier thus reduces the queueing to a very low level.

It is of interest to notice what would happen if the number of customers now increased so as to bring the utilisation of the cashiers back to the original level. In this case $\lambda = 3{\cdot}6$, $\mu = 2{\cdot}0$, $c = 2$ and the utilisation is again 0·9. It will then be found that

$$Q = 7{\cdot}07, \quad P(>10) = 0{\cdot}33, \quad w(>2) = 0{\cdot}38$$

which may be compared with the values (8·1, 0·31, 0·60) derived before for half the number of customers and one cashier. Although the number of customers is doubled, the numbers waiting are about the same and the probability of a long wait is less. The system is thus not equivalent to the sum of two separate stores each like the original example.

These examples illustrate three features which are common to most queueing situations :

(*a*) Unless the utilisation of the service is fairly high (over $\frac{1}{2}$, say) the congestion is very slight ;

(*b*) When the utilisation approaches unity a small change in it makes a large difference to the congestion ;

(*c*) When service channels can be shared by one queue, the congestion is much less than it would be for the same rates of input and service, but with an independent queue for each service channel.

Service in Series

It is supposed that each unit entering the system requires a number of services, which are carried out in succession.[7] After receiving the first service at the first stage the unit joins a queue for the second service at the second stage, and so on. At each stage there may be one or more service channels. The input for one stage is thus the output from the previous stage. With exponential distributions of input and service intervals, it can be shown that the output is also exponentially distributed so that each stage has an exponential input. In this case, the state of the system is not completely described by the number of units present ; it is necessary to specify the number at each stage (i.e. the number waiting for service or being served at that stage).

All the units pass through each stage so that, for a steady state to be possible, the average rate of input must not exceed the total rate of service at any stage. In the steady state, the average rate of input to each stage is the same and the stages can be considered separately. Suppose the rate of input is λ and at the jth stage there are c_j channels each with average service time $1/\mu_j$. Then, as in the previous section, the steady-state probability of n_j units in the stage is

$$P(n_j) = v(n_j)P(0_j)$$

where

$$v(n_j) = \rho_j^{n_j}/n_j! \qquad (\text{for } n_j \leq c_j)$$

$$= c_j^{c_j - n_j}\rho_j^{n_j}/c_j! \qquad (\text{for } n_j \geq c_j)$$

and

$$\rho_j = \lambda/\mu_j$$

The probability of the state of the whole system of k stages with n_j at the jth stage is the product of the probabilities for each stage and can be written

$$(n_1, n_2, \ldots n_j \ldots n_k) = (0) \prod_{j=1}^{k} v(n_j)$$

As before, $P(0)$ can be found by equating to unity the sum of the probabilities of all states and it is

$$P(0) = \prod_{j=1}^{k} \left\{ 1 \Big/ \sum_{n_j=0}^{\infty} v(n_j) \right\}$$

Erlang Distributions

The examples so far have been limited to those where the arrival interval and service time both have exponential distributions. With these distributions the probability of an arrival at any instant is quite independent of previous arrivals, and the probability of service being completed at any instant is quite independent of the time when service began. It is this property which allows simple transition equations to be written.

In practice, the unrelated arrivals of customers might well be exponential, but the distribution of service times may be quite unlike the exponential. Transition equations cannot be written using such distributions because the probability of service being completed depends on when service began.

There are a number of ways of dealing with this. One way is to extend the use of the simple method by building up service time distributions from combinations of exponential distributions. This method is considered here.

Service is considered to be in a number of exponential phases immediately following each other ; all phases have to be completed before another unit can enter the service channel. The distribution of the whole service time is quite different from that of its components. When all k phases are identical, the service time has a chi-square distribution with $2k$ degrees of freedom. Because of this physical interpretation of chi-square, formulated by A. K. Erlang for just the reasons given above, the distributions are widely known as Erlang distributions. The family of Erlang distributions for $k = 1, 2, 3, \ldots$ is scaled so that the mean of the whole service time is the same for each. When $k \to \infty$, the service time is constant. Erlang also considered the cases where the phases are different from each other and obtained much more general distributions.

The transition equations for systems with such distributions of service times can now be written, because the probability of passing from the jth to the $(j+1)$th phase of service is $\mu_j\, dt$, and depends only on the presence of a unit in the jth stage.

For random arrivals and Erlang service time, a single-channel system is characterised by $P(n, j)$ which is the probability that there are n units in the system and the unit being

served is in the jth phase of service. The equations for the steady state, with an arrival λ and a service rate $k\mu$ for each phase, and therefore μ for the whole system, are

$$(\lambda + k\mu)p(0, 1) = \lambda p(0, 0)$$

$$(\lambda + k\mu)p(0, j) = k\mu p(0, j - 1) \qquad (\text{for } 1 < j \leqslant k)$$

$$(\lambda + k\mu)p(n, 1) = \lambda p(n - 1, 1) + k\mu p(n + 1, k) \quad (\text{for } n \geqslant 1)$$

$$(\lambda + k\mu)p(n, j) = \lambda p(n - 1, j) + k\mu p(n, j - 1) \qquad (\text{for } n \geqslant 1,\ 1 < j \leqslant k)$$

In general, the solution of these equations requires the introduction of probability generating functions and numerical computation. For $k = 2$ it can be shown that

$$p(n) = \sum_{i=0}^{n} \frac{(n + 1 + i)!(1 - \rho)}{(2i + 1)!(n - i)!} \frac{(\rho)^{n+i}}{2} \qquad \text{where } \rho = \frac{\lambda}{\mu}$$

but for larger values of k the probabilities have to be found numerically. Average results can be expressed generally and the main ones are given in the list in Appendix 1.

It is seen that the average queue length and average waiting time for negative exponential service ($k = 1$) are twice the queue length and waiting time for constant service ($k = \infty$).

A similar approach can be used to analyse systems with an Erlang arrival time distribution. Units are imagined to pass through k phases of an arrival channel before joining the queue. As soon as one unit is through the arrival channel, the next one is introduced.

The use of Erlang distributions does not necessarily imply that the phases of service and arrival have any physical reality, though that may be so.

Methods of solution of queueing problems with Erlang distributions are described in some detail by Morse.[8] He is working out tabulations of some of the more complicated examples, which are to be published in a later book.

Exponential Distributions with Varying Parameters

The queueing situations to be discussed in this section have input and service time distributions of the same form as those given above, but the parameters λ and μ are not constant. Differential difference equations can be derived in exactly the

same way as for exponential distributions, but λ and μ are replaced by the appropriate functions.

The parameters may depend on the number of units in the system. The values of the parameters when there are n in the system can be written $\lambda(n)$ and $\mu(n)$. The steady-state probabilities can be derived in the same way as before. It will be found that for a single channel

$$P(n) = P(0) \prod_{i=1}^{n} \frac{\lambda(i-1)}{\mu(i)}$$

and for multiple channels corresponding results apply. Some special cases have been examined. For example, if $\mu(n) = n\mu$ and $\lambda(n) = \lambda$ where μ and λ are constant, a fairly simple expression can be derived for the time-dependent probability,

$$P(n, t) = e^{-R}R^n/n! \quad \text{where } R = (1 - e^{-\mu t})\lambda/\mu$$

and $$P(0, 0) = 1$$

The average number in the system at time t can be derived directly from the differential difference equations and is equal to R. It may be noted that, except for the waiting time, this case is equivalent to a system with an infinite number of exponential service channels.

The other type of variation which has been studied is dependence on time; there can then be no steady state in the sense used previously. It is necessary to solve the equations to give the time-dependent probabilities of the state n (which depend also on the initial state of the system). This has been done for a single service channel and distribution parameters which are general functions of time; the results are rather complex.[9]

Other Mathematical Methods of Solution

The preceding sections have used the classical approach to the theory of queues, by the formation of differential difference equations for the distribution of queue size. In his book on queueing theory, Morse uses only this approach.[8]

Other analytical techniques have been used to give both general and particular results. Some of the more important of these techniques are briefly described in this section.

Integral Equations

One approach leads to integral equations for the distributions of the waiting time. It is valid for general distributions of arrival and service time, but it is not easily applicable to transient states of the system. It was first used by Lindley[10] for a single service channel. The analysis for such a queue is given here.

If w_n is the time the nth unit has to wait for service if it waits at all and is minus the time the server has been idle if the nth unit does not have to wait, a_n is the interval between the arrival of the nth and $(n+1)$th unit, with distribution $a_n(t)\,dt$, and b_n is the service time of the nth unit, with distribution $b_n(t)\,dt$,

then
$$\begin{aligned} w_{n+1} + a_n &= w_n + b_n && (\text{for } w_n \geqslant 0) \\ &= b_n && (\text{for } w_n < 0) \end{aligned}$$

and on writing
$$\begin{aligned} u_n &= b_n - a_n \\ w_{n+1} &= w_n + u_n && (\text{for } w_n \geqslant 0) \\ &= u_n && (\text{for } w_n < 0) \end{aligned}$$

The distribution of w_{n+1} can be simply written down in terms of the distribution of w_n and u_n. If these are $w_n(t)\,dt$ and $u_n(t)\,dt$ respectively, then

$$w_{n+1}(t) = \int_0^\infty w_n(x)u_n(t-x)\,dx + u_n(t)\int_0^\infty w_n(x)\,dx$$

In writing this equation, it is assumed that the time the nth unit waits is independent of the subsequent service time and of the time when the next unit arrives.

The service time and arrival interval distributions are known, so the distribution of u_n is known,

$$u_n(t) = \int_0^\infty a_n(x)b_n(t+x)\,dx.$$

In the steady state the distributions of arrival, waiting and service times are the same for each unit, and it is possible to write

$$w(t) = \int_0^\infty w(x)u(t-x)\,dx + u(t)\int_0^\infty w(x)\,dx$$

where
$$u(t) = \int_0^\infty a(x)b(t+x)\,dx$$

The integral equation for the waiting time distributions is soluble in terms of the distributions of u, which in turn is completely given by the known arrival time and service time distributions. Although the equation is soluble analytically for simple forms of $u(t)$, numerical computation would usually be required. The equation is of the Wiener–Hopf type, and also arises in the study of random walks and sequential tests of hypotheses.

With this technique, Lindley studied queues with the following distributions :

(*a*) regular input and Erlang service,
(*b*) Poisson input and general service,
(*c*) Erlang service and general service.

Markov Chains

A sequence of events forms a Markov chain if the probability of each event can be expressed in terms of the previous event, irrespective of any information about earlier events. Hence a knowledge of any event makes all information about previous events irrelevant to the prediction of subsequent events.

In a similar way, a continuous process is Markovian if a knowledge of any instantaneous state of the process makes all previous history irrelevant in predicting its future behaviour.

The instantaneous state of a queueing system may be described by the number of units in it. If the arrival and service distributions are exponential, this number of units is all that is needed to determine the probabilities of future instantaneous states, and the process is Markovian. For other distributions these probabilities depend also on the previous history, and the process is non-Markovian. However, the process can still be regarded as Markovian if the times since the previous arrival and the last completion of service are included in the description of the instantaneous state of the system.

If the queueing system is described in such a way, it can be studied as a discrete-time system, using the well-developed theory of Markov chains.[11] The approach used in this chapter also makes use of Markov chains, for if the size of a queue when a unit leaves is known, then previous information on

queue size is of no predictive value. This is true about all instants when a unit leaves. This set of particular instants can then be thought of as a Markov chain " embedded " in the queueing process.[12]

The analysis of embedded Markov chains has proved a fruitful approach to the analysis of queueing systems. It was used, for instance, by Bailey in the analysis of queueing for bulk service. (See later section on bulk service.)

An interesting use of the Markovian properties of a queueing process is made in a recent paper by Winsten.[13] He considers the equilibrium condition of a queue with regular input and a single exponential service channel. The numbers in the system immediately before each arrival form a Markov chain. The probability of a transition from N_a before one arrival to N_{a+1} before the next (where both are greater than zero) depends only on $N_{a+1} - N_a$. A sequence of transitions, such that the number does not drop below its initial value, has a probability which is the product of the individual transition probabilities. This probability is the same whatever the number in the system initially, provided the sequence of changes at each transition is the same. From this property it can be shown that $p(n + 1) = kp(n)$ where k is a constant ; that is, that the number in the system has a geometric distribution.

The method can be extended to consider arrivals with a probability distribution about the regular time, but if the distributions of successive arrivals overlap, the calculation is complicated. The usual method of specifying the input to a queueing situation is by the probability distribution of the arrival intervals ; unless the distribution is exponential, this implies that the time of an arrival depends on the previous arrivals. In many problems this is not so and a distribution about a regular arrival time is a more appropriate description. Winsten's method of analysis may therefore prove to be a useful general approach.

" Ad Hoc " Methods

Some quite general results can be obtained without carrying out the analysis necessary to give a complete description of the system. A good example of this type of *ad hoc* approach is D. G. Kendall's derivation of the average queue length and

waiting time for a single queue with random arrivals and general service time distribution.[14]

Let λ be the arrival rate
μ be the service rate
q be the size of queue a departing unit leaves behind it
q' be the size of queue the next unit leaves behind it
v be the service time of this second unit
r be the number of units arriving during the service time v (r is conditionally a Poisson variable of mean value λv)

Then $q' = q - 1 + \delta + r$ where $\delta = 1$ when $q = 0$
$\delta = 0$ when $q \neq 0$

For the steady state, it is possible to equate the expected values of both sides of the equation and the expected values of the squares of both sides. As $\delta^2 = \delta$ and in the steady state,

$$E[q'] = E[q] \quad \text{and} \quad E[q'^2] = E[q^2]$$

$$E[\delta] = 1 - E[r] = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

and $$E[2q(1 - r)] = E[(r - 1)^2] + E[\delta(2r - 1)]$$

Noting that r is independent of q and δ it follows that

$$E[q] = E[r] + \frac{E[r(r - 1)]}{2(1 - E[r])}$$

and hence $$E[q] = \rho + \frac{\rho^2[1 + \mu^2 \operatorname{var}(v)]}{2(1 - \rho)}$$

Special Situations

The theory of queues has been extended to take in more and more features which appear in real-life queueing situations, whether the queues be people, manufactured parts, repair jobs or perishable goods. Some of these situations are dealt with in this section.

Limited Input

It may be that units arrive from a limited pool of potential customers. Once a unit joins the queue, there is one less unit which could arrive and therefore the probability of an arrival is lowered. When a unit is served it rejoins the pool of

potential customers, and the probability of an arrival is thereby increased.

The theory of queues with limited input was first formulated to give a model of machine interference. In some industries, banks of machines are supervised by one operator. When a machine stops it has to wait for service if the operator is already busy. If an operator looks after too few machines the labour cost is high, and if the operator looks after too many machines the cost of lost output becomes appreciable. These costs have to be balanced so that the total cost per unit of production is minimised.

Systems with limited input have been analysed for random arrivals and negative exponential service. In the general case there are m units in the system, and c servers each with mean service rate μ. Each unit in the pool of potential customers has a probability $\lambda\, dt$ of joining the queue in the short interval dt. This means that the time spent in the pool between completing service and joining the queue again is exponentially distributed. Writing $p(r)$ for the probability that there are r customers waiting or being served, and noting that the probability of an arrival is then $(m - r)\lambda\, dt$, the steady-state equations are then

$$
\begin{aligned}
m\lambda p(0) &= \mu p(1) \\
[(m-r)\lambda + r\mu]p(r) &= (m-r+1)\lambda p(r-1) + (r+1)\mu p(r+1) \\
&\qquad\qquad r = 1,\ 2,\ \ldots\ c-1 \\
[(m-r)\lambda + c\mu]p(r) &= (m-r+1)\lambda p(r-1) + c\mu p(r+1) \\
&\qquad\qquad r = c,\ c+1,\ \ldots\ m-1 \\
c\mu p(m) &= \lambda p(m-1)
\end{aligned}
$$

from which it follows that

$$
\begin{aligned}
(r+1)\mu p(r+1) &= (m-r)\lambda p(r) \qquad r < c \\
c\mu p(r+1) &= (m-r)\lambda p(r) \qquad c \leqslant r \leqslant m
\end{aligned}
$$

The solution is

$$
\begin{aligned}
p(r) &= \frac{m!}{(m-r)!\, r!}\left(\frac{\lambda}{\mu}\right)^r p(0) \qquad \text{for } r \leqslant c \\
&= \frac{m!}{(m-r)!\, c!\, c^{r-c}}\left(\frac{\lambda}{\mu}\right)^r p(0) \qquad \text{for } c \leqslant r \leqslant m
\end{aligned}
$$

$p(0)$ being found by the condition that $\sum p(r) = 1$. As the

maximum queue length is finite there is no limit to the ratio λ/μ.

As in other systems, it is better for two or more servers to work on the pool together as a team, rather than to split the pool into parts and to work independently.

The various measures of effectiveness of the system can be

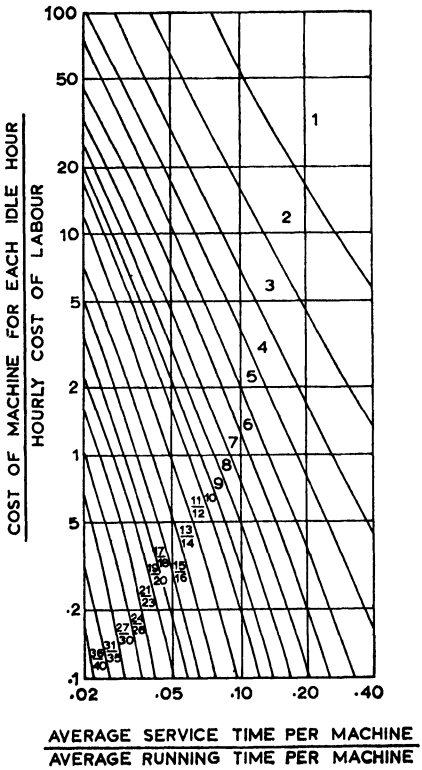


FIG. 3.2

derived from the formula as before, though they are not of simple form. Naor[15] has given the solution in terms of tabulated Poisson functions and their cumulatives, and has discussed the solution in these terms. Cox[16] has given a table of the utilisation of the service for a single service channel when the average time in the pool is at least eight times as long as the service time. A modified form of this table is given in table 3.3,

TABLE 3.3. *Machine interference : utilisation of a single service channel*

$\frac{m\lambda}{\mu}$ \ $\frac{\lambda}{\mu}$	·005	·01	·02	·03	·04	·05	·06	·07	·08	·09	·10	·11	·12
0·1	0·099	0·099	0·098	0·097	0·096	0·095	0·094	0·093	0·092	0·092	0·091	0·090	0·089
0·2	·199	·198	·195	·193	·191	·189	·187	·185	·184	·182	·180	·179	·177
0·3	·298	·296	·292	·288	·285	·282	·279	·276	·273	·270	·268	·266	·263
0·4	·397	·394	·388	·382	·378	·373	·369	·365	·361	·357	·353	·350	·347
0·5	·495	·491	·482	·475	·468	·462	·456	·451	·446	·441	·436	·432	·427
0·6	·593	·586	·575	·565	·556	·548	·541	·534	·527	·521	·516	·510	·505
0·7	·689	·680	·665	·652	·641	·631	·621	·613	·605	·598	·591	·584	·578
0·8	·783	·770	·750	·734	·720	·708	·697	·687	·678	·670	·662	·654	·647
0·9	·872	·854	·828	·809	·793	·779	·766	·755	·745	·736	·727	·719	·711
1·0	·946	·924	·895	·874	·856	·841	·828	·816	·805	·795	·785	·777	·768
1·1	·989	·972	·946	·926	·908	·893	·880	·868	·857	·846	·837	·828	·819
1·2	·999	·994	·978	·962	·947	·934	·922	·910	·900	·890	·880	·872	·863
1·3	.	.	·994	·984	·973	·963	·952	·943	·933	·924	·916	·907	·900
1·4	.	.	.	·994	·988	·981	·974	·966	·958	·951	·943	·936	·929
1·5	.	.	.	.	·996	·992	·987	·981	·975	·970	·964	·958	·952
1·6	.	.	.	.	.	.	·994	·990	·986	·982	·978	·973	·968
1·7	.	.	.	.	.	.	.	.	·993	·990	·987	·984	·980
1·8	.	.	.	.	.	.	.	.	.	.	·993	·990	·988
1·9	.	.	.	.	.	.	.	.	.	.	.	.	·993

m: the number of machines λ: the mean rate of breakdown per unit running time μ: the mean service rate

The machine utilisation is given by (utilisation of service channel) $\Big/\left(\frac{m\lambda}{\mu}\right)$. The appropriate values can be found from the table by interpolation. For example, with $\lambda = \cdot 9$, $\mu = 10$, $m = 5$, service utilisation is 40%, and machine utilisation is 89%

In an earlier paper, Ashcroft[17] gave tables of results for machine interference with constant service time.

In his original paper, Palm[18] gave a diagram showing the optimum number of machines to use in terms of the hourly cost of idle time, the hourly cost of labour, the arrival rate and the service rate for groups of machines tended by one operator. A version of this is given in Fig. 3.2 for exponential service and running time.

The assumptions of the theory are an idealisation. With more machines to care for, travelling time between them increases, and so service takes longer. This becomes even more marked if the operators work as a team. It might well be that when much ground is to be covered, the increase in service time could outweigh the theoretical advantage of working as a team. This ought to be checked before limited input—machine interference—formulae are applied, quite apart from the need to check that it is reasonable to use exponential service and running times.

The theory of queues with limited input is applicable to many situations, for example :

breakdown of coalface machinery
maintenance of a pool of equipment
allotment of maintenance workers
shorthand writers attending a limited number of dictators.

Bulk Service

Service of batches of units may take a number of different forms :

the server waits for s units, and then serves them together, e.g. passengers on a coach tour.
when ready, the server takes the units waiting, though not more than s, and serves them together ; e.g. passengers on public transport.

In some instances the time spent waiting may be all that matters, the service time being irrelevant. The interval between successive services could then be regarded as being completely taken up by serving the previous batch, or to be the time when the service channel is ready for a new batch, the actual service time being shorter.

One case has been analysed by Bailey.[19] Units arrive at random and are served in batches of not more than s, the server taking zero units if none is waiting. The distribution of intervals between service is chi-squared with $2p$ degrees of freedom and mean p/a. The analysis uses embedded Markov chains and probability generating functions. There is no simple analytic solution, but a useful inequality for the average waiting time W is

$$0 \leqslant W - \frac{p+s}{2a(s-m)} \leqslant \frac{p(s-1)}{2am}$$

where m is the average number of arrivals per service interval. For example, if the arrival rate is nearly equal to the service rate, say $s = m + 1$, then the average waiting times, measured in service intervals, is given by

$\frac{m+1}{2} + \frac{1}{2} \leqslant W \leqslant \frac{m+1}{2} + 1$ for exponential service ($p = 1$)

and $\frac{1}{2} \leqslant W \leqslant 1$ for regular service ($p = \infty$).

In general,

the greater the traffic intensity, m/s, the greater is the advantage of regular service over irregular service,
the larger the batch size for a given traffic intensity, the greater is the advantage of regular service over irregular service.

The theory was developed to give a model of customers at a clinic, where requests for attention are effectively random and clinics are held each week or at irregular intervals. The theory of bulk service might also be applied to

rubbish removal
bus and train services
bulk shipment of goods.
train despatch from marshalling yards
lifts.

Transient Solutions

Most of the previous analysis has been concerned with the steady state of queueing systems. For systems where this condition can occur, the steady-state probability of a given

state can be defined as the probability of finding the system in that state at a time sufficiently far in the future ; that is, the steady-state probability of the state n, $p(n)$ is the limit of the time-dependent probability $P(n, t)$ as $t \to \infty$. If such a system is known to be in the state k at a given instant, then immediately afterwards states near k are more likely than would otherwise be expected. The state probabilities tend gradually to the steady-state probabilities. These changes in the probabilities are of interest in some problems, particularly after the occurrence of no units or a large number of units in the system ; that is, after idle or busy periods.

Transient probabilities are also of interest for systems which operate only for limited periods. In this case, there is no need for a steady state to be possible. Indeed, Bailey[20] gives an instance where it is quite suitable for an appointment system to have $\rho = 1$, because the period for which it operates is short enough for the waiting time of customers to remain small.

The transient solutions can be obtained by solving the time-dependent differential equations of the system. Morse[8] gives one way of doing this and Saaty[9] quotes transient solutions for a few types of systems.

For systems having input or service rates which vary with time, all solutions can be regarded as transient. Random distributions with varying parameters have been considered earlier. Another more common problem arises from the rush-hour type of situation ; input, and perhaps also service, rate increases for a limited period. This problem can be tackled by splitting the total period into intervals during which the rates can be taken as constant, and solving for each interval in turn. Cox[3] has used this method for a simplified analysis of a rush-hour. Usually, however, simulation methods are employed for such problems.

Queue Discipline

So far, in the queueing models discussed, units have been served in the order in which they arrive. The queue discipline is thus first-come first-served. This is the discipline which normally comes to mind when queues are thought of, but it is not the only one.

Units may be selected for service at random from those

waiting. The average waiting time of units is the same as for first-come first-served queues. But, because a unit may now be served " out of turn " and have to wait an " excessively long time ", there is a greater proportion of very long and very short waiting times, and the variance of the waiting time is greater than for first-come, first-served queues. An example of random selection for service is the selection of trunk callers by the operator at an exchange.

A further queue discipline is last-come first-served. The average waiting time is the same as for queues which are first-come first-served. But the variance of the waiting time is even greater than for random selection, and there is a correspondingly greater proportion of very long and very short waiting times. An example of last-come first-served discipline is the handling of goods which have to be stacked ; foodstuffs in a warehouse or letters in an in-tray.

In short, for these three types of queue discipline, the average waiting time is the same, but the variance of the waiting time, and the proportion of units waiting a long time, increase in the order first-come first-served, random, last-come first-served.

Queue discipline might be extremely important in the handling of goods which deteriorate, the discipline which may be convenient to the stock-room being the one causing the greatest loss.

A practical variant is that there may be an advantage in giving priority to those with short service time regardless of arrival sequence. This principle is sometimes applied in shops by giving up one service point to " quickies ".

Priority

The units waiting for service may belong to two or more classes having an order of priority. The priority will be either

(*a*) priority of entry when the service is free, or
(*b*) pre-emptive priority ; priority to displace units being served if they are of lower priority.

The average waiting time of all the units remains unaltered if some of them are given priority (unless the priority is related to the service time), but the average waiting time of the various classes is decreased or increased according to their order of

priority. In the case of pre-emptive priorities, the behaviour of the class with first priority is as though there was no other type of unit in the system.

It may be that the high-priority classes have a shorter average service time than those of lower priority. In this case the average number waiting for service and the average waiting times are reduced. Similarly, if priority is given to classes having longer average service time, the average number waiting for service and the average waiting time are increased. Having a shorter or longer service time may be an incidental characteristic of the classes selected for priority, or it may be the reason why they are given priority.

Saaty[9] gives references to three systems in which priorities have been studied :

Poisson arrival, r classes of priority, non pre-emptive service, general service times,[21]

Poisson arrival, 2 classes of priority, pre-emptive service, exponential service times,[22]

Poisson arrivals, a continuous number of priorities according to shortness of service, general service time.[23]

Impatient Customers

Another feature of real-life queues is the impatient customer who does not want to queue for too long. Two types of impatient customers have been studied :

the one whose decision to join is influenced by queue length,

the one who decides to leave the queue if the wait is too long.

When the decision to join the queue depends on the length of the queue, the arrival probability in the queue is effectively queue dependent—the situation described earlier. In the case of exponential arrival and service times when the probability of an arriving unit joining the queue is proportional to e^n, the average queue length remains finite, however large the arrival rate is compared with the service rate.[8]

Another case studied is that of customers leaving after waiting for a time T in an exponential arrival and service system.[24]

But here, as in the previous section, although some particular cases have been studied, there does not as yet appear to have been a systematic study of the loss of customers who are impatient.

The applications of such a study are not only to human queueing systems where impatience is understood, but also to such queueing systems as

storing perishable goods
processing heated metal
communication of information,

in each of which the unit loses its value if it waits too long.

Cyclic Queues

In closed systems, units may pass from one service facility to the next in continual circulation, either being served or waiting for service. At each stage there may be a single server or a number of servers in parallel.

Koenigsberg[2] has analysed a cyclic system with m units in the system and k single-channel exponential service facilities of mean service rate $\mu_1, \mu_2, \ldots \mu_k$. There are no arrival rates to consider, as the output of one stage is the input of the next stage. The equations governing the transition probabilities can be formulated as in previous examples. The probability of there being n units waiting or being served at the first stage, n_2 at the second stage, and so on, is given by

$$p(n_1, n_2, \ldots n_k) = \frac{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}}{Z_k^m}$$

where $x_i = \dfrac{\mu_1}{\mu_i}$

and $$Z_k^m = \sum x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k} = \sum_{i=1}^{k} \frac{x_i^{m+k-1}}{\prod_{\substack{i=1 \\ j \neq 1}}^{k} (x_i - x_j)}$$

The sum being over all partitions of m

If further ${}_iZ_{k-1}^m$ is the same sum with $n_i = 0$

then the idle time of the ith stage, D_i, can be written as

$$D_i = \frac{{}_iZ_{k-1}^m}{Z_k^m}$$

The throughput of the ith stage is $(1 - D_i)\mu_i$. As the throughput of all stages must be equal, it follows that this expression has the same value for all values of i. It is of interest, too, that the throughput is independent of the order of the servers in this case.

The simplest particular case is that when all k servers have identical service times, μ. In this case

$$Z_k^m = \frac{(m + k - 1)!}{m!(k - 1)!}$$

$$_iZ_{k-1}^m = \frac{(m + k - 2)!}{m!(k - 2)!}$$

$$D_i = \frac{k - 1}{m + k - 1}$$

The rate of throughput is therefore $\frac{m}{m + k - 1}$. More difficult cases have been tackled.

The theory of cyclic queues was developed as a model of the cycle of operations in a group of nine coal faces in room and pillar working. The faces were considered to queue for service by blasting, filling, timbering and drilling teams. The effect of doubling one service stage was also considered.

An obvious application of the theory would be the movement of trucks on a closed loop. For most circulating systems it will probably be necessary to consider some more general distributions of service times.

Concluding Remarks

The special systems so far studied in the literature of the subject (summary in[25]) have considerable variety and provide powerful means of dealing with processes and systems which appear, at first sight, to be intractable because of the random element they contain. Compared with reality, it must be admitted that, like most tractable models of real situations, these are very simple and, in consequence, cannot be expected to provide exact analogies of real-life situations. This is by no means uncommon in Operational Research ; nevertheless, general conclusions can be drawn from the models, trends

perceived, and comparisons made which are true for the real situations. Where planning has taken place without regard to those theoretical results—as must often be the case at present—there is much to be gained even from the qualitative results yielded by these simplified models.

An example of a useful, general result is the very rapid increase in the queue size and corresponding waiting times for systems with traffic intensities approaching unity—though still less than unity. For a single server and simple exponential distribution of service time, the average queue size for a traffic intensity of 0·95 is double that for an intensity of 0·90 which in turn is more than double that for intensity 0·80. The waiting times correspond. Such a general conclusion is sufficient guide for planning the number of service points in cases where it is important to keep the queue size low—as, for example, for aircraft arriving at an airport, urgent messages at police centres, or lorries arriving for loading in a congested street.

Although the actual queue sizes and their ratios under differing traffic will be different for different systems—bulk service, multiple servers and so on—it is generally true that at high traffic intensities near to unity any system is on the verge of the non-stationary state associated with traffic intensities of unity or more discussed earlier.

Another useful generality concerns the effect of making service times less variable than if their beginning and end are at random. Thus, at a traffic intensity of 0·8 it can be shown that an increase in service time of up to 10% is to be preferred as regards waiting time of customer to allowing the service time to become variable whilst retaining the same average. The effect is, of course, less at the high traffic intensities when the server is fully occupied anyway and the rate at which demands are dealt with depends only on the average service time. The figure of 10% quoted is for the comparison of constant with exponentially distributed service time, the usual form of distribution taken by such times.

Thus it can be said, in general, about almost any congestion problem that steps taken to make the service more constant will reduce the congestion even though the demand remains at random.

The Weaknesses of the Methods

Since the problems arise from a random event, the demand, a complete and certain answer should not be expected. The statistical situation in which it is not known for certain what will happen next, but only the pattern of happenings over a period, leads to corresponding statistical answers in the form of a pattern, or distribution, over a period. This situation is more common than is generally supposed and can be dealt with by common statistical methods. The resulting information can yield—besides average values of the quantities of interest like waiting time and size of queue—the values which are seldom exceeded or the frequencies with which stated values are exceeded. " Seldom " can be taken to mean any particular low frequency desired and is a measure of the risk of these values being exceeded.

This uncertainty in the form of answer is not so much a weakness of the method as of the original information available. The fact of random demand introduces the uncertainty. It is mentioned here to prepare the user of O.R. for the form the answer must take.

A major discrepancy between the simple models and reality is the lack of constancy in rate of demand. In many real situations it is not to be expected that the rate of demand will be steady ; instead it will fluctuate cyclically or randomly. Some attempts have been made to deal with models having these rates as simple functions of time,[3] but such models—while needed—are obviously more difficult to handle ; the results are no longer continuing results which can be expressed as relative frequencies or relative lengths of time during which the system is in given states. The results are instead tied to a particular zero of time and the statistical statements of results are relative frequencies among the occurrence of such zeros.

Some situations of changing rate with time may be dealt with by dividing time into periods of approximately constant rate. Where the periods have to be short to achieve this constancy, stationary states are not likely to be achieved and the usual results are not applicable.

Future Development

The development in the immediate future in Operational Research on congestion problems is clearly in the use of high-speed computers to simulate, not real situations necessarily, but complex models close enough to real situations to yield useful answers of the kind already available from the simpler models for simple situations.

The value of using the computer lies firstly in the possibility it gives of using peculiar distributions of the various intervals involved and of various combinations of such distributions when a multi-phase system is studied : secondly in the possibility of generating directly the distributions, instead of the average values, of the quantities in which the Operational Research is interested : and, finally, in enabling fairly rapid exploration of a field of parameter values in a search for a near optimum.

The computer, too, would lend power to investigations of changing rate systems with traffic intensities greater than unity. Various techniques of generating distributions on machines and of storing intermediate information are already in the literature and a few Operational Research projects of this nature are already in progress.

Appendix 1

STEADY STATE FORMULAE FOR QUEUES

Most of the quantities of interest in queueing problems in the steady state are defined, and formulae for them are given, for the simple case ; a selection of the more important results for other cases is given. For brevity, the word probability has been abbreviated to *P*. In all the formulae $\rho = \lambda/\mu$.

Contents List

Unless otherwise specified, the following conditions are satisfied :

(*a*) Unlimited input with Poisson distribution,
(*b*) Unlimited number in the system,

(*c*) Single queue serves one unit at a time,
(*d*) Each unit requires service by only one channel (i.e. single-stage service).

(*i*) *With service in order of arrival*

Type		
1	Single channel	Exponential service
2	Single channel, limited number in system	
3	Multiple channels	
4	Multiple channels, limited number in system	
5	Multiple channels, multiple stages	
6	Single channel, service rate proportional to number in system	
7	Single channel, Erlang service	
8	Single channel, general service	

(*ii*) *With priorities*

9	Single channel, general service	Priorities in order of service time
10	Single channel, exponential service	
11	Single channel, exponential service, two classes, one with pre-emptive priority	

1. Poisson input, unlimited queue length, single service channel with exponential service time distribution. Average rate of input λ and average rate of service μ.

Traffic intensity	$\rho = \lambda/\mu$
P of n units in the system	$P(n) = \rho^n(1 - \rho)$
P of more than n units in the system	$p(>n) = \rho^{n+1}$
Average number of units in the system	$\mathcal{N} = \rho/(1 - \rho)$
P of n units in the queue	$q(n) = \rho^{n+1}(1 - \rho)$ (for $n > 0$)
	$q(0) = 1 - \rho^2$
P of more than n units in the queue	$q(>n) = \rho^{n+2}$
Average number of units in the queue	$\mathcal{Q} = \rho^2/(1 - \rho)$

P of no waiting (up to start of service)	$w(0) = 1 - \rho$
P of waiting time t to $t + dt$	$w(t)\,dt = \rho(\mu - \lambda)e^{-t(\mu-\lambda)}\,dt$
P of waiting longer than t	$w(>t) = \rho e^{-t(\mu-\lambda)}$
Average waiting time	$W = \rho/(\mu - \lambda)$
P of time from t to $t + dt$ spent in the system	$s(t)\,dt = (\mu - \lambda)e^{-t(\mu-\lambda)}$
P of more than time t spent in the system	$s(>t) = e^{-t(\mu-\lambda)}dt$
Average time spent in the system	$S = 1/(\mu - \lambda)$
P that a queue exists	$p(>1) = \rho^2$
P that a unit needs to wait for service	$(p>0) = \rho$
Utilisation of the service channel	$u = \rho$

The formulae for $q(n)$, $q(>n)$ and Q, if divided by ρ^2 apply to the time during which a queue exists; those for $w(t)$, $w(>t)$ and W if divided by ρ refer to the units which need to wait.

2. Number in the system limited to L but otherwise as type 1 with no restriction on λ and μ.

$$p(0) = (1 - \rho)/(1 - \rho^{L+1})$$

$$p(n) = \rho^n p(0) \quad (\text{for } n \leqslant L) \qquad \mathcal{N} = \frac{\rho - (L+1)\rho^{L+1} + L\rho^{L+2}}{(1-\rho)(1-\rho^{L+1})}$$

$$u = 1 - p(0) \qquad Q = \mathcal{N} + p(0) - 1$$

3. Poisson input, unlimited single queue, c service channels in parallel, each with exponential service time distribution. Average rate of input λ and average rate of service by each channel μ. $\lambda < c\mu$.

$$u = \rho/c = \text{traffic intensity per channel}$$

$$p(0) = 1 \Big/ \left\{ \sum_{i=0}^{c-1} \rho^i/i! + \rho^c/(c - \rho)(c-1)! \right\}$$

$$p(n) = \rho^n p(0)/n! \qquad (\text{for } n \leqslant c)$$

$$= c^{c-n}\rho^n p(0)/c! \qquad (\text{for } n \geqslant c)$$

$$p(>n) = c^{c-n}\rho^{n+1}p(0)/(c - \rho)c! \qquad (\text{for } n \geqslant c - 1)$$

$$Q = \rho^{c+1}p(0)/(c - \rho)^2(c-1)!$$

$$\mathcal{N} = Q + \rho$$

$$w(>t) = p(>c-1)\, e^{-\mu(c-\rho)}$$
$$W = Q/\lambda$$
$$S = W + 1/\mu = N/\lambda$$

4. Number in the system limited to L, no restriction on λ or μ, and otherwise as type 3.

$$p(0) = 1 \Big/ \left\{ \sum_{i=0}^{c-1} \rho^i/i! + \rho^c(1 - (\rho/c)^{L-c+1})(c-\rho)(c-1)! \right\}$$
$$p(n) = \rho^n p(0)/n! \qquad (\text{for } n \leqslant c)$$
$$= c^{c-n}\rho^n p(0)/c! \qquad (\text{for } L \geqslant n \geqslant c)$$

In the particular case where $c = L$ (no queue allowed)

$$p(n) = \rho^n/n! \Big/ \sum_{i=0}^{c} \rho^i/i! \qquad (\text{for } n \leqslant c \text{ and } c = L)$$

5. Poisson input of average rate λ, k stages of service in series with c_j channels at the jth stage each with exponential service time distribution of mean rate μ_j. $\lambda < c_j\mu_j$ for all j. $\rho_j = \lambda/\mu_j$.

P of $n_1, n_2, \ldots n_j \ldots$ units at stages $1, 2, \ldots j \ldots$ is

$$p(n_1 n_2 \ldots n_j \ldots n_k) = p(0) \prod_{j=1}^{k} v(n_j)$$

where
$$v(n_j) = \rho_j^{n_j}/n_j! \qquad (\text{for } n_j < c_j)$$
$$= c_j^{c_j - n_j}\rho_j^{n_j}/c_j! \qquad (\text{for } n_j \geqslant c_j)$$

and
$$p(0) = \prod_{j=1}^{k} \left\{ 1 \Big/ \sum_{n_j=0}^{\infty} v(n_j) \right\}$$

when $c_j = 1$ for all j (one channel at each stage) the results for type 1 apply to each stage (putting μ_j for μ).

6. Poisson input of average rate λ, unlimited queue, single service channel, with probability $n\mu\, dt$ of service finishing in the interval t to $t + dt$.

$$p(n) = \rho^n e^{-\rho}/n! \qquad W = Q/\lambda$$
$$N = \rho \qquad S = 1/\mu$$
$$Q = \rho + 1 - e^{-\rho}$$

7. Poisson input of average rate λ, unlimited single queue, single service channel with Erlang service time distribution of mean rate μ. $\lambda < \mu$.

P of service time from to $t + dt$ is

$$\mu(t)\,dt = (\mu k)^k\, e^{-\mu k t} t^{k-1} dt/(k-1)!$$

$$u = \rho$$

$$p(0) = 1 - \rho$$

$$N = Q + \rho \qquad W = Q/\lambda$$

$$Q = (k+1)\rho^2/2k(1-\rho) \qquad S = N/\lambda$$

When $k \to \infty$ the service time is constant as $1/\mu$ so that $Q = \rho^2/2(1-\rho)$.

8. Poisson input, unlimited single queue, one service channel with a general service time distribution of mean $1/\mu$ and coefficient of variation v. $\lambda < \mu$.

$$p(0) = 1 - \rho$$

$$N = Q + \rho \qquad W = Q/\lambda$$

$$Q = \rho^2(1+v^2)/2(1-\rho) \qquad S = N/\lambda$$

9. Poisson input of average rate λ, unlimited single queue, single service channel. Units in the queue are served in order of increasing service time (which is known accurately in advance). General distribution of service time, with probability $F(t)$ of a time less than t.

Average waiting time of a unit with service time t is

$$W_t = W_0 \Big/ \left[1 - \lambda \int_0^t t'\, dF(t')\right]^2 \qquad \text{if } t \leqslant t_0$$

$$N = W_0 \int_0^{t_0} \left\{ dF(t) \Big/ \left[1 - \lambda \int_0^t t'\, dF(t')\right]^2 \right\}$$

where units with service time greater than t_0 never reach the service channel, t_0 being given by

$$\lambda \int_0^{t_0} t\, dF(t) = 1$$

and where

$$2W_0 = \lambda \int_0^{t_0} t^2\, dF(t)$$

($W_t = \infty$ if $t > t_0$ and $N = \infty$ if $t_0 < \infty$.)

10. As for **9** but exponential service of mean rate μ.

$$W_t = W_0/\{1 - \rho[1 - e^{-\mu t}(1+\mu t)]\}^2 \qquad \text{if } t \leqslant t_0$$

$$N = \lambda\rho \int_0^{\infty} e^{-\mu t}\, dt/\{1 - \rho[1 - e^{-\mu t}(1+\mu t)]\}^2 \qquad \text{if } t_0 < \infty$$

where $\rho[1 - e^{\mu t_0}(1 + \mu t_0)] = 1$

and $W_0 = \lambda\rho$ if $t_0 = \infty$

$= (1 - \lambda t_0 e^{-\mu t_0})/\mu$ if $t_0 < \infty$

11. Two Poisson inputs, 1 and 2, of rates λ_1 and λ_2. Unlimited single queue, single service channel. Units 1 have pre-emptive priority over units 2. Exponential service of rates μ_1 and μ_2.

The results for units 1 are exactly as for queues type 1 (with λ_1 for λ and μ_1 for μ).

For units 2, the average number in the system is

$$N_2 = \rho_2[1 + \mu_1\rho_1/(1 - \rho_1)\mu_2]/(1 - \rho_1 - \rho_2)$$

REFERENCES

[1] BROCKMEYER, E., HALSTROM, H. L., and JENSEN, A. (1948). "The Life and Works of A. K. Erlang", *Trans. Dan. Acad. Techn. Sci.*, **2**.

[2] KOENIGSBERG, E. (1958). "Cyclic Queues", *Operat. Res. Quart.*, **9**, 22–35.

[3] COX, D. R. (1955). "The Statistical Analysis of Congestion", *J. R. Statist. Soc.*, **A 118**, 324–35.

[4] BENSON, F. (1952). "Further Notes on the Productivity of Machines Requiring Attention at Random Intervals", *J. R. Statist. Soc.*, **B 14**, 200–10.

[5] DOIG, A. (1957). "A Bibliography on the Theory of Queues", *Biometrika*, **44**, 490–514.

[6] EDDISON, R. T., and OWEN, D. G. (1953). "Discharging Iron Ore", *Operat. Res. Quart.*, **4**, 39–50.

[7] JACKSON, R. R. P. (1958). "Random Queueing with Phase-Type Service", *J. R. Statist. Soc.*, **B 18**, 129–32.

[8] MORSE, P. M. (1958). *Queues, Inventories and Maintenance* (New York, John Wiley).

[9] SAATY, T. L. (1957). "Resumé of Useful Formulas in Queueing Theory", *Opns Res.*, **5**, 161–200.

[10] LINDLEY, D. V. (1952). "The Theory of Queues with a Single Server", *Proc. Camb. phil. Soc.*, **48**, 227–89.

[11] MEISLING, T. (1958). "Discrete-time Queueing Theory", *Opns Res.*, **6**, 96–105.

[12] KENDALL, D. G. (1953). "Stochastic Processes Occurring in the Theory of Queues", *Ann. math. Statist.*, **24**, 338–54.

[13] WINSTEN, C. B. (1959). "Geometric Distributions in the Theory of Queues", *J. R. Statist. Soc.*, **B 21**, 1–22.

[14] KENDALL, D. G. (1951). "Some Problems in the Theory of Queues", *J. R. Statist. Soc.*, **B 13**, 151–73.

[15] NAOR, P. (1956). "On Machine Interference", *J. R. Statist. Soc.*, **B 18**, 280–7.
NAOR, P. (1957). "Some Problems of Machine Interference", *Proceedings of the First International Conference on Operational Research—Oxford 1957* (English Universities Press, London), pp. 147–64.
[16] COX, D. R. (1954). "Table for predicting the Production from a Group of Machines under the Care of One Operator", *J. R. Statist. Soc.*, **B 16**, 285–87.
[17] ASHCROFT, H. (1950). "The Productivity of Several Machines under the Care of One Operator", *J. R. Statist. Soc.*, **B 12**, 145–51.
[18] PALM, C. (1958). "The Distribution of Repairmen in Servicing Automatic Machines" (translated from the Swedish), *J. indust. Engng*, **9**.
[19] BAILEY, N. T. J. (1954). "On Queueing Processes with Bulk Services", *J. R. Statist. Soc.*, **B 16**, 80–7.
DOWNTON, F. (1956). "On Limiting Distributions arising in Bulk Service Queues", *J. R. Statist. Soc.*, **B 18**, 265–74.
[20] BAILEY, N. T. J. (1955). "A Note on Equalizing the Mean Waiting Time of Successive Customers in a Finite Queue", *J. R. Statist. Soc.*, **B 17**, 262–3.
[21] COBHAM, A. (1954). "Priority Assignment in Waiting Line Problems", *J. Opns Res. Soc. Am.*, **2**, 70–6.
COBHAM, A. (1955). "Priority Assignment—A Correction" (Letter to Editor), *J. Opns Res. Soc. Am.*, **3**, 547.
[22] STEPHAN, F. F. (1958). "Two Queues under Pre-emptive Priority with Poisson Arrival and Service Rates", *Opns Res.*, **6**, 399–418.
[23] PHIPPS, T. E. (1956). "Machine Repair as a Priority Waiting line Problem", *Opns Res.*, **4**, 76–85.
[24] BARRER, D. Y. (1957). "Queueing with Impatient Customers and Indifferent Clerks" and "Queueing with Impatient Customers and Ordered Service", *Opns Res.*, **5**, 644–56.
[25] COHEN, J. W. (1957). "A Survey of Queueing Problems occurring in Telephone and Telegraph Traffic Theory", *Proceedings of the First International Conference on Operational Research—Oxford 1957* (English Universities Press, London), pp. 138–44.

Chapter IV

STOCK CONTROL

Introduction

STOCKS of goods are held as a means of smoothing demand and as a means of supplying in convenience quantities differing from quantities made. Thus raw materials, goods being processed or finished articles may be stocked. Stock holding, however, is costly. It ties capital for both stock and buildings to house the stock. It induces deterioration and obsolescence risks, not to mention handling charges and insurance. In general manufacturing industry perhaps 90% of firms may have stockholding costs of between 12% and 24% p.a. by cost. It is therefore essential to balance the costs of holding stock against other costs, e.g. the costs of running out of stock. As so frequently happens in Operational Research problems, the object is to calculate best levels and not necessarily to minimise any particular one cost.

In a manufacturing concern, the interests of some departments may favour high but others low stock levels. For example, Production likes high stocks of raw materials and spare parts so as to avoid stoppages. At the same time production departments should not favour too many goods in progress standing on a factory floor because this may also lead to interference with production. At any one time sales departments like to draw from stocks of finished material to meet any demands from any customer, but on the other hand they may wish to reduce stocks when faced with the problem of redundant goods. There are a whole range of stock control problems where it is necessary to bring together, in some optimal way, two or more conflicting interests, and a scientific approach has been successfully made to these problems. All the factors which enter into any stock-control problem can be taken into account in this approach and are the variables in the situation.

The inequalities between supply and demand in any system may arise from either or both of two causes and it is as well to have them clearly distinguished. In the first place, stock may arise at a point of change between a batch process and a continuous process or a process with different size of batch. For example, a steel melting furnace may produce a batch of 300 tons of steel which has to be teemed into 15-ton ingots ; the steel must be held as " stock " in the ladle during teeming. This sort of situation is a very common one and in a controlled production process it may occur without any element of uncertainty. More frequently, however, stock problems arise when, in addition to a change in batch size, there is some uncertainty about the rate of supply or demand or both. In these circumstances there is a need to forecast. A typical type of problem on which much work has been done[1] concerns dams. The inflow arises in batches of variable size and at uncertain times ; how big must the dam be to ensure that a certain constant outflow can be satisfied with a given probability ? Similarly a warehouse orders goods in batches in order to supply them to retail outlets in smaller batches ; if the warehouse manager knows exactly how long it takes for goods to reach his warehouse from the moment he places his order (the lead time), and exactly what is the rate of demand for the goods from the warehouse, his problem is a fairly simple one : he will place orders for quantity Q at such intervals that immediately before receipt of each consignment his stock is zero, rising to Q when the goods come in. His average stock therefore is $Q/2$ and he has only the problem of determining the most economic batch size Q. If, however, as is more likely, the lead time is not known exactly but varies about a mean value and the rate of demand also varies, it is necessary to carry a buffer stock to reduce the risk of running out of stock because of lead time being longer than average and/or demand greater than average. An over-cautious warehouse manager can easily run up an excessive stock holding bill by working with a large buffer stock. The calculation of the most economic size of buffer stock depends on a knowledge of the distribution of demand, lead time and run-out cost. It is quite a separate problem from the calculation of the optimum batch size, Q.

A Warehouse Problem

In Chapter III an example was given of the solution of a simple stock problem by the application of Erlang's loss coefficient. In the present chapter a more complex example of an actual case study will be described, expanding the discussion from the particular to the general at appropriate points in the story.

The case concerns a warehouse, supplying a large number of outlets in a well-defined area, and re-ordering its stock via the organisation's head-office buyer. The number of items carried is of the order of 6,000.

The case was first brought to the notice of the company's O.R. section in order to determine whether the facilities provided by an electronic computer would enable worthwhile improvements to be made in the existing stock control and re-ordering system. The existent system of stock control and provisioning was based on a cyclical review, i.e. every item was reviewed at fixed periods and an order placed, which could be expressed by :

$$Q = D(L + 1) - (S_1 + S_2)$$

where Q = quantity ordered

$D(n)$ = forecast of demand over the next n periods

L = lead time in terms of review period

S_1 = stock in hand

S_2 = stock in pipeline.

The forecasting rule for $D(n)$ included a calculation based on the experience of demand in the same periods the year before and in fact was :

$$D(n) = \sum_{t=1}^{n} d'_t \sum_{t=-n}^{-1} d_t \Big/ \sum_{t=-n}^{-1} d'_t$$

where d_t = actual demand in period t this year

d'_t = actual demand in period t last year

with starting point the time of review.

The length of the cycle was four weeks for practically all items.

Stock Control System

In general, stock control systems can be divided into two classes, "Constant Interval" systems, of which an example was just given, and "Low Stock Level" (or "fixed order") systems, in which an order for a fixed quantity is raised, whenever stock falls to a pre-set "re-order" level. The re-order level must of course be sufficiently high to allow for lead time. There is an equivalence between the two systems, as regards number of orders placed over, say, one year. If the year is divided into n periods, then n orders will be placed, if the constant interval system is in use, and on average the same number of orders will be placed in the low stock level system, provided that the order quantity equals the average demand over one period. Which system to choose in a particular case depends mainly on technical considerations. The low stock level system is particularly suitable in circumstances where up-to-date information on stock levels comes to hand continuously, and the number of items to be reviewed is manageable, whilst the constant interval system is of advantage when there are large numbers of items to be dealt with simultaneously (e.g. many items to be ordered together from one merchant) and information on their movement becomes available only periodically.

As already mentioned, one of the basic items of information necessary for the design of a stock control system is the pattern of demand. Demand patterns can be roughly divided into four groups:

(i) interdependent, i.e. demand over a given period depends in some way on demand in previous periods;
(ii) time-dependent, i.e. demand changes with time;
(iii) random, i.e. demand over a given period is a random variable with a given probability distribution;
(iv) constant, i.e. demand is a fixed quantity.

There are, however, demand patterns which are a combination of two or more of the types mentioned above.

Forecasting

Forecasting rules depend on the demand pattern, hence the first step is to see into which of the above groups any actual

data can be fitted. Constant demand, easily recognised, does not require any forecasting; there is, therefore, no need to discuss this type any further at this stage. Again, inter-dependent and time-dependent demand are hardly ever found in their pure state, but almost always with a random element superimposed. Also certain types of demand pattern could be equally well grouped in either of these classes, e.g. seasonal patterns. The purpose of the grouping is solely a guide for the choice of a forecasting method. All patterns which show a periodic fluctuation will be assigned to class (i) above, those showing a trend over time in class (ii).

An easy way to recognise patterns belonging to class (i) is the computation of serial correlation coefficients.[2] The serial correlation coefficient for log i, V_i is defined by:

$$V_i = \sum_{t=1}^{n} \delta_t \delta_{t-1} \Big/ \sum_{t=1}^{n} \delta_t^{-2}, \text{ where } \delta_t = d_t - \bar{d}, \text{ and } \bar{d}, \text{ the mean}$$

of the d_t's.

The significance of the serial correlation coefficient, i.e. the probability that an interdependence between the data exists, can be obtained from statistical tables.[3]

If then any significant coefficients are found, an easy forecasting model is:

$$\hat{\delta}_t = \sum_{i}^{k} a_i \delta_{t-1}$$

if the first k serial correlation coefficients are significant.

The a_i are obtained by the methods of least squares and are themselves functions of r_i. The equations that have to be solved[3] for the a_i are:

$$\begin{aligned}
r_1 &= a_1 + a_2 r_1 + a_3 r_2 + a_4 r_3 + \ldots a_k r_{k-1} \\
r_2 &= a_1 r_1 + a_2 + a_3 r_1 + a_4 r_2 + \ldots a_k r_{k-2} \\
r_3 &= a_1 r_2 + a_2 r_1 + a_3 + a_4 r_1 + \ldots a_k r_{k-3} \\
&\vdots \\
r_k &= a_1 r_{k-1} + a_2 r_{k-2} + a_3 r_{k-3} + a_4 r_{k-4} \ldots + a_k
\end{aligned}$$

Time-dependent patterns can be recognised by computing moving averages and fitting a suitable curve. The number of

periods used for one average depends on circumstances; for example, if there is a periodic variation, say a seasonal fluctuation, the moving average would extend over the length of the entire fluctuation; on the other hand if no such periodic variation is suspected one would extend the moving average over the main planning interval, say, one year. In the case of time-dependent or mixed time-dependent, interdependent demand patterns a recommended forecasting method is that known as "Exponential Smoothing". A description of this method is given by A. Muir,[4] and the following formulae are taken from this article (or see reference 12).

A smoothing function of the demand figures d_t is defined as:

$$u_t = pd_t + (1 - p)d_{t-1}$$
$$= p[d_t + (1 - p)d_{t-1} + (1 - p)^2 d_{t-2} + \ldots]$$
$$0 < p < 1$$

u_t is therefore a sort of moving average of the d_t's, which gives decreasing weights to older data.

Let us assume now that a linear trend has been found in the demand pattern, say:

$$d = a + bt + e, \quad \text{where } e \text{ is a random element.}$$

The forecasting rule for T periods ahead becomes then:

$$D(T) = u_t + b_t(1/p + T - 1)$$
$$b_t = p(u_t - u_{t-1}) + (1 - p)b_{t-1}$$

b_t being a least squares estimate of b.

Other forms of time-dependent trends can also be used in conjunction with the smoothing function, including mixed ones.

The value of p is arbitrary and is best obtained by experiment. It is obvious that if p is large, the effect of the older data will quickly vanish and vice versa. By choosing a number of values for p, and comparing the forecasts obtained from these with actual data, one value for p which fits the data on hand best can be selected.

If neither interdependency nor time-dependency can be found, the assumption that the demand behaves as a random variable is justified. One distribution that has often been found as a suitable description of the demand pattern for

consumer goods is the Poisson distribution. In most cases of random demand the best forecast that can be made is independent of the actual form of the distribution, and is the average demand. This distribution, however, is important in the computation of economic safety stocks, which will be dealt with below.

In the case that is here used as an example, it was soon discovered that the existing rule was not a good one and the expert in charge in fact often found it necessary to change the forecast obtained from it. A course of action, on the lines described above, was, therefore, followed and the demand pattern for the majority of items in the warehouse was found to be a random one, with a certain complication, namely that the average demand could change fairly suddenly and permanently to a different level, and also had a long-term drift in some items. The reason for the changes in mean level lay in the particular type of customers of the retail establishments which formed a sort of floating population.

The model of demand for the ith item over the tth period was therefore assumed to be :

$$d_{it} = K_{it}m_i + \varepsilon_{it}$$

where m_i = mean demand of i^{th} item

K_{it} = a constant, which is a seasonal factor.

For items not exhibiting a seasonal fluctuation,

$K_{it} = 1$ for all t

ε_{it} = a random variable,
with mean = 0, variance = σ_i^2 ;
and ε_{it} independent of $\varepsilon_{i,t+j}$ for all j.

This model is independent of the length of period, provided all periods are of equal length.

The next step was to find the lead time pattern. Here again, one can either have a constant lead time, which unfortunately is very rare, or one conforming to a statistical distribution. In the case at hand this was found to be normally distributed with a mean of 8 weeks for almost all items.

Application to the Warehouse Problem

On the basis of the demand model and lead time distribution

and the fact that the computer could hold continuously up-to-date stock figures, the following low stock level system was examined :

I. For items not in seasonal demand, i.e. items whose $K_{it} = 1$ for all t.

Fixed order quantity, Q initially fixed arbitrarily as equal to average 4-weekly demand

$$Q = m_i$$

Re-order level equal to average demand over 8 weeks plus a buffer stock to minimise risk of run-outs (the computation of this buffer stock is dealt with below).

$$R = 2m_i + B$$

The computer to hold continuously the following for each item :

R (re-order level), S_1 (stock in hand), S_2 (stock in pipeline). S_1 can be negative (i.e. when there are some orders which cannot be met due to lack of stock on hand).

Whenever a change in S_1 due to demand occurs $S_1 + S_2 - R$ is computed and if the result is equal to or less than 0, a signal is given to raise an order for m_i of the item in question.

In order to discover any permanent changes in the level of average demand, it was also suggested that control charts be set up in the following way :

Let $$X_{i,t} = \tfrac{1}{2}(d_{i,t-1} + d_{i,t})$$

then $X_{i,t}$ will be approximately normally distributed with mean m_i, and variance $\simeq \sigma_i^2/2$.

Therefore a control chart can be run on the X_i, and the order quantity and re-order level can be adjusted, whenever the chart indicates with a given level of risk that a permanent change has occurred. As the computer accumulates total demand, the X_i can be produced regularly and if necessary at intervals of less than 4 weeks.

II. The only changes required for items in seasonal demand, i.e. items where K_{it} is different from 1 for some t, are :

Order quantity in period t :

$$Q_t = K_{i,t+2}m_i$$

Re-order level in period t:

$$R_t = m_i \sum_{j=0}^{1} K_{i,t+j} + B$$

And the mean of the control chart in period t:

$$\bar{X}_{it} = K_{it} m_i$$

In order to demonstrate to the responsible managers that the proposed system was workable in practice and would have positive advantages, a simulation of it was carried out.

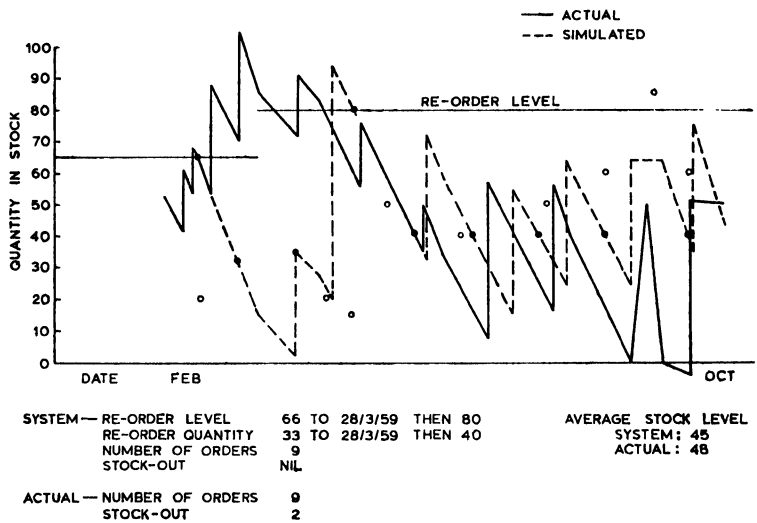


FIG. 4.1

Fifty items were selected at random, and, starting from the stock levels at 1st January 1959, they were, on paper, subjected to the suggested stock control system for the first 10 4-week periods of the year. The simulated stock history thus obtained was compared with the actual stock history; an example is shown in Fig. 4.1. It was found that for all 50 items the suggested system would lower the average stock level by about 20% and the average number of orders placed by 10%, and there would be fewer incidents of "stock-out".

At this point the first part of the problem—the principles of a stock control system to fit in with a computer—was solved. It was, however, by no means the end. On the evidence of

the simulation the suggested system was both feasible and better than the one in use ; to test whether it was also less costly, it was necessary to turn to an investigation of costs associated with a given system.

Economic Order Quantity

In stock control problems the costs associated with stocks are split into three parts :

1. The cost which is a known function of the number of orders placed and the size of the order.
2. The cost which is a proportion of the stock value.
3. The cost which is a known function of the number of occasions when the stock is not available.

These three costs correspond roughly with ones which are termed ordering, holding and stock-out costs.

The cost of ordering is interpreted differently according to the type of stock ; to determine the quantity to be produced in one batch, ordering cost is synonymous with the set-up cost, whereas in determining the input of raw material it includes the cost of sending out an enquiry, checking tenders, placing an order, receiving delivery of raw material in the storehouse and clearing payment.

The cost of holding stock depends on the materials stored, their rate of obsolescence, methods of handling in the storehouse and the interest charges on the capital tied up in the material (or the opportunity cost if capital is short and there are competing demands for it). This cost is often proportional to the value of the stock, in which case the price per unit of stock enters into the problem.

Any stock item has associated with it a cost which is incurred whenever that item is not available. The make-up of this cost depends on the particular stock control problem ; stock-out cost of raw material or spare parts is made up of plant down time and special delivery of the necessary material. Stock-out cost of a finished article is made up of a cost of dissatisfied or lost customers or of emergency action to obtain supplies in a hurry.

All these costs are somewhat involved and to determine them an appropriate form of cost analysis must be made to

suit the circumstances of each particular case. It is not possible to discuss here the costings applicable to the case being considered. Suffice it to say that it is essential to management to reach an acceptable estimate of cost. It is no good saying, for example, that the cost of a lost customer cannot be determined ; for that would be to say that a proper stocking policy cannot be determined. Any decision on stock policy implicitly involves an assumption regarding stock-out cost.

For any particular problem, therefore, one must determine three sets of costs :

$C_1 = f(N, Q)$ Order cost is a function of number and perhaps size of orders.

$C_2 = y(S)$ Holding cost is a function of average stock level, which in turn depends upon the size of order and the amount of buffer stock.

$C_3 = p(B)$ Stock-out is a function of the buffer stock, since the larger the buffer stock the less the risk of stock-out in any given situation. It is the number of occasions out of stock or possibly the length of time out of stock according to the context.

Setting $C = C_1 + C_2 + C_3$, it is required that C be a minimum.

The first two of the above functions can be made explicit by letting :

h_1 be the cost of raising one order.
D the total demand, known or forecast, over the planning period.
h_2 the cost of holding a unit in stock.
Q the order quantity.
B the buffer stock.

$$C_1 = h_1 D/Q \; ; \quad C_2 = h_2(B + Q/2)$$

since the expression in the bracket is the average stock level. For the time being C_3 is left in functional notation. To minimise C, differentiate C separately with respect to Q and B, and obtain values for Q and B, which will make C a minimum

by setting the partial derivation equal to 0 and solving the two simultaneous equations for Q and B.

Thus
$$C = h_1D/Q + h_2(B + Q/2) + p(B)$$
$$= h_1D/Q + h_2Q/2 + h_2B + p(B)$$

Differentiating with respect to Q

$$h_2/2 - h_1D/Q^2 = 0 \quad . \quad . \quad . \quad . \quad (1)$$

Differentiating with respect to B

$$h_2 + \frac{dp(B)}{dB} = 0 \quad . \quad . \quad . \quad . \quad (2)$$

Hence the ordering quantity which will minimise total cost is independent of the buffer stock and can be obtained from equation (1) above.

It is
$$Q = \sqrt{\frac{2h_1D}{h_2}}$$

This quantity Q is often called the economic ordering quantity, and it is applicable to the constant interval system as well as to the low stock level system. In the former it fixes the length of the review cycle and in the latter the re-order quantity.

The length of the re-order cycle is obtained by dividing D/Q, the number of orders that it is most economic to place, into the planning interval: this is k, the length of the review period (or average time between orders in the case of low stock level system) in, say, weeks. Some adaptations however need to be made in case the ordering cost depends partly on the size of the order; for example when suppliers give a quantity discount. If in order to obtain discount d, the order quantity is increased by n weeks average supply the average stock level will rise by $nD/104$ (assuming 52-week planning period). Setting the cost value of the quantities ordered over the planning interval, $uD = A$, gives:

a saving in order cost equal to $52h_1\left(\frac{1}{k} - \frac{1}{k+n}\right)$

$= 52nh_1/(k^2 + kn)$,

a saving in the cost of supplies equal to Ad,

but on the other hand increases stock-holding costs by:

$h_2nD/104$

Therefore, as long as the saving achieved is greater than the increase in carrying cost, i.e.

$$Ad + 52nh_1/(k^2 + kn) > h_2nD/104$$

it pays to increase the computed Q by n weeks average supply, given a quantity discount d.

Turning now to the cost of stock-outs, and thence to the optimal size of buffer stock, again two types of cost have to be balanced against each other, namely the cost of holding a buffer stock and the cost of run-out.

Let h_3 be the cost of being short one unit

$P_1(d_i = N)$ the probability that demand over i days equals N.

$P_2(i)$ the probability that lead time equals i days.

(These probabilities are obtained from the distribution of demand and lead time),

then the probability of shortage of n units, v_n, during a lead time of i days and buffer stock B is given by

$$v_n = \sum_i P_1(d_i = B + n)P_2(i)$$

the i ranging over all possible values of lead time in days.

The average cost of shortage is therefore $h_3 \sum_n v_n n$.

The cost of holding a buffer stock B is h_2B.

Hence minimise $h_2B + h_3 \sum_n v_n n$.

This can be done by differentiation with respect to B ; the explicit form of the cost function depends, of course, on the distribution functions and differentiation may not always be simple. An easier way would be to find the minimum graphically.

The average cost of shortage is a decreasing function of B, whilst the holding cost is an increasing function of B, the latter in fact is a straight line through the origin, forming an angle with the horizontal. Graphing both functions on a common graph, it is not too difficult to draw the curve which is the sum of the functions, and thence read off the value of B at which that curve reaches its minimum.

These formulae establish the model of optimal stock control (optimal in the sense of least cost) in *most* cases. The model is applicable to both types of stock control system.

It remains to set the re-order level, R, for a system of the fixed order type, and the actual order quantity, O, for a system

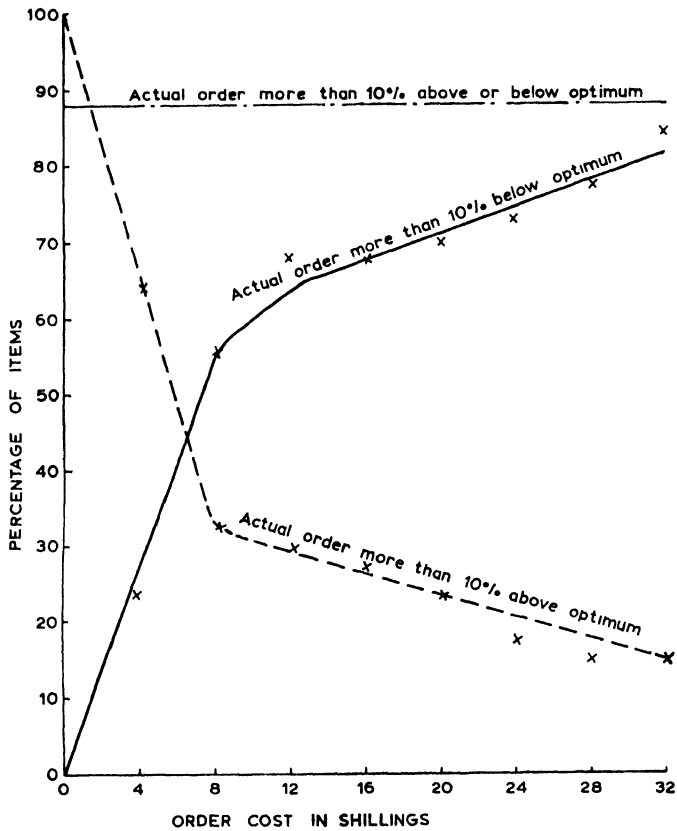


FIG. 4.2

of the cyclical review type. These turn out to be quite naturally :

$$R = Q + B\,; \quad O = Q + B - S$$

where S = free stock, i.e. stock on hand plus stock on order.

Returning now to the case study, it will be remembered that the system suggested as a first solution did not concern itself with optimisation at first. One of the difficulties was to estimate the order cost, C_1, as the introduction of the system itself was bound to change this, and therefore, the cost data

at hand did not supply a reliable guide. Hence it was decided as a first step to enquire whether that sort of optimisation would in fact be important, or, in other words, and considering the large number of items to be dealt with, whether the currently used 4-weekly review period is not sufficiently near the optimum. A random sample of 50 items was already at hand and the holding cost, C_2, was known as a percentage of retail value. Q, the economic re-order quantity, was computed for various assumed levels of order cost (4/– to 32/– in steps of 4/–). The percentage of items, whose Q was more than 10% above or below the mean 4-weekly demand (which would be equivalent to Q, if in fact 4-weekly intervals were optimal), was then plotted against order cost (Fig. 4.2). It was found that for nearly 90% of all items the actual order quantity differed from the theoretical optimum cost by more than 10%, regardless of order cost.

It was proposed therefore to install the stock control system in two stages ; firstly a system based on an O & M study but to bring into play the principles as described above with a 4-weeks average order interval. During that stage data would be collected to estimate order cost and thence values for Q and B. The second stage would not change the actual system but would introduce the computed optimal values.

More Complex Systems

It is quite often necessary to deal with a central stores and some subsidiary stores. Thus, an army will have central storehouses and depots in the home country and subsidiary storehouses, fed from the central stores, wherever forces may be posted ; or a large company may have a large central stores organisation with separate stores at factory or plant level. This is illustrated in Fig. 4.3.

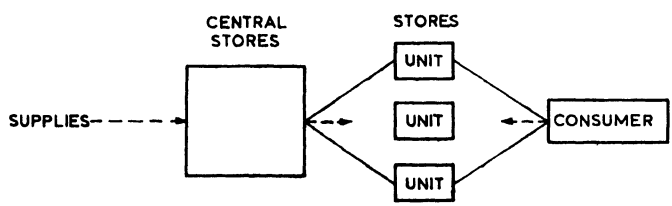


FIG. 4.3

In this type of organisation there are problems of finding which items should be centrally stocked or how often should items held centrally be distributed to the subsidiary stores. Very often centralised holding leads to economy in stock-holding and operation of the stores and to better care of material. In such instances, factors against central stores would be poorer availability of goods and added transport costs. In any given situation it is possible to enumerate the factors for and against, and to deduce conditions that determine the stores location.

More centralised holding or control of expensive stores items may produce savings in total costs by stock-holding reduction. Suppose an item with a cost P and delivery time L has a Poisson demand with average D in each of n depots. What condition would allow centralised stores holding for the n depots? For a given level of protection, each depot would hold a stock reserve of $K\sqrt{LD}$ where K is a constant from the Poisson distribution depending on probability of stock-out. Since there are n areas, the total reserve is $nK\sqrt{LD}$. For the same level of protection, stock reserve at one centralised point is $K\sqrt{nLD}$.

Thus the stock-holding reduction is at least $K\sqrt{nLD}(\sqrt{n}-1)$. If stock-holding cost at the depots is C_1, and is the same at the central point, the money saving is $C_1K\sqrt{nLD}(\sqrt{n}-1)$.

This saving is gained on the expense of some increase in transport cost. The transport cost is likely to vary with the amount of goods involved, so let it cost T on average to move goods from the central store to each of the depots. $\therefore$ Total extra transport cost $= nTD$.

Centralised holding follows the condition

$$nTD \leqslant C_1K\sqrt{nLD}(\sqrt{n}-1)$$

i.e.

$$\frac{C_1^2L}{D} \geqslant \frac{nT^2}{K^2(\sqrt{n}-1)^2}$$

The expression has on the right-hand side a quantity which can be calculated for a given group of stores and regarded as a constant value for that set-up. The condition for centralised stock-holding can therefore be shown on a diagram; the position of an item on the graph is set by its price and con-

sumption, and if that item falls in the shaded portion, that is the portion above its own particular lead time, then it is worthwhile stocking that item in a central store (Fig. 4.4). A more complex situation is illustrated in Fig. 4.5.

In optimising ordering and holding for stocks (*A*) and (*B*), a buyer sets certain ordering quantities. At the extreme right-hand end of the diagram, in a new role as a seller he tries to find the best delivery sizes. Ideally one model would link

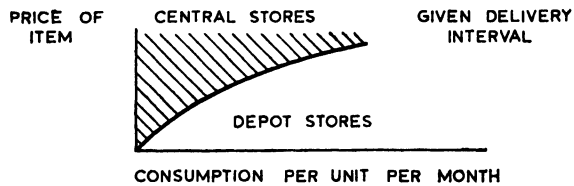


FIG. 4.4

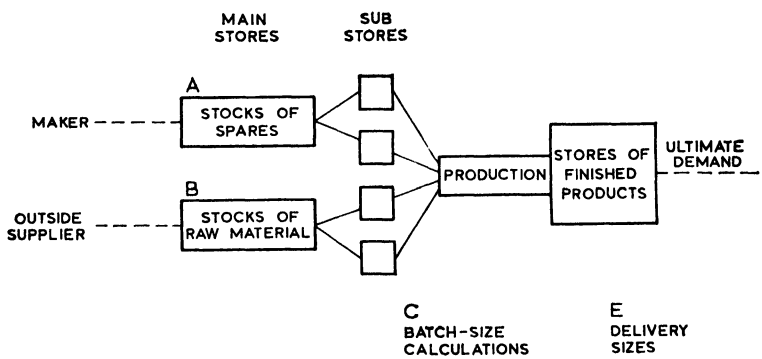


FIG. 4.5

the whole process, but the models apply to individual stocking points.

The first example of economic ordering quantity calculation would be applicable for batch size calculation under *C*, or for stocks *B* for certain restricted conditions.

The second example, where order quantity and re-order level are calculated for variable demand and lead time, would apply to stores *A* and *B* in the diagram, though the problem might well be made easier through having constant demand from stocks of raw material. The method of the second

example would also be applicable in stores *D*, of finished product, when there should be no lead time but a variable demand ; these stores exist in order to allow constant production in the face of fluctuating demand. An exercise on discounts would be applicable to stocks *B*, of raw material. Conditions for the location of stores would place goods in the main stores or the sub-stores.

Associated Problems

Stock control based on these mathematical methods is not generally being used at the present time. The models are not easy to interpret or to apply but they do have obvious benefits. It is possible in practice, though not to be recommended except as a preliminary measure, to make considerable improvements on existing routine by part-application of a scientific approach to stock control. For example, where run-out costs are not known, it can be shown that rationalising the routine reduces the number of run-outs without increasing ordering frequency or stock-holding.

Stock control is often effected by stating that " n " months' supply of a certain group of articles should be kept in stock. Any change of policy, for example, regarding stock-outs or stock levels means changing the number of months' supply, " n ". If any of the methods outlined above were in use such a change of policy would only involve changing the appropriate figure in a formula or table and the effects of the policy would be readily understood and measurable. For example, a reduction of stocks could be obtained within a few months by more frequent ordering, or much more quickly by reducing re-order level and thus increasing the probability of stock-out. Scientific routines, when in use, would find the effect of changes in policy.

Models of stock control form a theoretical basis for starting an investigation. Immediately on application added problems arise. Distributions of demand are not always Poisson. Lead times are nearly always indeterminate, and there may not be sufficient data to check their distribution (often assumed to be normal). Costs of stock-outs are difficult to obtain or evaluate. A large number of stock items may be encountered, and—from purely a time and investigation cost basis—each

of these items cannot be treated individually by the recognised methods. Order costs have, so far, been assumed to vary directly with the number of orders, whereas this is not necessarily the case; similarly total holding costs are not always a fixed percentage of the stock value.

This complexity of factors points to the need for application of further operational research to stock control problems. There are simple applications but these should be done as a matter of course and should not need any investigation. Existing mathematical models should be fully known and understood by the research worker, whereas the main results outlined in the previous section are all that are required by the practitioner of provisioning. These results have to be supplemented by ready reckoners for re-order quantity and the use of nomograms for finding the re-order levels is recommended. These practical aids are discussed very fully by W. E. Welch.[5]

A formula with a square-root sign is not readily acceptable in many provisioning departments, and it is preferable to provide tables or graphs where the re-order quantity is read from price and demand. For a department whose costs are 10/– for placing an order and 10% per annum for holding stock the best re-order quantity

$$= \sqrt{\frac{2 \times 10 \times \text{Demand p.a.}}{0{\cdot}1 \times \text{Price (shillings)}}} \simeq 4\sqrt{\frac{\text{Monthly demand}}{\text{Price (shillings)}}}$$

It is preferable to provide a table of this result for convenient price and demand ranges than to recommend using the result itself.

Re-order levels can conveniently be obtained from the average demand and lead time on nomograms which can be drawn for any stock-out probability level. Proper re-order levels require knowledge or assumptions about the pattern of demand and lead times. A point to bear in mind here is that lead times are often not random; they are influenced from time to time by customer–supplier relations, and often chance deliveries are mixed with phased deliveries, where, in the latter case, orders are placed well in advance, and longer lead times are accepted when deliveries are promised for given dates. Often a buyer will confirm that lead times are very variable,

but can give an assurance that goods are nearly certain to be available in some longer period than the average lead time. In practice, records will show very variable lead times because they represent both routine and urgent orders.

Two important interactions come into play and are most conveniently said to arise in connection with the re-order level. One is that, in the mathematical models, probability of run-outs has been determined, and these run-outs are only liable to occur after the re-order level is reached and an order has been placed. Thus, there is an apparent probability of run-out but the real effect depends on the frequency of ordering. So a margin of a 1 in 100 probability of run-out may be required when stocks must be re-ordered monthly, whereas the probability level can be relaxed to 1 in 10, say, if stocks are re-ordered annually. It is advisable to check re-order frequency first, and then to think about probability of stock-out, and to do this in terms of failure per year instead of failure per order.

The second interaction effect is between the average size of issue of an item and the re-order level. If an article is issued singly to one outlet of consumption and in dozens to another, the re-order level not only has to allow for average consumption in the lead time but also for the number of demands for one dozen batches. With extreme differences, the simplest solution may be to provision separately according to the unit of issue. The unit of issue is usually marked on every stock record card and it has two uses ; one is in stock control, the other is in stock evaluation, where it is convenient to give the unit of issue that the article is usually priced at. For stock control purposes, the recorded unit of issue is often misleading and useless.

Refinements are being made in the mathematical stock control models. For example, in *Introduction to Operations Research* by Churchman, Ackoff and Arnoff,[6] the second and third chapters introduce price breaks (discounts) and restrictions on lot size into the models. Setting up suitable models is the first important aspect of applying these various techniques to stock control. The second is the practical application, and the final, and often the most difficult one, is to obtain the correct parameters to use in the models.

It is often useful to make a detailed study of the factors affecting ordering, stock-holding and stock-out costs in order to reduce the costs themselves as well as optimising them, which is the main O.R. contribution to stock control. Although ordering cost, for example, may not always vary directly with the number of orders placed, it is often convenient to obtain a rough estimate of it by dividing the total cost of an ordering department by the number of orders placed. Bailey[7] suggests that ordering costs are more likely to vary directly with the number of orders placed in large than in small organisations. The first approximation to the ordering cost can be used to apply the scientific ordering quantity formula, and observations will show whether this changes stock levels and present ordering practices.

Apparent difficulties in ascertaining exact ordering, holding or stock-out costs will not appear to be so serious if large savings are made when the rational techniques are tried out in practice. Large anomalies are removed, such as monthly deliveries of trivial amounts, say, less than £1 worth of goods, at a receiving into stores cost of perhaps 8/– per article ; or at the other end of the scale placing and receiving in one batch a large order for a year's supply of goods worth £500 to £1,000. These effects may be seen in descriptive statistics of the items held in the storehouse. If 1,000 different items are held in a store, from a stock control point of view their physical descriptions are no more important than their price and consumption. Three useful distributions are those showing the number of items falling in various ranges of price, value of stock held, and value of monthly consumption. If stock records are on punched cards, these distributions are easily obtained, and in addition the amount of stock-holding can be shown for ranges of price and values of stock held. It would probably be worthwhile studying the mathematical properties of some of these distributions as they are likely to be generally applicable.

If a change of $p\%$ is made on the stock value of a subset of items accounting for $q\%$ of total stock value, then the overall effect on total stock value is a change of $\frac{pq}{100}\%$. It is easy to lose sight of these simple points when for instance a group of about 1,000 stores items are controlled by a broad rule which

says that "*n*" months' stock should be held for the whole group. Without completely rational provisioning and stock control techniques each article comprising the 1,000-item storehouse may be restricted to "*n*" months' stock with resulting excessive buying costs. Again, if stocks are to be held at "*n*" months, this usually refers to the average stock level, so that there is no need to set a maximum stock level of "*n*" months on each article.

If stocks are controlled by target levels of this sort, the distribution of monthly consumption will show what the target level ought to be according to the scientific provisioning techniques outlined earlier.

Often a reduction in the number of orders will not lead directly to a reduction in staffing, and fewer stock-outs are not readily appreciated. When stock target levels have been set, reduction of stocks is the only rapid and positive result. A process of optimising does not guarantee this result, because the whole policy behind the stock target levels may have been erroneous in the first place.

A short list of references is attached. Published work on this subject goes back about forty years, when attention was first paid to optimising order quantities and production lot sizes. Recently, and over a relatively short period of time, the treatment has become considerably more involved in mathematics. Amongst the references it is worthwhile paying special attention to Whitin's book *Theory of Inventory Management*[8] because this is a good general introduction to stock control problems in their widest application and traces the development of mathematical models. The chapters on stock control in the textbook *Introduction to Operations Research* by Churchman, Ackoff and Arnoff[6] attempt to consider all the factors that enter into the problem and to combine these into mathematical models likely to have a wider application. The difficulty with this approach is that slight variations in the problem lead to large differences in the models and it may be preferable to defining the likeliest problems rather than the generally applicable models. The treatment in this textbook is invaluable to the Operational Research worker but unlikely to be of much help to the practitioner. The textbook which covers most of the problems very ade-

quately for the practising stock-control executive is one called *Scientific Inventory Control* by W. E. Welch.[5] The author here is very conscious of the difficulties of applying stock control solutions and introduces numerous devices. Most papers on the subject are of a specialised mathematical nature ; two useful introductory papers are ones in the *Purchasing Journal* by Bailey[7] and in *Accountancy* by Hepburn.[9] The proceedings of a Conference held in January 1954 at the Case Institute on Operations Research in Production and Inventory Control[10] give case studies and an interesting and general introduction to the whole topic. A very useful comprehensive discussion of the subject is provided by Magee's *Production Planning and Inventory Control,*[11] and Brown's *Statistical Forecasting for Inventory Control.*[12]

REFERENCES

[1] MORAN, P. A. P. (1959). " The Theory of Storage ", *Methuen's Monographs on Applied Probability*, London.

[2] WIENER, N. (1949). " Interpolation and Extrapolation and Smoothing of Time Series ". New York.

[3] Notes on Operations Research Summer School, M.I.T. 1953.

[4] MUIR, A. (1958). " Automatic Sales Forecasting ", *The Computer Journal*, Vol. 1, No. 3.

[5] WELCH, W. E. (1957). *Scientific Inventory Control.* Management Publishing Corporation, Greenwich.

[6] CHURCHMAN, C. W., ACKOFF, R. L., and ARNOFF, E. L. (1957). *Introduction to Operations Research.* New York : John Wiley & Sons Inc. London : Chapman & Hall Ltd.

[7] BAILEY, P. (1957). " Purchasing for Stock ". Part 1 : " When to Buy " ; Part 2 : " How Much to Buy ", *Purchasing Journal*, September and October.

[8] WHITIN, T. M. (1953). *Theory of Inventory Management.* Princeton University Press.

[9] HEPBURN, J. H. (1957). " The Control of Stock ", *Accountancy*, Vol. 68, March, pp. 106 10.

[10] Proceedings of a Conference on Operations Research in Production and Inventory Control, Cleveland, Case Institute of Technology, January, 1954.

[11] MAGEE, J. F. (1958). *Production Planning and Inventory Control*, New York : McGraw-Hill Book Co.

[12] BROWN, ROBERT G. (1959). *Statistical Forecasting for Inventory Control.* New York : McGraw-Hill Book Co.

Chapter V

SIMULATION IN BUSINESS STUDIES

THE simile is a literary trick as old as words. With its help the best of writers have been able to project an image into the reader's mind with such sudden clarity that every detail of a complex situation or mood is felt and understood.

The fruits of science do not fall quite so quickly but the simulator designer also hopes to throw light on complicated problems by considering more familiar systems.

Just as different poets will use different similes to evoke the same image, so different simulators may be used to study the same problem. Each simulator will obey the same mathematical laws and give the same answers to the problem, although one may be a mechanical simulator with moving parts and another an electronic type.

No one would take "All the world's a stage . . ." as a denial that the earth is round, and observers must be equally careful not to interpret the simulator's behaviour as giving significant results outside the field for which it was originally designed.

It would be as well to start by listing some of the things which a simulator can and cannot do.

Computation. Solves complicated equations.

Research. Tests how successfully a proposed expression describes the behaviour of a given system and indicates modifications to the expression to give closer agreement.

Development and Decision. Allows new ideas to be tested before being tried in practice, thus facilitating experiment and informed decision.

Training. Duplicates the behaviour of the real system, possibly on a faster time scale, so that trainees can quickly, cheaply, and painlessly acquire experience.

Finally, and very important, a simulator can never be a substitute for original thought, although it will usually be a very potent stimulus to such thought.

Digital Computation. Monte Carlo Simulation

There are two types of machine available for the solution of difficult mathematical problems: the digital computer and the analogue computer. Although the former is not used in simulation proper it is sufficiently often employed in related fields to merit discussion here, if only to dispel misconceptions about its use as a simulator.

The digital computer handles all its data in the form of numbers which it manipulates by counting, like a child doing arithmetic.

The basic unit in such a computer may be a wheel with ten positions each of which represents one of the ten digits, as in a mechanical desk calculator. In an electronic computer the basic unit might be four valves each of which may be switched on or off, so that the $2^4 = 16$ different combinations of valve states represent sixteen different numbers.

The greater the number of such units the greater the number of significant figures which the machine can handle. Thus the accuracy of a digital computer is always well defined and can be increased at will, by using more numerical units, to any extent justified by the accuracy with which the data of the problem is known.

The high-speed digital computers which have become common in the last ten years are usually electronic. Essentially they perform the operations of addition, subtraction, multiplication and division, but they have two extra features, "memory" and "logical facility", which have led to the misleading name of "electronic brain".

Children learning arithmetic use squared paper and are taught to write one digit in a square. In an electronic digital computer specific instants of time are used rather like the squares in the child's book. A special electronic clock produces pulses of voltage, perhaps a million per second, and numbers in the machine are represented by other voltage pulses occurring at the same time as the clock pulses. Each clock pulse is like a square in the arithmetic book and a digit or a 0 is written into the square by the generation or not of a voltage pulse at the same instant as the clock pulse. Numbers in the machine are thus represented by trains of pulses and

these can, if necessary, be stored for long or short periods in a " memory " which may be magnetic tape.

The logical facility of a machine might consist essentially of its ability to select one of two operations to perform next, depending upon whether the number resulting from the previous arithmetical operation was greater or less than another number stored in the machine.

There is a separate store in which instructions to the machine, coded as numbers, are kept. Such instructions might be : " select the number in store x and add it to that in store y ", or " test the number in store a and if it is bigger than that in store b repeat the previous instruction, but if it is smaller than the number in store b proceed to instruction seven ".

In the absence of any instructions the machine will perform perhaps a dozen very simple operations in order. Programming is the invention of a set of instructions which will modify this simple machine cycle in such a way that the required calculation is performed.

In general the programming of a particular problem will take hours while the time taken by the machine to solve the problem after the programme has been inserted will be only minutes. Certain operations will be common to a large number of programmes and these sub-routines are stored in a programme library so that they may be used *en bloc* in new problems.

The digital computer provides an answer to a given problem with a single set of initial conditions. If the initial conditions vary then the computer will have to solve the problem afresh for each different set of conditions. We shall see later that the analogue computer provides a continuous solution to such a problem, the solution displayed at any instant corresponding to the conditions existing at that instant. The digital computer will provide a discrete number of solutions, depending on how many separate times it is required to solve the problem.

Thus the output of the analogue computer may be considered to be a continuous curve for varying conditions, while that of the digital computer is a set of straight lines joining points, each point representing a separate solution but the state of affairs between the points is not explored.

For this reason the digital computer is not used in true

simulators, but its role in certain operational research problems is loosely referred to as simulation. These are the problems which are attacked by what is known as the " Monte Carlo " method.

A mathematical expression is invented which describes the process under investigation for one set of initial conditions only. Different sets of initial conditions, for example the arrival of customers in a shop, occur in a random manner or according to some other known distribution law. The digital computer can be used to generate a set of random numbers which define a set of initial conditions, each of which will give a different solution to the mathematical expression. All these separate solutions together describe the general behaviour of the system and the more solutions there are, the more complete is the picture. By using a digital computer first to generate very large quantities of random numbers and then to solve the equation the corresponding number of times, an accurate description of the system may be obtained.

Thus, suppose a shop has 500 customers per day and it takes an assistant 5 minutes to serve each. Then, for a given number of assistants, the time a customer must wait to be served can be calculated if the arrival pattern of customers is known. Assuming a random arrival pattern a very large number of calculations of the waiting time can be made for a very large number of possible customer arrival situations. The result will be a set of waiting times which are likely to occur in practice and from which the longest waiting time, typical waiting time, etc., can be found together with the intervals for which assistants are idle, and the effect of changing the number of staff.

The calculations performed by the digital computer thus produce a set of results which are similar to those which could have been obtained by observing the shop over a long period. To this extent the computation is a substitute for experiment and for this reason the procedure is often referred to—wrongly in the writer's opinion—as simulation.

Analogue Simulation

Simulation proper depends on the analogue method. Here the system under investigation is analysed and a mathematical

equation obtained which describes its behaviour. Next an electrical circuit is invented whose behaviour is described by an identical mathematical equation. This circuit is the electrical analogue of the original system and experiments made with it will yield results identical with those which would have been obtained by carrying out the same experiments on the real system.

Suppose we have a system where we are interested in a quantity Q, which might be the distance travelled by a rocket, or the number of tons of a material held in stock, or the sales of a particular product. We will usually be concerned with the way in which Q behaves when the various factors which determine it are changed by known amounts.

If it is a relatively simple system we might find that the value of Q is determined by three variables. In the case of the sales of a product these variables might be the price of the product, the number of retailers stocking it, and the amount of advertising. Let these three variables be a, b, c.

Thus we can say that in the system being simulated Q depends upon a, b, c, and in addition upon time, t. Time will be part of the expression relating Q with the three variables because in general changes in the variables will not at once produce their full effect upon Q which will usually take a significant time to reach its new value.

Q will thus depend upon a, b, c and t in a manner which may be simple. In real technical or business systems, however, the relation is likely to be complex with the values of a, b and c themselves interdependent to a small or large extent.

Thus $Q = f(a, b, c, t)$, where the function f is probably of some complexity, is the mathematical expression which describes the behaviour of the system.

An electrical circuit must now be invented whose behaviour is described by exactly the same function f. In particular, at some point in the circuit there will be a voltage V which will represent Q. In a simple case the equation representing the behaviour of the circuit might be written $V = f(r_a, r_b, r_c, t)$ where r_a, r_b, and r_c are resistances which appear in the electrical equation in the same form as a, b and c in the original equation.

Thus the electrical circuit with its equation $V = f(r_a, r_b, r_c, t)$ is the analogue of the real system with its equation

$$Q = f(a, b, c, t)$$

Further, V is the analogue of Q, r_a the analogue of a, r_b of b, etc.

If values of r_a, r_b and r_c are set in the electrical circuit—and this will be done by turning one dial for each—corresponding to the values of a, b and c in the real system, then the voltage V will correspond to the value of Q in the real system. If now one of the quantities, say r_a, is changed in a certain way then the voltage V will vary in precisely the same fashion as Q would have done in the real system had a undergone the same change.

The voltage V will normally be measured with a pen-recorder voltmeter so that in the simulator the important quantity Q is continuously displayed and a permanent record kept of its fluctuations.

In many of the systems studied the behaviour of Q over periods of many months may be of interest. In such cases it is convenient that the electrical circuit shall be designed to operate on a speeded-up time scale so that a year of fluctuations in the real system is performed by the simulator in a minute or so.

Suppose we are concerned with the behaviour of Q in a real system over a period of a year, and suppose also that one of the quantities which determines Q, say a, is altered once per week in reality.

Then the time scale chosen for the simulator must be such that the whole " year " of experiment is conducted in a reasonably short time. But since a has to be changed once per week the time scale must be slow enough to enable the operator to make an alteration in the r_a dial every simulator " week ".

It might be convenient to use a time scale of five seconds to one week so that a run covering one year of the real system would take less than five minutes. Apart from the practical requirement of being slow enough to allow manual setting of the controls, this sort of time scale has the advantage that the operator is able to follow intelligently the fluctuations in Q as they occur. At the end of five minutes he has acquired much the same experience as he would have done by plotting the results of the real system over a year.

Having discussed the principles of analogue simulation it would be as well to mention briefly some of the equipment

which is used. In principle an analogue simulator could be electrical, mechanical, hydraulic, etc., but we shall continue to restrict our discussion to the first type since analogue simulation is most commonly and conveniently electrical.

Simulation Equipment

The commonest electronic business aid at the moment is the digital data handling or computing machine, and most executives are aware of the size, cost and complexity of such machines. It might be useful to make a very rough general comparison between an electronic digital computer and an analogue simulator.

The cost of an electronic digital computer is of the order of many tens or a few hundreds of thousands of pounds, whereas an analogue simulator suitable for investigating management problems would cost a few thousands, or say ten to twenty thousand pounds for a very versatile machine. The simulator will be easier to maintain, easier usually to operate, and smaller in size. In general the results of simulation experiments will have an accuracy of about 1% whereas a digital calculation would probably be accurate to 0·01%. The digital computer will perform calculations at very high speeds whereas the analogue simulator is comparatively slow in operation, although in simulation this is not a disadvantage.

The basic unit of the analogue simulator is an electronic voltage amplifier of high gain. If a voltage V_1 is applied to the input, then at the output terminals a voltage $V_0 = -GV_1$ appears. G is a constant called the gain of the amplifier and the reversal of sign is a technical detail associated with the way in which the amplifier works. The amplifier will handle steady or fairly rapidly varying voltages. Such an amplifier is called a D.C. amplifier (direct coupled) and it is often drift corrected, i.e. it incorporates a mechanism to ensure that there is no spurious output voltage due to second-order effects in the amplifier. The number of such amplifiers in a machine is a guide to its versatility. A thirty-amplifier machine is a fairly large general-purpose model.

This basic amplifier is modified by connecting suitable resistances and condensers between input and output terminals so that the original relation $V_0 = -GV_1$ is modified. Suitable

choice of components enables a given amplifier to give an output which is : a given fraction of the input, the sum of several different input voltages, the integral of the input, etc.

These basic facilities of multiplication by a constant, addition, and integration are provided by amplifiers which are supplemented by other special devices such as multipliers, where the output is the product of two inputs, or function generators, where the output is related to the input by some arbitrary function which can be set up by the operator.

A number of amplifiers, multipliers, etc., together with associated power supplies and pen-recorder are usually gathered into a single piece of equipment with an easily accessible board where the required interconnections between the units can be made. The same machine may be used for many different simulations by making up the interconnections for any problem separately so that they can be plugged into the board as required.

Once the problem has been set up the controls available to the operator are exactly those which would exist in the real system. Indeed, in certain technical simulators, such as those used for training pilots or nuclear reactor operators, the simulation equipment is attached to the actual control system used in the aircraft or reactor.

For technical details of the equipment used in electronic analogue simulation, and for an account of how to build a simulator from computing " bricks " like amplifiers and multipliers, the reader is referred to standard textbooks such as *Electronic Analog Computers*, Korn and Korn (McGraw-Hill).

It remains now to see how analogue simulation may be used to assist in the field of business studies, under the two headings research and development, and training.

Research and Development

If a complex system is being studied in an attempt to formulate an expression describing its behaviour, then it is usually comparatively easy to invent from first principles a simple expression of just a few terms which is a plausible starting-point. It will be rare luck if this is the equation required. Painstaking modification of this original expression, with each slightly different form laboriously compared with the properties

of the system under study, usually leads to a final successful expression.

If a simulator is available, then the first expression may be set up and compared in a minute or so with the past behaviour of the real system. Different weight can now be given to one of the terms, simply by turning a knob, and in a minute the simulator record examined to see whether it is closer or farther than the original from reality.

Similarly extra terms or modifications of existing ones can be tried quickly to see whether they successfully shape the response of the simulator to the same form as the real system.

Simulation may again be used when we have proceeded beyond research to development and now wish to propose a course of action which will modify the response of the real system to something more profitable.

All ideas may be tried quickly on the simulator to see whether they show promise of " tailoring " the response to the required shape.

The choice of the idea to try in practice is in this way likely to be a much more informed decision than one taken in the confusion of committee argument.

Training

There are obvious advantages in using a simulator to replace a nuclear reactor when training reactor operators.

Less obvious are the advantages of simulation for training business executives to take good decisions. The diehards assert—and they are right—that only experience will produce the good executive. The principal advantage of a training simulator is that it can give a lifetime of experience in a week.

It might reasonably be said that decisions in simple systems hardly need specially trained executives, whereas complex systems lack a well-established law on which to base a simulator.

In training at Staff level, however, the object is seldom the production of an automaton, and the complete accuracy of the fundamental law is less important than versatility in the machine. The situations presented to the trainee, by real life or the simulator, will seldom demand a unique, determinable set of decisions. The object is rather the encouragement of confident, rational and flexible mental processes within very

broadly defined boundaries. For this the fast time-scale simulator is ideal.

The following references are suggested as further reading on the topic of simulation.

REFERENCES

[1] MEYER, HERBERT A. (editor) (1956). *Symposium on Monte Carlo Methods* (John Wiley & Sons, New York ; Chapman & Hall Ltd., London).

[2] WILKINSON, R. I. (1953). " Working Curves for Delayed Exponential Calls Served at Random ", *The Bell System Technical Journal*, Vol. 32, March, p. 361.

[3] KAHN, H., and MARSHALL, A. W. (1953). " Methods of Reducing Sample Size in Monte Carlo Computations ", *Journal of the Operations Research Society of America*, Vol. 1, No. 5, November, pp. 263–78.

[4] BOWMAN, E. H., and FETTER, R. B. (1957). *Analysis for Production Management* (Richard D. Irwin, Inc., Homewood, Illinois), pp. 304–9.

[5] CHURCHMAN, C. W., ACKOFF, R. L., and ARNOFF, E. L. (1957). *Introduction to Operations Research* (John Wiley & Sons, New York ; Chapman & Hall Ltd., London), pp. 498–9.

[6] MCCLOSKEY, J. F., and COPPINGER, J. M. (1956). *Operations Research for Management*, Vol. 2 (The Johns Hopkins Press), pp. 376–400.

[7] FISHER, R. A., and YATES, F. (1943). *Statistical Tables for Biological Agricultural and Medical Research* (Oliver & Boyd Ltd.).

[8] THE RAND CORPORATION (1955). *A Million Random Digits with 100,000 Normal Deviates* (The Free Press, Glencoe, Illinois).

[9] TOCHER, K. D. (1954). " Symposium on Monte Carlo Methods ", *Journal of the Royal Statistical Society*, Series B, Vol. 16, pp. 39–60.

Chapter VI

CONTROL SYSTEMS

Part I: Introduction

EVERY system is under some sort of control. The solar system, the economic system, the system called the internal combustion engine, and the physical system of the human body are all controlled. Each behaves in a stable fashion, operating within fairly narrow limits (in the body these are called " physiological " limits).

Governers have been fitted to engines, feedback connections to anti-aircraft guns and transfer machinery. Yet these machines could hardly be called responsive. Anyway, this kind of mechanism has only to be controlled over a narrow range of activities. The sort of industrial machine which has unpredictable elements in its behaviour, which can only be described probabilistically, is more difficult. Technologically, methods of control are satisfactory. They ensure a sound product. The trouble is that in order to get the sound product most of the uncertainty is concentrated in the system, where it is difficult to handle. This is the point where industrial control demands integration of various bits of plant to create a smoothly running factory.

To take a readily understood example : the Bessemer converter is not a deterministic, but a probabilistic, machine. To operate this converter in a steelworks with maximum control over the steelworks as a complex system, an arrangement might be considered where the converter emptied automatically into a moving conveyor of moulds, at precisely pre-determined timings. The uncertainty in the Bessemer system, however, would then become manifest in the product. The steel made in this way would certainly not be to the proper analysis very often. To ensure a good product, it is necessary to load the process's fundamental uncertainty on to the timing. It is this necessity which undermines the control of production flow throughout the steelworks.

The fluctuations in the melting cycle of a Bessemer converter are not very great compared with many other industrial processes. This case was quoted just because it is *not* macabre. Control of a whole factory becomes a difficult prospect on the strength of just such slight indeterminism as this.

This chapter is concerned with the control of such complex systems as factories. It should be clear that the very notion of " control " in this context is extremely difficult and involved. There is in addition to the probabilistic plant the quasi-social system constituted by the manning (including management) of the factory. A control system has to manipulate a mechanism of this intricate kind.

The most difficult feature of the whole job of creating a control system is that people will insist on taking up two attitudes from the beginning.

The first of these is that it is " given " what the system is that has to be controlled. It is said that a function of *management* is to define the system requiring control ; the job of O.R. is to do as it is told. It is certainly management's job to delineate the area or portion of the factory in which the control is required. But it is part of the O.R. study, and probably the major part, to work out exactly where the controls are necessary. Consider the job of creating a control system for a portion of a factory in which many raw materials are first prepared ; all are then brought together for a critical process ; from this process products fan out into a variety of finishing operations. This factory looks rather like an egg-timer passing sand through a narrow neck. One management group will say that, if the critical process at the neck is planned and controlled, all will be well. A second asserts that to start the right balance of materials off at the beginning will force material through the critical process at the maximum rate, whence it will flow out for finishing automatically. A third group says that if the control is exerted over despatch, the system will " suck through " the optimum balance of work. Operational Research ought not to be confronted with these sorts of prejudgment : it must find out how to control this system " on the job ". It must define the system itself.

The second fallacy takes one of two opposing forms. Firstly, there is the outlook which holds that what is required for a

control system is plenty of paperwork. The degree of control is sometimes thought to be proportional to the amount of form filling. This is a dangerous aphorism when wielded by a congenital bureaucrat. On the other hand it contains a germ of truth ; for the control exerted over a system is proportional to the logarithm of the amount of information in the system. The other form of this fallacy says that all will be well " if you trust the foremen ". It is the whole job of the O.R. attack to decide how much and what sort of information has to be in the system in order to obtain control ; where and when it appears, how fast it moves, how accurate it needs to be : all these are critical questions, and all are part of the operational research.

The science which has gradually evolved to deal with this topic is known as cybernetics. The word was first used in 1947 as a name for the study of the organisation, communication and control of a system ; it originated with Wiener's group in America.[1] The results of this kind of research are being used by O.R. in the study of factory control systems.[2]

In all this work, the O.R. scientists have been seeking methods of control which would, as it were, liven the organism to be controlled and make it responsive. They were trying to create the analogue of a man riding a horse : where the information systems of the two components of the system (the horse and the rider) are put into a very special relationship, so that one is almost an extension of the other. A similar argument can be advanced about the pilot of an aeroplane. He does not control his machine by the exercise of brute force ; he uses servo-mechanisms which amplify the movements of his hands and feet. The machine feeds back to him on dials information about its internal states to which the pilot responds. Remember, too, that an automatic pilot can be switched in to control an aircraft. The important fact is that there should be rich connectivity between the information systems in both the component parts, so that the states of one affect the states of the other.

The idea of this rich interconnection, and especially of what is called an error-controlled regulator operating on the very well known principle of negative feedback, is the biggest single cybernetic contribution to the Operational Research into con-

trol systems for industry. The idea is not a new one : but it is certainly not one which has been properly employed. A manager who reviews results at the end of the month, through his cost statement and his reports on the level of machine utilisation, rejections, and so on, and then holds a meeting to award praise and blame, is operating a feedback circuit. Something can be learnt from comparing the characteristics of this feedback with those in mechanical and biological models of control. The variety, feedback, speed, subtlety, effectiveness and many other facets of the feedback system repay careful thought.

It is small wonder the factories so controlled should not be highly adaptable to the changes from inside and outside the plant with which the company has to deal. Somehow these defects need eliminating. Then the breakdown of one piece of plant will be readily accommodated by the automatic adaptation of the factory's programme to the new state of affairs. The time might come when a factory would respond in advance to changing market conditions. For a suitable control system, deducing the coming state of the market from the changes in its order book, might adapt its programming of the plant suitably to prepare for the change. Such possibilities can no longer be discounted as belonging to the realm of science fiction. They are no more than pale imitations of quite simple biological mechanisms.

Part II: Forecasting with Feedback

The Machine-Job Time

A forward plan for a given department consists of order loads on its constituent machines. The appearance of an order on a given machine will be called a machine-job ; thus an order consists of at least one machine-job, usually several, and sometimes a very large number indeed—possibly many more than the total number of machines available, in the sense that an order may return repeatedly to the same machine. For instance, in the steel industry, it is very common for a machine carrying out work on the steel (such as a rolling mill or a forge or a drawing machine) to do its given job in a considerable number of " bites ". Between each bite the steel has to be annealed, or to undergo some other process.

If these machines are visualised as being listed on the vertical axis of a plan, on which the horizontal axis is time, each order may be plotted on the plan as a series of machine-jobs.

The "planning board" just described must be the basic component of many industrial control systems, in so far as a large number of proprietary boards can be purchased on which to display work in just this way. Unfortunately, no proprietary system of display supplies a means of forecasting the time for a machine-job. This is the basic brick from which the whole edifice is built, and it is essential to the effectiveness of the control system that an accurate forecast should be available. In heavy industry in particular, the virtual impossibility of providing these machine-job bricks cut to anything like the right lengths explains the litter of defunct planning boards in our workshops.

One of the plants in which the work described in this section was done is a large steelworks, some two miles in length. It makes steel by three different processes (open hearth, electric arc, and electric high frequency) ; it specialises in special alloy and stainless steels. Thus the number of different *kinds* of steel is very large indeed. Further, the number of possible shapes into which the steel is made is virtually infinite. It may be hot rolled into blooms, billets, slabs, bars, rods or sheets. It may be cold rolled into fine, narrow strip of high precision ; or cold drawn into wire of any diameter down to a thousandth of an inch. A large number of special products is also manufactured in the steelworks, especially springs of all kinds, from watch springs to laminated springs for road vehicles. The variety of machinery is enormous : even in a shop containing perhaps twenty cold working machines, no two will be identical. On the product side, however many different analyses of steel are listed, a new combination of the basic elements can be inserted at any time. However many combinations of width and thickness and length are listed, a new combination can be demanded at any time. Thus the task of providing adequate estimates of machine-job times is formidable ; even if one were to provide Work Study services on a quite unprecedented scale, for the purpose of measuring all these times, there would still be two big problems to overcome. Firstly, it would be virtually impossible to keep them

all up to date ; secondly, so many new orders are received that have never been manufactured before (although something fairly like them may have been made) that even the most exhaustive list of products would not be exhaustive enough.

" Models " of the Machine-Job Time

Some method is required for estimating the machine-job time for any job on any machine throughout the works. Reasons have already been given why this task cannot be attempted by empirical means ; some specially constructed measure is required. First of all obtain a first approximation. This will be done by constructing an idealised version of how long a job will take—the " objective " time. It should be very clearly understood that the objective time is *not* to be used in any practical planning work. The objective time is defined as the time the job ought to take ; and by " ought " is meant " in the best of all possible worlds ".

If ten factors are involved in a job, and each factor may take on one of ten states, then it is much easier to tabulate the one hundred relevant values of factors than to enumerate the 10^{10} possible final combinations. Thus, if the O.R. scientist can formulate a model for the process he is discussing, and if he can tabulate the factors which are to be applied to the model, then the vast and quite unmanageable numbers of possible machine jobs can be reduced to a few simple lists. Someone, supplied with the model and the appropriate tabulations, could then evaluate an " objective " machine-job time by setting up his model, choosing the factors from the tabulation that affected the job for which an estimate is required, and doing the arithmetic. Consider how this works in practice.

A great many steel-making processes are of the " throughput " type. That is to say, a piece of steel (whether it be an ingot of 10 tons, or a spool of, say, fine wire) has to be processed *through* a piece of plant (whether this be a rolling mill or a continuous furnace). In this kind of process, the fundamental factors in the time taken are the speed with which the material can be passed through the plant and the quantity of material. Select, for a given machine, the maximum speed at which it is possible to process a set of jobs. The only constraints admitted are those of plant engineering (namely, that

the equipment concerned is in fact physically capable of processing the material at this rate) and those of the relevant production science, in this case metallurgy (namely that the process must produce a sound product at this speed). The time the material takes to pass through the process once will then be a function of this speed and the cross-sectional area of the product together with its weight ; the resulting time must be multiplied by an integer if it has to pass through the process more than once, or by a fraction if it should be possible to pass more than one item of the product through the process at the same time. Providing allowance is made for acceleration and deceleration of the machine, this gives a measure of the amount of objective time this product will spend actually being put through the machine. Associated with this element of time, even in an idealised measure such as this, there must be further elements concerned with handling : mounting the product on the machine, setting the machine, testing the run, and removing the product at the end. If it is convenient (as it has been found to be) to assess this handling element in two parts (one part characteristic of the piece being processed, and the other characteristic of the " pass " through the machine) there must be added two auxiliary estimates to the original throughput component. The result of all this is the most elementary kind of all Operational Research models : a simple formula.

Let there be :

c = number of coils of steel
A = a cross-sectional area in square inches
p = number of times through a plant
v = speed in feet per minute
d = the density of the steel in pounds per cubic inch
w = total weight in pounds
h = handling time associated with each pass
H = handling time associated with each coil.

Then the objective time t_i for the machine job is given by :

$$t_i = chp + cH + pw/12dAv$$

This formula can be used for any throughput process. For example, for a coil of strip, width and thickness would be multiplied together to form A, while the steel's density would

also be known. Since it will be desired to produce a certain weight, tabulations of coil weights for various classes of product will have had to be prepared, and these are now consulted to give a value to c in the light of w. Schedules will also have been prepared from which appropriate values of p, v, h and H may be selected. The objective machine-job time can then be evaluated from the formula.

Even a model of this restricted kind requires much research to support it. The schedules that are required as factors with which to feed the model when it is desired to evaluate the objective time for a particular machine-job consume a great deal of effort. The O.R. technique most extensively invoked in their preparation is that of statistical analysis. It is necessary to establish a statistical account of shop practice, unless indeed the basis of that practice is already a schedule. In a traditional industry like steel it is unusual for a key factor to be supplied from a table. Usually, there is folk-lore about what to do, and it is this that has to be investigated and categorised. Hence regression studies, analysis of variance, and all kinds of significance testing come into free use. When the model and its supporting schedules are ready for action it is possible to calculate an objective machine-job time for any job on any machine, whether such a job has been done on such a machine before, or whether such a job has *ever* been attempted before. In addition, as a management by-product, the schedules which have been set up become available in the shop as formalised (and therefore teachable) alternatives to the " little black book " and " inherited skills " of aged employees.

There are difficulties that could be encountered by a scientist who attempted to apply this technique. In fact, the model must undergo some sophistication if all the processes of the steelworks are to be accounted for in an analogous way. In the first place, there is the more advanced version of the model already given, where complications in determining the schedule have led to ingenious modifications in the original throughput model.

Consider the rolling of stainless steel sheet, and the particular process known as " hot finishing ". In this case, an attempt to analyse the schedule of passes was unsuccessful by routine methods that had been established in other departments. It

was, however, clear that the number of times the sheet would pass through the rolls must be a function of the reduction from the starting to the final thickness, and the statistician concerned eventually discovered this function empirically in terms of a standard reduction of thickness as measured by a proportion r. Further, since in this process the length of material increased very greatly, it proved necessary to incorporate the geometric means of the gradually increasing length. Thus, the original simple formula was modified to read :

$$t_2 = h_s + (\log t_2 \quad \log t_1)\sqrt{l_1 l_2}/(v + hp) \log (1 - r)$$

This throughput model will account for a very large portion of steelworks' processes.

There are, however, other kinds of processes. For example, some annealing is carried out by batch processing : a given amount of steel is inserted in a pot, the pot is lowered into a furnace, cooked for a certain time and then withdrawn. To deal with such cases, a completely new sort of model was required. As previously, it is possible to break down the time consumed in making a cast of steel into component elements. In this case, these are : the time to charge the furnace, to melt the charge, to oxidise the bath, to reduce the bath, to tap the furnace, and to fettle the furnace after tapping. After these six episodes have been completed, the cycle may begin again. Thus, in a sense, a simple formula for the time taken to produce a cast of steel is the sum of six component factors. Difficulty arises because there is no idealised speed at which any of these processes ought to take place—especially those three that are chemical reactions. The O.R. approach in this case was again to make statistical studies of the six component processes. The analysis was made for various categories of steel and the resulting distributions were studied. Suppose, for one moment, that one of these six component times was independent of all the others, and that it was normally distributed. In this case, the objective time for that component could be construed as a point of the distribution on its lower side. If this assumption could be said to hold for each of the six component times, then the sum of these parameters would constitute the objective machine-job time for making a cast of steel of a certain analysis. This was in fact the basis of the

theory used. It had to be modified in practice because the conditions were not as simple as those suggested here. In particular, the distributions were not normal, but tended to be lognormal; secondly, there was a good deal of interaction between the component times.

In some further work of a similar kind in a second steelworks, these particular models were found inadequate, and another sort of technique had to be used. The difficulty arose because the shop concerned was engaged in highly diversified engineering; a new model was required for almost every job, and only a small proportion of the factors involved in the model could be categorised at all. Accordingly, a specifically cybernetic idea known as the "Black Box" was drawn on.

A Black Box is a jargon-word in cybernetics, used to refer to a situation too complex to be described in all its connections. This situation is said to be enclosed in a black box: its mode of working, that is to say, cannot be discovered. What can be observed are the inputs to the Box, and the outputs from it. For any state of input, the output may be observed; if the input state can be altered, then the changes in the output situation can be watched. To the extent that by this means a description of the Box's effective part in the scheme of things may be discovered, no problem remains—despite the undoubted fact that no one has ever seen inside the Black Box. Consider an extremely complex production set-up as a Black Box, accepting none the less that it is desired to apply the model-building O.R. technique for objective time to it.

Despite the "jobbing" character of the shop now discussed, one or two jobs carried out here were of a repetitive nature. The study began by examining these. Notionally, one of these jobs would be carried out in the same time whenever it was done. In practice, there was indeed a mean time of this kind, but it was statistically distributed. By a statistical analysis of this kind of work, it proved possible to obtain a basic statistical variance for a machine-job time in this shop. To the extent to which such a variance can be measured and validated in one or two cases, a measure of *maximum discrimination* has been created. It is meaningless to try to take account of a quantity which is below the threshold of one's power to discriminate. If, for instance, it were a demonstrable fact that a certain man

could not tell the difference between the temperatures of two rooms that were in fact five degrees apart, then it would be meaningless to instruct him to go into the cooler of two rooms having a temperature differential of three degrees.

Having recognised the measured variance of a so-called "repetitive" job as a measure of one's ability to discriminate time inside this shop, the next step is to calibrate time as a function of this value. This time scale is then calibrated for the greatest possible discrimination that is built into the set-up. A relatively small number of "times" results: namely, the total range of time taken between the shortest job done and the longest job done divided by the unit of discrimination.

Thus, according to this theory, the measuring rods at one's disposal are adequate for the recognition of this finite number of different machine-job times. To try and nominate any more would be meaningless. Therefore, although the orders on hand in this works constitute a Black Box for the O.R. analyst, there is in fact a set of descriptions of work being done that can be perfectly correlated with the number of units of time provided on the scale just calibrated. To catalogue jobs in this way is really a ranking process. Work is described so that it falls into a finite number of ranks, and so that the output of the system (as measured by keeping deliveries) is zero-deliveries in line with forecasts. Such an output from the Black Box would measure one-one correlations between the inputs of time units on the one hand and complexity units on the other. This method is clearly highly complicated, and its full comprehension requires some knowledge of cybernetics. However, enough may have been said to indicate that it is not impossible for an O.R. scientist to build a model of objective machine-job times—even when production is so complex and so varied that it is literally indescribable.

The Coefficient of Correction

It was emphasised from the beginning that the machine-job time was not an attempt to estimate what really happens on the plant, and therefore it could never be used for planning purposes. This concept of time, however, was constructed as a first approximation to the estimate really required.

At this stage it is possible to begin the analysis of current

production in terms of the approximation that can now be made. That is to say : if a particular job is selected from a particular machine this morning, there can be calculated, through our objective time model, the time it should take to produce. The shop floor records disclose how long it has *in fact* taken to produce. If a ratio of these two times is made, this gives a measure of the constructed times—for that particular job. In other words, if the application to the model of the details of this particular job results in an estimate of one hour, whereas the job which was actually done on a machine this morning took two hours, then the ratio will be 0·5. The machine is only half as effective as was thought—in this particular instance, at any rate. Note that this ratio should be less than unity, because the methods adopted in establishing parameters for the models (the schedules and so forth) have always erred on the side of behaviour " in the best of all possible worlds ". If this has been properly done, it will be most difficult for the model to produce an objective time which was actually longer than any given actual time for that job on a real machine.

The next move is to carry out an analysis of the behaviour of a machine in terms of this coefficient. A statistical sample of work must be taken, sufficiently long to embrace any cyclic variations which may exist in the rate of production. It is impossible to say that the sample must include all classes of work : by definition the control system under development must deal with many products in the future which have not been met in the past. However, it is possible to take a sample record of production which is, within the normal statistical meaning of the word, " representative ". The many pitfalls which attend the selection of a sample of this kind are well known to applied statisticians, and there is no reason to think that the problems of successful and valid sampling in this piece of research are either more or less difficult than is usual. In practice, the samples taken for this primary analysis have varied between three weeks and a year.

The technique of analysis involves the application of the machine time model to every one of the jobs in the sample. Since these constructive times are available together with the logged times of these same jobs when they were actually

produced, it is a simple matter to calculate a coefficient of correction for every job included in the sample. The next move is to analyse the complex distribution of correction factors covering the machine's whole behaviour. The difficulty remaining with this set of results considered as a descriptive tool is that it is almost certainly not recognisable as one of the standard statistical distributions.

If the setting of objective factors for the model of objective time has been done with uniform stringency for any one machine over the total range of its products, then distribution of sample coefficients may be expected to be clustered unimodally around a mean value. In cases where the real effort put into the job in both a technical and a labour sense is more or less uniform over all products, this is in practice found to be the case. It constitutes the simplest analytic situation encountered by this approach. In fact, however, the distribution although unimodal is not expected to be Gaussian. This is fairly obvious. The coefficient is not a free variable, but a ratio which has an upper bound of unity. The effect of such an upper bound on the distribution is well known : it undoubtedly has a skewing effect. This skewness may be largely removed by application of the inverse sine transformation. Thus any value of the coefficient p is converted to the angle θ, where :

$$\theta = \sin^{-1}\sqrt{p}$$

In practice, this transformation was always made to the coefficient before any further work was done, and all further references to the coefficient in this chapter will be implicitly taken to refer to the *transformed* factor. Thus, in the simplest case—that of the unimodal distribution—the transformed set of values is likely to yield a distribution which can be described as " normal ". This has the useful property that the behaviour of the machine described in this way can be set down in two statistics, namely, the mean and variance of the distribution of coefficients.

Needless to say, it is in practice much more likely that the complex factors affecting output effectively group the work done on any machine, with the result that the empirical distribution under study contains within it a number of discrete

populations which have been mixed together. The simplest case of this, in turn, is a bimodal distribution. In practice, again, the machine behaviour may be discretely spread over as many as six modes, whereupon the realisation that this is the case is not at all obvious to the eye when the distribution is viewed graphically. In any case, whether the distribution is bimodal or multimodal, and whether it is apparent or concealed that this is the case, more advanced statistical analysis has to be performed in order to cut the distribution up into its component discrete populations.

This stage of the work is classic applied statistics. Discussions with managers, technicians, foremen and workpeople will suggest to the O.R. worker a large number of possible reasons why the productive capacity of the machine should be higher for some categories of work than for others. Having made a collection of these possible criteria, the O.R. man must formulate likely hypotheses about the reasons for multimodality in his distribution, and test these statistically. Full-scale Analysis of Variance may be involved, more often it is possible to test hypotheses by simple significance methods.

At any rate, the outcome of this attack has been a set of descriptions about the plant and its behaviour in remarkably simple terms. In practice there are usually two or three discrete distributions for the adequate description of any one machine. Each of these distributions refers to a class of production which may be quite clearly defined in words ; and, of course, it is possible to *quantify* the definition of each of these classes in a fully statistical sense, simply by quoting the mean and variance of the typical distribution involved.

In this way, a detailed understanding and record of the way a complex department is behaving may be obtained and stated on a single page of paper. All the vast number of variables involved in manufacture, and the great number of quite unforeseeable risks which attend production, are subsumed within the O.R. model ; whether in the formula for objective time, or the statistical evaluation of the correction factor. The inevitable complexity which must be taken into account in describing a production situation is thus removed from the description itself, where it cannot be handled because of its magnitude, and put into the work of Operational Research,

where it is sorted out and accounted for by the most economic means available.

Now up to this point the analysis has been of an existing situation. The outcome is a set of descriptions of complex activity, for which the claim has already been advanced that it manages to take everything into account while being extremely concise. The technique involved has two distinct components : firstly there is the fairly coarse approximation from a formula or perhaps a statistical model : and, secondly there is the fine " tuning " of the estimate by means of the correction coefficient, which is a highly sophisticated measure. It will be noticed at once that the first component of this technique is invariant with time—unless vital changes in plant or the whole technique of production occur. Thus, as far as this component of the technique is concerned, the validity of the set of verbally and statistically defined production groups will remain constant despite fluctuations in the tone or morale of the shop. Trends will none the less be discovered which will affect the validity of this description as a whole, since better management in all its many forms should result in steadily improving performance.

According to the measure constructed, these changes will be detected in the *second* part of the measure, that is by changes in the behaviour of the correction coefficient. A given kind of change on the shop floor must affect the descriptions set up in terms of this measure. The reaction will be felt on one or more of the statistical groups, and will be felt as a change in the mean value of the statistical distribution, or of its variance. Hence, if the description of the plant in these terms is not only to *be* right but to *remain* right, a further technique is required by which to keep the means and variances of each of the statistical groups under surveillance. Fortunately, this can very easily be done.

The technique of statistical control is very well known, and there is no need to go into details of its operation here. Let it suffice to say that a statistic can be maintained economically by routine S.Q.C. procedures, undertaking to detect a change within some pre-ordained level of risk. The prerequisites are : it must be possible to define the population of which the statistic is a parameter ; it must be possible to sample this

population ; the distribution of the value to be controlled must be known, and (for the use of *standard* techniques) must be Gaussian. In the case of group correction coefficients, all these requirements can readily be met. It remains merely to set up a control chart for each of the groups, and to set up a routine sampling procedure by which to maintain the charts.

No special difficulties are likely to be met in establishing this kind of control. In the case study that has been considered in these pages many hundreds of control charts were required. Accordingly, special display cabinets were designed, each holding a display of eighteen charts. These were (and are) continuous graphs, on which the last twelve months' results are perpetually visible. " Massed batteries " of these displays are used by the control statistician to detect changes in the behaviour of the whole of a steelworks employing some 7,000 men. This system of control has two very great advantages.

Firstly, by these means the set of models used in estimating or forecasting machine-job times is kept perpetually up to date. There is never any fear that since the original study was made something might have changed, and therefore, the figures on which everything is planned are quite possibly meaningless. This is achieved, moreover, without setting up the expensive mechanism required to re-analyse production on a continuous basis. 5% sampling is quite sufficient to keep these control gauges sensitive. Of course, if a change is detected, or is even suspected (for example, a movement is indicated at a low level of significance), then the rate of sampling can be stepped up to examine this possible change more closely. In the actual case reported, routine sampling is on a four-weekly basis. But if in this period a sample point on one of the control charts crosses the 1-in-40 limit line, there is certainly no need to wait another month to see whether this trend is confirmed. Instead, daily sampling can be undertaken to see whether the chart concerned returns to its accustomed course (in which case the point detected can be ignored as an acceptable chance variation) or whether the first point will be followed by others even more unlikely. Once a genuine change is detected at some high degree of statistical significance, a re-analysis is undertaken. It is often possible to determine, by examination of the figures fed into the model and the way the model operates,

why the change has occurred. In any case, once a stable level is reached the parameters of the group concerned are altered in the records and the set of descriptions can be regarded as being kept up to date.

Secondly, and quite separately from the question of production planning, the system described provides a direct form of management control. This index is in fact an extremely advanced kind of measure, which takes a great deal of understanding. Managers often fail to understand how sensitive it is to the things in which they should be interested. It should not, however, require much rumination to note that *all* factors affecting industrial production have been reflected somewhere in this index. In the case history quoted, at any rate, the results of controlling this index are supplied to management regularly. They do constitute a form of " control by exception "—of a highly scientific kind. For instead of being confronted with a vast mass of data, the manager receives each month a very short statement on which are recorded details of the very few groups (4 or 5 usually) in his department whose parameters have moved or are suspected of moving. A statistical undertaking is given to him on the same sheet of paper that all the rest of his activities, that is those not mentioned on the report, are operating at normal levels within such chance fluctuations as can be tolerated.

Through this piece of Operational Research, control is exercised in these two distinct ways. And it is control of the type discussed in Part I—or a step towards it—the sensitive detection of change.

The Forecast and the Feedback

Forecasting in operational research implies estimating the *most likely* outcome of any situation, based on understanding of the nature of the process and scientific evaluation of historical facts. Thus, when people seek to contrast scientific method with experienced intuition, as if these two methods of estimating were quite dissimilar, they are somewhat misled. The scientist's job is to analyse and codify history, and thereby to acquire an artificial " experience " which can readily be tapped. If his models are bad and his analyses inept, he may be much less successful in his estimates than the old foreman whose

brain is doing these things for him unconsciously. But a good scientist can usually better the old foreman, because his tools are better.

In forecasting a machine-job time, for example, a foreman is cutting his way through a tangle of reminiscence in an attempt to recollect what happened on some other occasion when he was in fact undertaking something rather like the job he is now being asked to consider. He may or may not be lucky in this search of memory and feat of comparison. In situations as complex as those considered in this case history, at any rate, a careful experiment on the ability of departmental foremen to estimate machine-job times showed that they simply could not do this with anything like the accuracy required to make a plan built up of individual machine-job time bricks a practical possibility. The average forecast was heavily on the pessimistic side ; but, and this was much more important, the *spread* of these errors was extremely wide. Forecasts ranged from 10% of the actual to more than 300%. It is small wonder that a detailed planning system had never been operated in this environment. The O.R. scientist, however, using the sensitive and controlled measures described in this chapter, was able to institute methods of forecasting of a simple and extremely accurate kind.

In order to make a forecast, both components of the descriptive technique must be invoked. The key factors affecting a given job are fed into the model by which it is described, together with those relevant factors from the schedules compiled by O.R. means. From this computation the objective time may be derived. It is merely necessary to identify the statistical group to which this job belongs, to look up its current correction coefficient, and to apply this. An estimate has now been made which is based on the latest up-to-date information from the shop floor.

The accompanying diagram (Fig. 6.1) tries to give a clear representation of the way the forecast system works, and the various feedback circuits by which the whole system is stabilised as time unfolds. The extent to which this operational research has produced results of a basically cybernetic character should be appreciated in terms of the ideas put forward in Part I of this chapter. This is a fundamentally organic system :

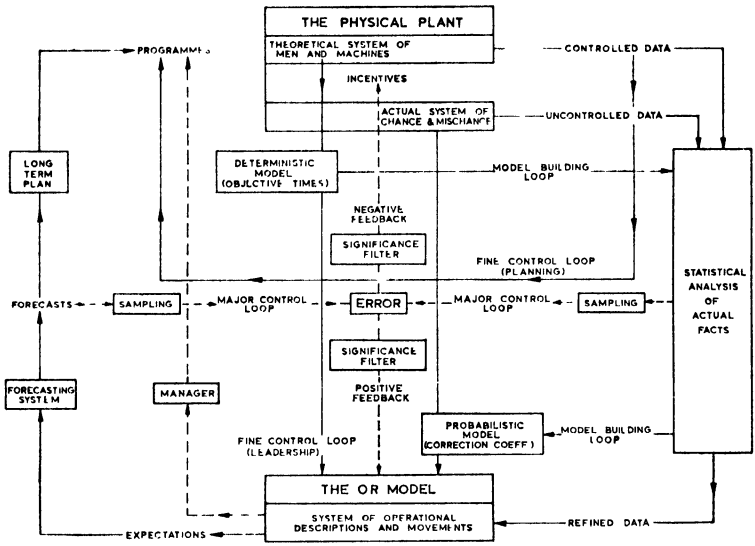


FIG. 6.1

one which is *responsive* at least to its own internal changes. The reader should not be misled, by the straightforwardness of some of the statistical techniques used, into thinking that the value of this work lies in its attack on the problem as involving questions (for example) of classification, or probability. The really important feature of the work is the way in which control procedures can be *vitalised.* The point becomes most clear in terms of results. A judgment of these will depend upon which effects are most required by the present state of affairs.

A self-stabilising system of this kind should aid any managerial action taken, to improve matters in any direction—whether to increase output, reduce stocks, or to improve some other feature of the system. But of many aspects of such improvements which could be isolated and quoted as examples of results, the key one is usually the aspect of deliveries. Any manufacturing concern is faced with the problem of forecasting its deliveries, and then keeping the promise to its customer which has been based upon that forecast. A control system which is characterised by the feature of "forecasting with feedback" is specifically designed to treat this problem in a cybernetic fashion. It is no longer a question of a fight be-

tween a scientifically unsound and remorselessly threatening forecast on the one hand, and an unlucky backsliding excuse-offering production on the other. Rather is it a collaboration between forecast and production jointly to arrive at a date which satisfies the customer. Careful perusal of the diagram will reveal how this comes about *automatically*, so long as the rules are adhered to.

In practice, the effectiveness of delivery promises has been increased beyond recognition in an industry where the exigencies of production were generally accepted as involving erratic deliveries. To quote an actual example : one department, which had for a long time been fulfilling only 27% of its orders on time, soon came to be delivering almost 90% on time. The remaining 10% were up to about 6 weeks late, compared with delays for the minority of over a year under the previous method of control.

Other benefits can be deduced as bound to accrue from the proper application of this approach, and they did accrue in practice. A reduction of stock levels at the input side of the system can be effected once the rates of consumption have been defined in probabilistic terms. A reduction of work in progress on the shop floor also follows better forecasting and planning. This in turn will almost inevitably increase production—by cutting down waste time and sorting out muddled supply lines. The main requirement, once the Operational Research has been done, is a managerial understanding sufficiently broad and clear-sighted to take advantage of these benefits and of the *organic* qualities of the resulting control system towards self-optimisation.

The sense in which the system described is self-optimising most certainly includes the controllers of the system as organic parts of it. Many people would probably prefer not to talk about the system's self-optimisation until the human part of the control has been replaced by automatic components (a prospect discussed in the next Part). But there is no good reason for this attitude. If the control system is designed so that a planner who is part of it is now presented with a clear indication that he should do A rather than B, instead of guessing between the two alternatives as heretofore, then the system itself has a self-optimising element. Admittedly the concept

requires, and is getting, a great deal more consideration than that. Basically, however, it is true that a control system which contains enough information to plan ahead so that agreement between the plan and actual events when they occur is better than chance agreement, is self-optimising. This point will be mentioned again at the end of this chapter.

Meanwhile it is vital to appreciate that this self-optimising quality *is* dependent on the human components of the control system (until they can be automated), and that they include the manager. Not only is his insight into the real nature of the system essential if management is to exploit the immediate advantages offered it, but it would be possible for a manager who fully comprehended the feedback mechanisms involved to exercise his control in a subtle and more far-reaching way. As described so far, and as operated in practice, this system faithfully reflects changing performance on the shop floor in terms of programming. It says: performance has fallen at this point by this amount ; sad but true ; in the next run allowances must be made for this new and depressed level of effort. Conversely, of course, less time will be allowed next run if the productivity this run is significantly high. The danger in this situation lies in the effect of a general slackening of activity (from whatever cause). Once a general downward trend in productivity begins, the control system becomes " machinery for running down and stopping the plant ", because it has positive feedback. In so far as the delivery promises throughout this process will be accurate and maintained, all is well ; but in so far as production is lowered, controls ought not actively to encourage the depression.

The answer to this difficulty is that falling productivity should be applied to the general plan as positive feedback, and to the departmental incentive schemes as negative feedback. The positive feedback ensures that plans and hard facts do not get out of line ; it is applied in quanta determined by the statistical significance adopted in the filtering part of the feedback system. The negative feedback would delicately encourage recovery in just those places and by just those amounts required to stop the falling away in effort. This could easily be done, *providing* that the incentive schemes were designed as a by-product of the productivity measures.

Unfortunately, this arrangement could not be reached in the case of this history quoted. The use of these general methods of estimating values, together with their feedback controls, has, however, been adapted successfully to problems of costing.

Part III: Further Developments

The Speed of Response

In the O.R. context of the steelworks hitherto described, a systematic attempt is being made to endow the control circuits described with responses a great deal more speedy than those attainable by present hand methods. Since it would be a gigantic task to attempt the full mechanisation of the whole control system in any one department, this stage of development is divided among a number of separate production units. In one, for example, methods are being studied of taking the necessary data from the plant directly from the electrical systems of the production units concerned. Thus the starting and stopping of a mill, for instance, would be duly recorded. Once a history of this type has been successfully recorded on punched paper tape, it is obviously possible to feed the information directly into an electronic computer. There are, however, some processes which may be difficult to record automatically in unambiguous terms, and the cost of doing this exhaustively throughout a factory would clearly be rather high. However, enough has been done in several different steelworks to demonstrate that the kind of record required in such a control system is certainly obtainable.

Secondly, research is going on into the problem of calculating correction coefficients for any given Operational Research model by computer methods. It is now clear that this work can be done on either a standard digital computer, or on a special-purpose electronic analogue machine—providing sufficient effort and money is spent on obtaining (or avoiding the necessity for) unusual accuracy in the potentiometers. This piece of research is concerned with the statistical analyses, and the storing and modification of the system.

Next comes the question of the major feedback loop itself. An "importance meter" has been proposed, the object of

which is simply described as exercising statistical control automatically. If one imagines 100% of certain data as the input to a machine whose output is that 5% of the data which exceeds the two-sigma limits of the population, where the machine itself is measuring the standard deviation of the data and computing its own confidence limits, the concept of the importance meter will be sufficiently understood. This project seems likely to succeed, and a piece of *ad hoc* machinery should not be unduly expensive. It is perfectly clear, of course, that the operation could be integrated with other operations on a standard digital computer; the point to be borne in mind is, however, that the use of a computer in these many ways could soon absorb its whole capacity. There is little doubt that a fully responsive cybernetic control of the factory would indeed justify the use of a whole computer; but this assurance does not diminish the need to examine alternative methods on cheap, special-purpose machines.

Finally the major effort at this stage is being concentrated in an attack on the ubiquitous problem of transmitting a programme to the shop floor. There is no point in having a high-grade information system if changes in policy, whose effects could be developed into programme modifications, cannot be implemented. At present, the indications are that a feedback control forecasting system could result in programmes modified up to the minute; yet typical steelworks practice, in which the operators themselves are instructed by a mixture of written and oral orders, may well mean that a desired change cannot be put into effect for several hours—or indeed several shifts—unless an emergency or even a crisis is declared. Consequently, two pieces of the research devoted to developing this work have been concerned with the automatic transmission of programmes.

Of these developments the major one is concerned with a new Bar and Rod Mill. Maximum production demands a rapid and efficient system for monitoring progress and for communicating with the operators. The mill is of an extremely modern, high-speed type; yet relatively small batch production will be undertaken. Operational Research studies have demonstrated that in these circumstances there is only one conceivable means of maintaining a high level of produc-

tivity: an abnormally high level of responsiveness will be required, and the benefits of this can be obtained only if information can travel round the mill at very high speeds. The scheme being installed consequently compiles its programmes on paper tapes. From them, any aspect or detail of bar and rod production progress can be extracted for office use, or fed back to earlier production departments automatically. Conservation of information outside the department is guaranteed by the possibility of using the tape to operate electric typewriters, or to be converted mechanically into punched cards, or used as the input for an electronic computer which it is recommended should be installed.

Internally to the department, details of the material coming through the mill are transmitted instantaneously from the tape programmes to visual displays at all the critical points in the mill where the operators can read all the relevant information about the batch at present being rolled and the one to follow it. At other points, such as the despatch bay, more details are required about the order and its destination, and they are required in a more permanent form. Consequently, the same impulses are used to operate a teleprinter mechanism, which types the required information on pre-printed stationery.

This is, of course, the briefest possible sketch of an expensive scheme already in advanced stages of preparation. But it does indicate the direction being taken by the research workers involved in this extensive case history, and the length to which Operational Research must be prepared to go if the problems it is set are to be truly solved and not merely used as sophisticated exercises which create still more problems for other specialists. In this example, some might consider that the work was being followed into a domain proper to the communications engineer. So it is: the point is that O.R. itself has delineated the true extent of the problem, and solved it in principle by a logical study of the information situation, and is now uniquely placed to see the job through to a successful conclusion—dealing simultaneously with its ramification for Operational Research in such concomitant problems as stock control and optimum batching. The proper way in which to deal with the specialised engineering aspects of this work is to add suitable specialists to the Operational Research

team engaged on the job : a course which has in fact been followed.

Feedback without Forecasting

The major part of this chapter has occupied itself with what was called " forecasting and feedback ". In introducing that section, one important question was carefully and deliberately begged. It was assumed that production control was to involve forward planning of a detailed order book through a detailed machine load. In the work which was then being described, a decision had been taken (for very good reasons) to attempt planning in this way. Therefore, it was necessary to operate a forecasting system—with feedback. Suppose, however, that a form of control is being exercised in which this detailed machine-job loading is not required ; Operational Research is still likely to cling to its cybernetic model. The *essential* feature of the system is not the detailed forecasting mechanism but its self-regulating feedback mechanism which works towards stability. Five different approaches to this kind of control will be mentioned briefly.

Firstly, one of the control problems was solved entirely in terms of organisation, once the concept of feedback was invoked in the design of the organisation. The problem was concerned with the ability of a large section of plant to satisfy its customers according to an allocation system that had been worked out and agreed with them. The total product to be delivered in any fairly lengthy period could in fact be produced ; no one could really understand how it was that only about 35% of the customers concerned were at any given time " up to allocation ". Since the total output was fairly steady, it was obvious that this lucky minority was receiving very much more than a fair share. This much was known, but all the attempts to rectify the position failed. Inquests always revealed excellent reasons why the unduly favoured customer had received so much more than his allocation : the issue turned usually on the factors of economic batch size, and the disposal of steel which did not come out to individual specification. O.R. study revealed means of handling the problem. Existing methods of planning were examined as a whole control system.

It was discovered that over the years a highly diversified system had grown up, in which 14 different sections, responsible to 8 different senior managers, were involved in controlling the production of this plant. A study of the information flow was fascinating: roughly speaking, each of these sections was instructing the other 13. The Operational Research answer to this problem, which was implemented, notably involved the abolition of all these sections and their re-creation as part of a new organic system under one head and in one place. The information in this system was designed to be highly regulated. The result of this O.R. study was that within six months of the organisational changes being made (including, that is, the lengthy period of "settling down") the percentage of customers in receipt of their full allocation had more than doubled.

The remaining four techniques involve *partial* measurement of the system, but seek to avoid the wholesale quantification undertaken in Part II. These are now introduced briefly under the title of the feature of the system for which measurement is in fact involved.

(i) **Controlling by Measured Importance.** Various attempts have been made within the framework of this history to exert control by measuring and ordering only the most important features of the system. It seems that sometimes this can be done by the detailed planning of only such machines or groups of machines as appear to have a key influence in the system. This has been done, for example, on a large mill where the influence of a number of factors is measured by the really serious effect of each on the mill's production, and their relative importance sorted out by the technique of Analysis of Variance. Little use has in fact been made of this idea, however, in the cases being discussed because of the obvious risk that the characteristics of the system would change with time. There is a particular danger of this in the steel industry, where it is not always clear which process of several is *the* process—to which the others are ancillary. One tends to view the most heavily loaded process as the critical one; in which case a feature of today's production may be regarded as purely ancillary and will next week become the key factor over which control ought to be exerted.

(ii) **Controlling by the Level of Demand.** This approach, which has been tried out most successfully, required the abandoning of a traditional policy in certain departments. It had been accepted previously that all work had to be done against a specific order : a fairly lengthy period was involved in production, and the cycle could not even be started until the order had been received, for fear of holding too large a stock of finished or semi-finished goods. This policy was abolished on the advice of O.R., and production was allowed to flow down several major " pipelines ", which were envisaged as delivering a smooth flow of goods in the general direction of the market. The feature of the whole system measured under this arrangement is the *continuous* level of demand. The receipt of an order involves " opening the tap " of one of the pipelines and drawing off enough material to meet the order. If this tap also operates a valve at the beginning of the system which allows an equivalent amount of material to enter the pipeline from the source (the previous department), then the pipeline would be regarded as perpetually full.

In practice, of course, the shortcoming of such a system tends to be that the level of demand is unsteady : a direct feedback would make erratic, irritating and uneconomic small and sporadic calls on the supply. Thus the valve under consideration needs to be more than a simulator of the tap ; it must include a statistical " smoothing " device. The pipeline control scheme does in fact work on this basis, using a statistical function of the continuous level of demand. Where it has been used, it has provided a most sensitive self-regulating feedback for the system and has of course—inevitably—reduced the work-in-process stocks.

In fact, in so far as it is essential in any case to have material in production before it can be despatched, there are no stocks in any orthodox sense at all. On the other hand, if the pipeline is itself called " a stock " then it is by definition inexhaustible : there is the infinitely long queue which is theoretically demanded by Queue Theory in a situation of this kind.

(iii) **Controlling by the Cost of Queues.** Another new development seeks to measure another particular aspect of the system rather than its whole detailed structure. This time, the O.R. group has turned its attention to the waiting time of

work which has not only to be produced but assembled in the course of its production. It is not so much the question of machine jobs which requires O.R. attention : the time an order spends actually on machines is likely to be a very small proportion of the total production time. What is now important, both in terms of meeting delivery promises and in terms of the capital invested in work-in-progress, is the queue waiting in front of each machine.

In one of the current jobs with which this O.R. group is concerned, measuring is being concentrated at this important point, and the control of the system will be sought through a control of the cost of queues. The O.R. techniques in use are naturally based on Queue Theory, but this has been extended in an interesting way to connect with the theory of costing which underlies the economics of the company concerned. Again, however, there is this common feature of partial quantification coupled with feedback. In this case, the feedback is obtained by the pressure of rising costs which operates a new queue discipline to reduce cost again. Thus, as ever, is met the self-regulating function of a well-designed control system.

(iv) **Controlling by Measured Priority.** Finally, in yet another plant, there is another new development based this time on measuring the importance of the customer, and the extent to which he is being satisfied. In this application, also on the verge of installation, waiting time in the steelworks is of more relevance to the control problem than the time spent on machines. Again, also, there is a special pressure exerted which tends to keep the plant occupied : this time it is not a logical pressure but a thermal one. That is, the project concerns a forge, where the conservation of heat is vital both economically and metallurgically. It is hoped to bring production in this plant under control by a system of priority numbers which do measure the characteristics of the order in which the company—as seller—is interested.

The numbers are subject to continuous modification so that they do at all times reflect the *importance* of the job next to be done for any particular order. Thus a simple rule is supplied, namely : always try and do the job with the lowest number next. The investigators concerned with this piece of Operational

Research have been surprised to discover how profoundly effective such a simple rule can be. The behaviour of these priority numbers has been simulated through a number of experiments, and the required level of control has been observed to exist.

Conclusion

Part III of this chapter has been divided into two main sections, the first dealing with the speed of response in a highly involved and fully quantified control system, and the second dealing with a rather more subtle but at times adequate form of control through partial quantification. This division has been deliberate, in so far as these are two lines of development which must both be pursued by Operational Research as applied to control systems. Both lines of development are drawn from the analogy of the cybernetic model, for both recognise two necessary and sufficient criteria of control : self-regulation within limits, and what may be loosely called " stability ".

For the moment, however, these are considered as objects of Operational Research, and in this role it would be appropriate to sum up each line of development in terms of the problems with which it still seems to leave us. There is reason to be suspicious of any Operational Research which claims to have disposed finally of any question.

Firstly, then, it is possible that the first line of development will involve the sort of advanced bureaucracy in which the system " seizes up " because of an excess of *unwanted* feedback—or auto-correlation—in the system. It is extremely difficult to work out the conditions under which such an elaborate type of control mechanism will go into oscillation. Again, cost has to be watched : the systems of this type so far sanctioned are certainly expensive. Thus, where this mode of thinking is concerned, there is a note of warning, and a feeling that the pursuit of control for its own sake up to the last crossed *t* and dotted *i* may simply not pay for itself. This feeling of caution is, however, almost totally offset by the contrary proposition that the maximum result can only be obtained from any system when it is under complete control.

Now this argument really resolves itself into the question

whether an attempt is being made to control industrial processes to the point where they are no longer probabilistic in their fundamental character, but deterministic. Is the idea to make them behave like clockwork, or would we recognise that they are basically exposed to the vagaries of statistical chance? The answer to this varies from industry to industry and from time to time. A fully automated plant is by definition deterministic: if statistical mischance enters into its processes, then there is no option but to say that it has broken down. There are, on the other hand, industries such as dairy farming in which it is impossible to conceive that chance will not always play a major part. In between, there are industries such as that with which these investigations have been concerned: few steel men really believe that their processes will ever be deterministic; on the other hand, the steady development of their technology has already slashed the number of situations in which luck produces adverse effects from one chance variable after another. Thus the nearer that any Operational Research man gets towards saying that the system he is studying is moving towards the deterministic state, the more he will tend to see his problem in terms of the first line of development proposed here. The more specifically unpredictable his system, on the other hand, the more he will tend to rely upon the mass effects of *statistical* predictability, and to seek his controls in feedback of a less highly anticipatory kind.

For convenience the first system was described as tending to determinism, and the second to probabilism; for convenience, too, the respective techniques were described as applying feedback to detailed forecasting, and as displaying feedback with only partial quantification. Now the cybernetic model *looks* more like the second type of system. Time unfolds irreversibly, and the pattern of tomorrow's events is not yet settled. Therefore, the argument would run, there is little to gain by detailed forecasting: this is in a basic philosophical sense impossible. Feedback control towards a generalised objective, a partially quantified optimum, will make the organism responsive, and turn it into an opportunist able to exploit events as they unfold.

The counter-argument to this is that man turns out to be

the superior of all animals because of his ability to synthesise experience into a forecast. This kind of forecast is not an attempt to look into the future, but an attempt to build a model of the environment and to operate it on the basis of present trends. In thus simulating a future, man—unlike other animals—becomes capable of influencing his own future scientifically.

He is able to organise matters so that likely events (the high probabilities) are allowed for in his plans and unlikely ones excluded from them. In this way his strategy has a much better pay-off than if he were to await events, and then wriggle.

REFERENCE

[1] WIENER, N. (1949). *Cybernetics* (New York, John Wiley & Sons Inc.).

[2] BEER, STAFFORD (1959). *Cybernetics and Management* (London, E.U.P. ; New York, John Wiley & Sons Inc.).

Chapter VII

LINEAR PROGRAMMING

IT is often necessary for a manager with a decision to make to have to choose between so many alternatives that he can never be absolutely certain the solution which he adopts is the best. It is usual for him to decide a particular planning or production policy with regard to some monetary function which has either to be minimised (in the case of cost) or maximised (in the case of profit).

In a certain class of decision problems there exists a technique which can be relied upon to give this maximum or minimum. This technique is known as linear programming.

As examples of the types of decision which have to be made by management, consider a firm which manufactures three qualities of pencils at four factories from the same basic raw materials. If the three qualities are made on different production lines there is no difficulty in deciding how many of each pencil to make. Should there be an unlimited supply of raw materials, then the production level can be set independently of the amount of wood or graphite that is used. If the firm's customers can be supplied with their requirements from their nearest factories then there are no distribution problems.

However, it often happens that none of these ideal situations occurs. For example, all the wood for the pencils might have to be shaped on the same machine and then the level of production of each quality will have to be set so that the capacity of this machine is not exceeded. There may be some restrictions on the amount of graphite available, so this will have to be divided so as to produce the quantities giving the greatest profit. Two customers' demands might exceed the capacity of the factory which is nearest to them and one must be supplied from a more distant factory at an increased cost.

These three decisions have one thing in common. In each case it is required to divide a number of limited resources

amongst a set of conflicting demands. The demands are connected, since if one is satisfied in a way a second cannot be satisfied in the same way, because of the total limitations. The aim of management might be to minimise cost or stock level or maximise profit or production whilst satisfying the limitations. For small-scale problems this planning can be done by experienced staff inspecting the facts and making decisions on them, but as the problem increases in size it is not usually possible to see how to obtain the best policy.

The technique of linear programming has been designed to aid in the making of decisions of this nature. It is a routine calculation procedure which enables the best solution to a problem to be obtained in a systematic way. There is no complicated mathematics in the method used (although there is in the proof of the method) and it can easily be taught to clerical staff.

Suitability of Problem

The main requirement for adopting this tool is that some function has to be minimised or maximised. The function will involve various proportions of certain variables (resources used) corresponding to (perhaps) the quantities of the final products made, the amount of raw material used, the size of stock at a particular place, the quantity of product sent by a particular route—or many other possible variables. The values of these variables will be controlled by other functions (which are known as restrictions). These might limit the amount of time available on certain machines and hence the quantity of final product manufactured, the tonnage of raw material available or required, the total investment, the number of men available and so on. Both the function which has to be maximised or minimised and the restrictions must be linear—in other words the profit from 100 articles is twice the profit from 50, or if 60 units can be made in 1 hour, 300 can be made in 5 hours.

Derivation of a Problem

As an illustration of the derivation of a problem, consider a firm which makes three grades of tyres, *A*, *B* and *C*. The firm uses two types of rubber, *L* and *K*, and each tyre has to

go through two operations, T and S. The profits per tyre are P_A, P_B, P_C respectively. It is required to obtain maximum profit. Assume x_A, x_B and x_C tyres are made of grades A, B and C. Profit from making x_A tyres of type $A = x_A P_A$ and similarly for tyres B and C.

Therefore total profit $= x_A P_A + x_B P_B + x_C P_C$. . (1)

Suppose tyre A requires l_A lb. of rubber, L, and k_A lb. of rubber, K;

tyre B l_B and k_B and tyre C l_C and k_C, then

total requirement of rubber $L = x_A l_A + x_B l_B + x_C l_C$

and total requirement of rubber $K = x_A k_A + x_B k_B + x_C k_C$

But if there is at most L' lb. of rubber L, and K' lb. of rubber K available the total requirements must not exceed the tonnages available and hence

$$x_A l_A + x_B l_B + x_C l_C \leqslant L' \quad . \quad . \quad . \quad (2)$$

and

$$x_A k_A + x_B k_B + x_C k_C \leqslant K' \quad . \quad . \quad . \quad (3)$$

In a similar way if there are t_A, t_B, t_C minutes required of operation T and s_A, s_B, s_C required of operation S and only T' and S' minutes available this gives

$$x_A s_A + x_B s_B + x_C s_C \leqslant S' \quad . \quad . \quad . \quad (4)$$

$$x_A t_A + x_B t_B + x_C t_C \leqslant T' \quad . \quad . \quad . \quad (5)$$

It is obvious that x_A, x_B and x_C cannot be negative

so

$$x_A, x_B, x_C \geqslant 0 \quad . \quad . \quad . \quad . \quad (6)$$

The problem is now in linear programming form and is to maximise (1) subject to the restrictions (2), (3), (4), (5) and (6).

Method of Solution

Linear programming problems can be divided into two main groups which have slightly different methods of calculation for each. The methods depend on the same general principle. In each case values are given to the variables which satisfy the restrictions, but do not necessarily give the optimum (maximum or minimum) to the objective (e.g. profit or cost) function. The values of the variables are allocated according to certain criteria and a " feasible solution " is said to have

been obtained. The feasible solution is tested by given rules to see if it is an optimum. If it is not, the feasible solution is altered in a systematic manner to give a better result, which is tested again. At some stage the feasible solution obtained will be the optimum and the problem will have been solved.

The two methods are called " transportation " and " simplex ". The latter will solve all linear programming problems, but entails more work than the specialised case of the former. The details of the computational procedures for the two methods are given later. The example of the previous section requires solution by the simplex method, whilst the first problem described in the next section on applications is solved by the transportation method.

Applications

The two examples which are now given illustrate actual practical applications of the technique and it is hoped will help in the greater understanding of the variety of problems amenable to solution by these methods.

Transport of Coal

In Great Britain the Central Electricity Generating Board uses about 40 million tons of coal a year. This coal has to be transported from some 800 collieries to 200 power stations and is carried by rail, sea, canal and road. At six-monthly intervals a plan is drawn up to show which collieries supply which power station. The coal cannot always be supplied from the nearest colliery and often has to be sent large distances across the country because the coalfields are concentrated in the North. Since the C.E.G.B. pays the cost of moving the coals, attempts have been made to calculate the allocation of coal which would give the minimum total transport cost.

Many complications arise, and perhaps the most important is that of size. It is almost impossible to solve a problem of this size because of the excessive amount of arithmetic involved. As a first stage the problem was solved for the various electricity divisions. The next stage will be to consider the country as a whole and then to assess the effects of the alterations on the efficiency of the stations and finally the production of electricity at least total cost. These problems are being con-

sidered by the industry. The size of the divisions varies from 5 power stations supplied by 20 collieries to 33 power stations supplied by 103 collieries.

The following table (Fig. 7.1) gives an illustration of the smallest divisional problem. The figures in the right-hand corners of the square are the transport costs per ton of supplying from each colliery to each power station and the other figures give the allocation of coal at minimum total transport cost. The row and column values are information which has been calculated and will be explained in the section on " Method of Calculations ". It is not necessary to know all the transport costs though the more that are known the more accurate the answer will be.

Mixing Problem

As an example of the application of linear programming in another type of problem, take the work by Duckworth, Kay and Morton at the Glacier Metal Company Ltd.

The company creates scrap metal in the form of turnings left over from machining bearings from white metal castings. The scrap metal contains a large proportion of tin and can be either melted down with virgin metal for re-use or sold to other firms who recover the metals. A number of factors such as percentage of impurities and cost of storage affect the decision of which alternative to use. Linear programming was applied to the problem and the results enabled the best procedure to deal with the fluctuations on impurities and prices to be found. As well as reduced material costs, the direct and interest charges of storing scrap were reduced.

Calculations were made for a trial survey of 10 weeks to compare the results of the linear programme with the previously used methods. The results showed that for every 1,000 lb. demand of white metal the usable recovery material was 462 lb., of which the old methods used 424 lb. and the linear programming method 446 lb. The average length of time in stock in this 10-week period was reduced from 7·5 days to 4·3 days.

The savings have outweighed the cost of calculation, since a trained computer takes less than 2 hours a day to perform the calculation.

Power Station	Pst	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Total
		14	73	38	62	55	87	112	56	50	54	55	74	92	39	91	90	46	55	98	69	
a	0		1200		420				1660				1430			1440						6150
		34	73	39	62	56	88	114	56	60	36	56	74	94	58	91	94	65	74	98	74	
b	÷ 9	1000							340						2010			550	380			4280
		23	70	30	54	46	79	104	65	42	46	47	65	87	48	84	89	55	64	95	66	
c	1			570		630	490		550					1340								3580
		34	74	39	65	56	88	114	57	52	56	58	78	93	59	94	94	65	75	98	71	
d	−8							850		460		1550				260					110	3230
		28	68	33	58	49	80	104	48	42	47	47	69	88	49	83	84	76	67	90	61	
e	−16				1280						1210						480			990		3960
		19	61	27	46	42	75	100	41	36	38	44	63	86	44	78	74	50	60	82	54	
Total		1000	1200	570	1700	630	490	850	2550	460	1210	1550	1430	1340	2010	1700	480	550	380	990	110	21200

FIG. 7.1

Transportation Technique

The Transportation Technique can be used to minimise the cost of distributing goods from a number (m) of despatch points to a number (n) of receiving points.

The following conditions must be satisfied :

1. The amount of goods to be sent from each point and received at each point are known.
2. The total amount of goods despatched equals the total amount received.
3. The cost of transporting each unit of goods from each despatch point to each receiving point is known.

A basic feasible solution satisfies the capacities and requirements of the despatch and receiving points by the use of $m + n - 1$ routes.

Example

There are three collieries which can supply coal in the following quantities :

Colliery	a	14	units	
,,	b	12	,,	(A unit here is 100 tons and
,,	c	5	,,	is used to reduce the ciphering)
	Total	31	,,	

There are three consumers at which these coals are required in the following quantities :

Consumer	A	6	units
,,	B	10	,,
,,	C	15	,,
	Total	31	,,

The costs for transporting one unit of coal from any one point a, b or c to any other point A, B or C is given in the following table (in £'s) :

	a	b	c	
A	6	7	4	(Cost is actual divided by
B	4	9	3	100 to reduce ciphering)
C	1	2	6	

The first step in the solution is to fill in on a blank chart an allocation of the coal which is a basic feasible solution.

Step 1

One systematic method is to start at the top left-hand corner of the chart (Figs. 7.2 and 7.3) and work from there allocating

	a	*b*	*c*	Total
A	6	7	4	6
B	4	9	3	10
C	1	2	6	15
Total	14	12	5	31

FIG. 7.2.—Basic data as to supply, demand and transport costs. Senders *a*, *b*, *c* supply 14, 12, 5 units respectively. Receivers *A*, *B*, *C* take 6, 10, 15 units respectively. The figures in the bottom right-hand corners of the cells are transport costs.

	a	*b*	*c*	Total
A	6 6	7	4	6
B	8 4	2 9	3	10
C	1	10 2	5 6	15
Total	14	12	5	31

FIG. 7.2*a*.—Step 1. Basic feasible solution obtained by N.W. corner rule $m + n - 1 = 5$ routes in use. Total cost=£136(00).

as much as possible by each route. This is called the North-West corner rule. The total cost of this allocation is £136.

Step 2

It is now assumed, in order to facilitate descriptions of the calculations, that the cost of transporting is made up of two parts, namely, a cost of despatch and a cost of reception. It is assumed (in Fig. 7.3) that one of the costs of reception is zero, say at *A* ; then, according to the cost in cell *Aa* (i.e. 6),

the cost of sending from a must be 6. The cost in cell Ba is 4 and thus the cost of reception at B is -2. This procedure is followed for cells in which there is an allocation so that all three costs of sending and the three costs of receiving are

		a	b	c	Total
		6	11	15	
A	0	$6-\theta$		$+\theta$	6
		6	[4] 7	[11] 4	
B	-2	$8+\theta$	$2-\theta$		10
		4	9	[10] 3	
C	-9		$10+\theta$	$5-\theta$	15
		1	2	6	
Total		14	12	5	31

FIG. 7.3.—Fill in fictitious costs of sending and receiving, starting at $A = 0$ and using cells with allocations only. Step 2. [11] indicates sum of part costs exceeds actual cost by £11 in square Ac if a quantity θ is sent by route Ac, i.e. total cost can be reduced by £$11 \times \theta$. Other changes required are shown. Ac has largest figure in square. $\theta = 2$ tons, biggest value possible see square Bb. Total cost = £114(00).

found. The unused cells can now be considered and, where the sum of receiving and sending costs for any unused cell exceeds the true cost which is marked within the cell, there is a saving in cost to be made by using this so far unused cell. The next step is to transfer goods into the unused cell which will give the largest saving (i.e. cell Ac in Fig. 7.4).

		a	b	c	Total
		6	0	4	
A	0	$4-\theta$		$2+\theta$	6
		6	7	4	
B	-2	10			10
		4	9	3	
C	2	$+\theta$	12	$3-\theta$	15
		[7] 1	2	6	
Total		14	12	5	31

FIG. 7.4.—Step 3. Repeat Step 2, but with $\theta = 3$ tons. In this case only Ca has a number in a square. Total cost = £93.

The goods to be transferred to cell Ac must come from either cell Aa or cell Cc. If θ units are transferred from Cc to Ac, then Aa must be decreased by θ units to keep the row total correct. Similarly, to maintain the correct row and column

totals, cells *Ba* and *Cb* must be increased by θ and cell *Bb* must be reduced by θ. The maximum value of θ, i.e. 2 units (determined by cell *Bb*), is entered into the solution. The resulting solution has a total cost of £114.

Step 3, etc.

The part costs must be adjusted and Step 2 repeated.

Successive steps are carried out until all the real costs in the cells either equal or exceed the sum of the part costs for sending and receiving. When this state is reached the optimum solution has been achieved. (Step 4, Fig. 7.5.)

		a	*b*	*c*	Total
		6	7	4	
A	0	1 6	7	5 4	6
B	−2	10 4	9	3	10
C	−5	3 1	12 2	6	15
Total		14	12	5	31

FIG. 7.5.—Step 4—shows no further improvement possible. (No numbers in squares.) Minimum cost £93.

It is essential that only one unused cell which can show a cost saving is dealt with at a time.

As an alternative to the North-West corner rule the first basic feasible solution can be constructed by commencing the allocation at the cell with the lowest cost and continuing to the cell with the next lowest cost, etc. A better first solution (i.e. lower cost) is generally obtained in comparison with the solution obtained with the North-West corner rule. This is illustrated in Fig. 7.6.

Degeneracy

If at any stage of the computation the allocation uses less than $m + n - 1$ routes the problem is said to be degenerate. With only, for example, $m + n - 2$ routes in use it is not possible to compute all the part costs.

The difficulty is overcome by increasing the requirements of

	a	*b*	*c*	Total
A	6	7	4	6
B	5	5		10
C	14	1		15
	4	9	3	
	1	2	6	
Total	14	12	5	31

FIG. 7.6.—Step 1. Basic feasible solution by commencing the allocation at the cell with the lowest cost and continuing to the next lowest cost cell, etc. Total cost = £118.

each customer by a very small quantity, x, and the production of one colliery by nx. There will always be $(m + n - 1)$ routes in use if the problem is worked with these slight adjustments. When the final answer has been obtained ignore the x's.

Demand and Capacity Unequal

If the total that can be sent exceeds the total that is required then introduce a " dummy " customer who requires the difference. The transport costs to this customer are all zero. Similarly if the total requirement is greater introduce a " dummy " colliery.

The Simplex Technique

Certain problems which involve the mixing of various constituents to form a number of products cannot be solved by the simple transportation techniques. As a result the method referred to as either the " General Method " or the " Simplex Method " is used. The general method may be applied to all linear programming problems, but it is not used if transportations or a similar method can be employed.

Example

Three grades of coal *A*, *B* and *C* contain as impurities phosphorus and ash. In a particular industrial process up to 100 tons of fuel is required which does not contain more than 3% ash nor more than 0·03% phosphorus. The National Coal

Board wishes to maximise its profit whilst satisfying these conditions. There is an unlimited supply of each grade.

In order to meet these requirements it is necessary to blend the three grades of coal. It is required to find the proportions of the grades which will give the maximum profit.

The percentage of the impurities and the profit of the grades are shown below :

Coal	% *Phosphorus*	% *Ash*	*Profit shillings/ton*
A	0·020	3·0	12
B	0·040	2·0	15
C	0·030	5·0	14

Suppose that x tons of A, y tons of B and z tons of C are used in the mixture.

The phosphorus content must be not more than 0·03%.

i.e. $(0{\cdot}02x + 0{\cdot}04y + 0{\cdot}03z) \leqslant 0{\cdot}03(x + y + z)$

$$\therefore -0{\cdot}01x + 0{\cdot}01y \leqslant 0$$

$$\therefore -x + y \leqslant 0 \quad . \quad . \quad . \quad (4.1)$$

Similarly the ash restriction gives

$$-y + 2z \leqslant 0 \quad . \quad . \quad . \quad (4.2)$$

Since demand is for not more than 100 tons

$$x + y + z \leqslant 100 \quad . \quad . \quad . \quad (4.3)$$

Also there are no negative tonnages

$$x \geqslant 0 \; ; \quad y \geqslant 0 \; ; \quad z \geqslant 0 \; . \quad . \quad . \quad (4.4)$$

$$\text{The profit} = 12x + 15y + 14z \quad . \quad . \quad (4.5)$$

The problem is now a mathematical one of maximising equation (4.5) subject to the conditions (4.1), (4.2), (4.3) and (4.4).

In order to solve the problem by the " Simplex " method the inequalities (4.1), (4.2) and (4.3) are made into equalities by adding new non-negative variables (called " slacks ") r, s and t to each equation. One is added to each and equations (4.6), (4.7) and (4.8) result. The " slacks " contribute zero to the profit.

$$-x + y + r = 0 \quad . \quad . \quad . \quad (4.6)$$

$$-y + 2z + s = 0 \quad . \quad . \quad . \quad (4.7)$$

$$x + y + z + t = 100 \quad . \quad . \quad (4.8)$$

A table (7.1) is drawn up containing the coefficients of the variables in these equations and the profit of each variable in function (4.5).

In Table 7.1, row w indicates the profit contributions of the variables (this row will be omitted from the succeeding tables). The second, third and fourth rows give the coefficients of equations (4.6), (4.7) and (4.8) respectively. The p-row (at this stage) contains the negatives of the elements in the w-row.

A basic feasible solution is selected by choosing one variable from each equation and making it positive—those variables not chosen are put equal to zero. This is usually done by equating the slacks to the right-hand sides of the equations.

TABLE 7.1

$w =$	12	15	14	0	0	0	P
	x	y	z	r	s	t	
$r =$	−1	1	0	1	0	0	0
$s =$	0	−1	2	0	1	0	0
$t =$	1	1	1	0	0	1	100
$p =$	−12	−15	−14	0	0	0	0

In the example $r = 0$, $s = 0$, $t = 100$. Although two of these are zero (as well as x, y, z) the problem can still be solved by imagining $r = \theta$, $s = \theta$ where θ is a very small positive number. This basic feasible solution is indicated in Table 7.1 by the r, s, t in the first column. One of the variables will be interchanged with one of the slacks and the profit increased. The variable to be introduced first is the one with the largest negative value in the p-row (i.e. y). To decide which variable to replace, the elements in column P are divided by the corresponding elements in column y. The variable replaced is the one lying in the row which gives the *smallest positive* value for this division, i.e. variable r in the first row yields $\theta/1$ where θ is substituted for 0 (see Table 7.2). This row will be called the pivot row and the element lying in the first row and the chosen column (y) the pivot element (is circled in Table 7.2). When y is introduced into the solution in place of r, the coefficient of y in the first row must be one and in the other rows

zero. This is achieved by first dividing the pivot row by the pivot element and then subtracting multiples of the pivot row from the other rows. For this problem the pivot element is

TABLE 7.2

	x	y	z	r	s	t	P	
r	−1	[1]	0	1	0	0	θ	Pivot row
s	0	−1	2	0	1	0	θ	
t	1	1	1	0	0	1	100	
p	−12	−15	−14	0	0	0	0	

1 so it is only necessary to add the elements of the pivot row to the elements of the s-row ; subtract them from the elements of the t-row and add 15 times them to the p-row. Now y is

TABLE 7.3

	x	y	z	r	s	t	P
y	−1	1	0	1	0	0	θ
s	−1	0	2	1	1	0	2θ
t	2	0	1	−1	0	1	$100 - \theta$
p	−27	0	−14	15	0	0	15θ

substituted for r in the first row. The value of the profit is shown at the intersection of the p-row and P-column (Table 7.3).

The first stage of the calculation has now been performed. It should be noted that the negative numbers in the p-row

TABLE 7.4

	x	y	z	r	s	t	P	
y	−1	1	0	1	0	0	θ	
s	−1	0	2	1	1	0	2θ	
t	1	0	$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$50 - \frac{1}{2}\theta$	Pivot row
p	−27	0	−14	15	0	0	15θ	

show by how much the profit will increase if one unit of the variable is introduced in place of one variable already in the solution. The positive numbers indicate how much the profit will decrease. The object is, therefore, to make all the terms

in the p-row non-negative. The second stage in the calculation is performed in exactly the same way as the first. The variable x will be introduced in place of t; $\begin{pmatrix} 100 - \theta \\ 2 \end{pmatrix}$ is the only positive value. Table 7.4 is the position when the pivot

TABLE 7.5

	x	y	z	r	s	t	P
y	0	1	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$50 + \frac{1}{2}\theta$
s	0	0	$\frac{5}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$50 + 1\frac{1}{2}\theta$
x	1	0	$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$50 - \frac{1}{2}\theta$
p	0	0	$-\frac{1}{2}$	$1\frac{1}{2}$	0	$13\frac{1}{2}$	$1350 + 1\frac{1}{2}\theta$

row has been divided by the pivot element and Table 7.5 shows the situation at the end of the second stage.

At the third stage the variable z is introduced in place of s; $\begin{pmatrix} 50 + 1\frac{1}{2}\theta \\ \frac{5}{2} \end{pmatrix}$ is the least of the positive value of the division.

TABLE 7.6

	x	y	z	r	s	t	P	
y	0	1	$\frac{1}{5}$	$\frac{1}{5}$	0	$\frac{1}{5}$	$50 + \frac{1}{5}\theta$	
s	0	0	1	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$20 + \frac{3}{5}\theta$	Pivot row
x	1	0	$\frac{1}{5}$	$\frac{2}{5}$	0	$\frac{1}{5}$	$50 - \frac{1}{5}\theta$	
p	0	0	$-\frac{1}{2}$	$1\frac{1}{2}$	0	$13\frac{1}{2}$	$1350 + 1\frac{1}{2}\theta$	

TABLE 7.7

	x	y	z	r	s	t	P
y	0	1	0	$\frac{2}{5}$	$-\frac{1}{5}$	$\frac{2}{5}$	$40 + \frac{1}{5}\theta$
z	0	0	1	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$20 + \frac{3}{5}\theta$
x	1	0	0	$-\frac{2}{5}$	$-\frac{1}{5}$	$\frac{2}{5}$	$40 - \frac{1}{5}\theta$
p	0	0	0	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{93}{5}$	$1360 + \frac{2}{5}\theta$

Again, Table 7.6 shows the position when the pivot row has been divided by the pivot element. In Table 7.7 it will be seen that all the elements in the p-row are positive and hence the answer has been obtained.

Hence the mixture is 40 tons of grade A
40 tons of grade B
20 tons of grade C

$$\text{The phosphorus content} = \tfrac{1}{100}\{40(0{\cdot}02) + 40(0{\cdot}04) + 20(0{\cdot}03)\} = 0{\cdot}03\%$$

$$\text{ash content} = \tfrac{1}{100}\{40(3{\cdot}0) + 40(2{\cdot}0) + 20(5{\cdot}0)\} = 3{\cdot}0\%$$

$$\text{Profit} = 40(2) + 40(15) + 20(14) = 1360/-$$

Summary of Rules for Transportation Technique

1. Set up cost information and capacities in form of a table, see Fig. 7.2.
2. Obtain a basic feasible solution (using $m + n - 1$ routes) either by (i) starting in top left-hand corner and assigning the minimum of the row and column totals ; then move to the next cell (in row or column) and allocate as much as possible to this cell and so on (Fig. 7.3), or by (ii) start at the least-cost cell and allocate as much as possible then move to the next available least-cost cell and so on (Fig. 7.6).
3. Assign a value zero to one of the rows or columns (Fig. 7.3).
4. For cells IN USE in the row (or column) chosen in (3) assign to the corresponding column (or row) the cost of those cells.
5. For all other cells IN USE make the sum of the row value and the column value equal to the cost of those cells (Fig. 7.3). (The row and column values may be negative.)
6. For all other cells (i.e. not in use) calculate (sum of row and column value minus actual cost). For positive or zero results of this difference write the difference in top left-hand corner of the cells (Fig. 7.3). Ring the largest of these.
7. Add a quantity θ into the cell with the largest difference.
8. Subtract θ from a cell IN USE in the same row as the cell in (7), add θ to another cell in use in this route's column.
9. Continue adding and subtracting θ to cells in use in rows and columns until the cell with the largest difference is

reached (Figs. 7.3, 7.4). It is always possible to find a unique way of doing this.

10. Make θ as large as possible so that all the quantities in the cells are positive or zero.
11. Repeat steps (3), (4), (5) and (6).

then either

12. If there are no positive values of the differences the answer has been reached. (A zero difference indicates that an alternative answer exists. To find this put θ in a cell with zero difference.)

or

13. If there is a positive value of the difference repeat steps (7) to (11).

Summary of Rules for General Simplex Technique

1. Write the restrictions in the form of equations by adding (or subtracting) a different non-negative variable (called slacks) to each. The contribution to the profit function from these variables is zero.
2. Set up a table of the coefficients of the restrictions (Table 7.1).
3. Select a basic feasible solution by making one variable in each equation non-zero and the remaining variables zero. The non-zero variable only occurs in one equation. This can usually be done by equating the slacks to the right-hand side of their respective equations.
4. Write the negatives of the profits in the bottom p-row. If the slacks are chosen for the first feasible solution do nothing with the p-row. (Variables in the solution have zeros in the p-row.)
5. To maximise, select the column with the largest negative value in the p-row and call it the pivot column.
6. Divide the elements in the P column by the corresponding elements in the pivot column. Call the row giving the smallest POSITIVE value of this division the pivot row. The element at the intersection of the pivot column and the pivot row is called the pivot element.
7. Divide each element in the pivot row by the pivot element.

8. For each row (other than the pivot row) take the element in the pivot column and multiply the pivot row by it. Subtract the corresponding multiples of the pivot row elements from the chosen row. All the entries in the pivot column (except the pivot element) will now be zero.
9. If all the elements in the p-row are non-negative the answer has been reached, otherwise repeat the process, starting at (5).
10. The answer is obtained by selecting (for each row) the variables which only occur in that row and equating it to the number in the P column.

Note

(i) To minimise, it is necessary to make all the elements in the p-row non-positive.

(ii) If there is a choice of which row to choose as the first row, choose the one with the bigger element in the pivot column. If these are identical choose any one.

Verification of the Method of Solution of a Transportation Problem

By considering the following small problem it is possible to verify the method of solution of a transportation problem given in this chapter. Three factories A, B and C can send a_1, a_2, a_3 respectively to three customers a, b, c who require b_1, b_2, b_3 respectively. The transport cost for the route from i to j is c_{ij}. It is required to find the minimum cost allocation.

Let x_{ij} be the amount sent from i to j. Then write a solution on the usual chart as shown below.

	a	b	c	Total
A	x_{11} c_{11}	x_{12} c_{12}	c_{13}	a_1
B	c_{21}	x_{22} c_{22}	x_{23} c_{23}	a_2
C	c_{31}	c_{32}	x_{33} c_{33}	a_3
Total	b_1	b_2	b_3	

FIG. 7.7

Where $x_{11} = b_1$
$x_{11} + x_{12} = a_1$
$x_{12} + x_{22} = b_2$
$x_{22} + x_{23} = a_2$
$x_{23} + x_{33} = b_3$

The cost of this solution is

$$x_{11}c_{11} + x_{12}c_{12} + x_{22}c_{22} + x_{23}c_{23} + x_{33}c_{33} = M \text{ (say)}.$$

As before, calculate the partial costs where the value of the second row is α.

		a	b	c	Total
		$c_{11}+c_{22}-c_{12}-\alpha$	$c_{22}-\alpha$	$c_{23}-\alpha$	
A	$c_{12}-c_{22}+\alpha$	x_{11} c_{11}	x_{12} c_{12}	c_{13}	a_1
B	α	c_{21}	x_{22} c_{22}	x_{23} c_{23}	a_2
C	$c_{33}-c_{23}+\alpha$	c_{31}	c_{32}	x_{33} c_{33}	a_3
Total		b_1	b_2	b_3	

FIG. 7.8

Call the difference between the sum of the row and column values and the actual cost d_{ij} then :

$$d_{13} = c_{12} + c_{23} - c_{22} - c_{13} + d - \alpha$$
$$d_{21} = c_{11} + c_{22} - c_{12} - c_{21} + d - \alpha$$
$$d_{31} = c_{11} + c_{22} + c_{33} - c_{12} - c_{23} - c_{31} + d - \alpha$$
$$d_{32} = c_{22} + c_{33} - c_{23} - c_{32} + d - \alpha$$

Suppose that θ_1 is introduced into the route (c, a), i.e. (3, 1) then the solution must be altered as shown below :

		a	b	c	Total
		$c_{11}+c_{22}-c_{12}-\alpha$	$c_{22}-\alpha$	$c_{23}-\alpha$	
A	$c_{12}-c_{22}+\alpha$	$x_{11}-\theta_1$ c_{11}	$x_{12}+\theta_1$ c_{12}	c_{13}	a_1
B	α	c_{21}	$x_{22}-\theta_1$ c_{22}	$x_{23}+\theta_1$ c_{23}	a_2
C	$c_{33}-c_{23}+\alpha$	$+\theta_1$ c_{31}	c_{32}	$x_{33}-\theta_1$ c_{33}	a_3
Total		b_1	b_2	b_3	

FIG. 7.9

With this alteration the total cost becomes :

$$(x_{11} - \theta_1)c_{11} + (x_{12} + \theta_1)c_{12} + (x_{22} - \theta_1)c_{22} + (x_{23} + \theta_1)c_{23} + \theta_1 c_{31} + (x_{33} - \theta_1)c_{33}$$

$$= x_{11}c_{11} + x_{12}c_{12} + x_{22}c_{22} + x_{23}c_{23} + x_{33}c_{33} - \theta_1\{c_{11} - c_{12} + c_{22} - c_{23} - c_{31} + c_{33}\}$$

$$= M - \theta_1 d_{31}$$

Hence if (i) d_{31} is positive, the total cost will be decreased
(ii) d_{31} is zero, the total cost is unaltered
(iii) d_{31} is negative, the total cost will be increased.

Since the rule which is performed is to alter the solution by introducing routes with positive d_{ij}s the cost will be reduced at each stage.

REFERENCES

Introductory

GASS, S. I. (1958). *Linear Programming* (McGraw-Hill).

HENDERSON, A., and SCHLAIFER, R. (1954). " Mathematical Programming : Better Information for Better Decision Making ", *Harvard Business Review*, Vol. 32, May–June, pp. 73–100.

WILKINSON, J. J. (1956). " Better Management ", *Journal of Industrial Engineering*, Vol. 7, No. 1, January–February, pp. 11–23.

ROSANDER, A. C. (1956). " The Use of Linear Programming to Improve the Quality of Decisions ", *Industrial Quality Control*, Vol. 12, No. 9, March, pp. 11–16.

VAJDA, S. (1958). *Readings in Linear Programming* (Pitman).

WAGNER, H. M. (1958). " The Simplex Method for Beginners ", *Journal of Operations Research Society of America*, Vol. 6, No. 2, March, pp. 190–9.

Applications

KAY, E., and MORTON, G. (1955). " Linear Programming—An Application to Swarf Recovery in a Foundry ", *Metal Working Production*, December 30th, 1955, p. 2215.

STRINGER, J., and HALEY, K. B. (1957). " Applications of Linear Programming to a Large Scale Transportation Problem ", *Proc. of First International Conference on Operational Research* at Oxford (E.U.P.), 1958.

Theoretical

KOOPMANS, T. C. (Editor) (1951). *Activity Analysis of Production and Allocation* (New York, John Wiley & Sons, Inc.).

CHARNES, A., COOPER, W. W., and HENDERSON, A. (1953). *An Introduction to Linear Programming* (New York, John Wiley & Sons, Inc.).

VAJDA, S. (1956). *The Theory of Games and Linear Programming* (New York, John Wiley & Sons, Inc. ; London, Methuen).

Further bibliography may be found in :

Journal of the Operations Research Society of America, Vol. 5, 1956, pp. 45–62, 555–62.

RILEY, V., and GASS, S. I. (1958). *Linear Programming and Associated Techniques* (Johns Hopkins).

Chapter VIII

DISTRIBUTION

IT is usually rather difficult to draw clearly defined classification limits for human activities. The definition of distribution problems given below is, therefore, necessarily an arbitrary one. For the purposes of this chapter questions of how optimally to get one's goods or services in the right amounts and at the right time to where they are required will be taken to constitute problems in distribution.

A simple question for a manufacturer to ask may be " What is the most economic route to use in order to get my goods into the consumer's hands ? " This is not a field that has been fully studied yet, but certain parts of it have been looked at beneficially. The normal chain of command is, factory, depot, wholesaler, retailer, consumer. Certain types of problem have attracted a great deal of attention ; in Chapter VII above, there is an account of the technique known as " Linear Programming ", and in this chapter an application of linear programming which is concerned with the cheapest way of delivering goods to a range of customers from a set of different origins.

There are other fields of distribution of a simpler nature, however, which may well repay attention ; the question of whether it may pay to have a depot in a specific town is one of these. It is not difficult to decide where depots could be placed beneficially, by considering the difference between transport costs in bulk and transport costs for small vehicles for direct delivery (but see below). It is not difficult to arrive at distances between depots from quite different considerations. Since transport in bulk is always cheaper than transport in detail, and since in general a 30-mile radius from the starting point is the limit that can be conveniently undertaken by a local delivery vehicle, it is not difficult to see how frequently depots should be placed. Bearing in mind the necessity of an overlap they would be presumably less than 60 miles apart.

This is in fact rather too generous, for utilisation of vehicles in bulk delivery is important, and if a vehicle is working a shuttle service it would appear that a depot should be a multiple of 2 hours away from the factory ; 2-hour journey from the factory—2 hours away—4 hours away—6 hours away is not very convenient—8 hours away is perfectly feasible. Six hours away may imply some juggling with 2-hour journeys ; but if a depot is 2 hours from a factory, and loading and unloading facilities are adequate, a bulk-delivery vehicle can make two journeys a day ; at 4 hours one journey a day, and at 8 hours an outward journey one day, return journey the next. A 6-hour journey may require an overnight stop at an intermediate depot. Any distances violently different from these stages in time cannot fail to cause uneconomic use of vehicles.

Another field of distribution which has yielded to elementary model-making is that of the worth-whileness in cost in delivering small orders. Models have been made of the cost of running a vehicle, in terms of cost per day and cost per mile. A portmanteau rule (which does not fit all circumstances) is to say that if the capacity of a vehicle is t tons its cost per day is $£t/5$—although of course this will depend to some extent on the make of the vehicle : the cost of running it is $1\frac{1}{4}t$ pence per mile ; the total cost of the vehicle is obtained by taking these two figures and adding the cost of the crew. In a rather different field the cost of delivery may be put as a standard cost of call which may vary from a matter of two or three minutes to half an hour plus so much per unit of weight delivered. The total cost of the delivery therefore is made up of the time taken to get to the customer, plus the time spent at the customer, which in its turn is dead time, plus so much per unit of weight delivered.

These simple fields seem to have made little impact on Operational Research literature, but can be very rewarding. It is not difficult to envisage a situation where the delivery cost for a small amount of goods may be grossly in excess of the profit on the goods ; for instance sending a six-ton lorry 30 miles in a fairly busy area implies something like a £2 cost. This can only be worth while if the margin on the goods is enormous or the value of the goods is perhaps ten times or twenty times more than the delivery cost. Work of this nature makes little

appeal to the deviser of advanced techniques but can be extremely profitable as a development in thinking on the subject of worth-whileness of small orders.

Another field of distribution that is not without interest is that of frequency of delivery. This is of course linked in with vehicle utilisation, and with stock-holding. It may be noted that doubling the frequency of delivery does not halve stock-holding levels ; it halves the cycle stock, but it only reduces the buffer stock by a factor of $\sqrt{2}$. This particular aspect of distribution again has been studied to some extent in conjunction with stock-holding ; it has also been studied in connection with purchasing.

Such questions, however, are only one aspect of business management, and whatever decisions are taken in respect of distribution, impinge on all other activities of an organisation. Hence it is often dangerous and sometimes impossible to consider such problems in isolation. These interconnections will be pointed out in the examples and case studies that are used as illustrations in this chapter.

Recently some distribution problems have been associated with one particular form of linear programming, that of the transportation model. Although the use of linear programming has marked a great advance in the solution of certain distribution problems, it is by no means the only tool of the O.R. worker in this field. Nevertheless, it is of great help in formulating the mathematical model of a problem. The parameters that are to be used in obtaining practical solutions from the model are, however, often the result of other mathematical techniques, in particular, statistics.

In a chapter of this length it is only possible to give an outline of the more frequent types of problem in this field, indicating the form of the mathematical model and the techniques used. It is, however, hoped that this will give a guide to the reader for further study.

All questions of distribution arise from the geographical and/or temporal separation of demand from supply briefly touched on above. The O.R. approach to the problem of how most efficiently to bridge this gap in space or time consists in describing the situation in abstract terms, that is setting up a mathematical model and then manipulating this model

by the rules of mathematics to yield a solution that is optimal by some criterion, previously agreed to between management and O.R. scientist.

In order to demonstrate this line of approach a very simple example will be used. Assume that a business expands its activity into a new area and requires to set up a centre in this area, from which to distribute its goods. The question of optimal location for the distribution centre then arises. The optimal location is defined as that point in the area from which

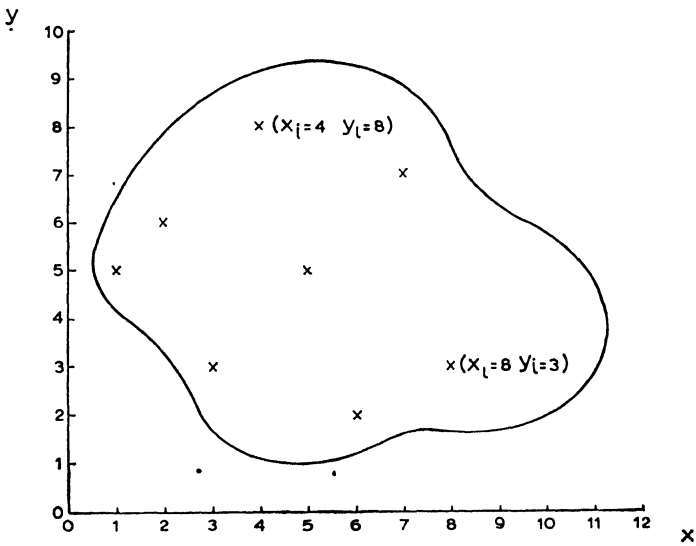


FIG. 8.1

the total distribution costs are a minimum. Assume further that distribution costs are proportional to demand and distance from the centre, and that the demand at all points in the area is known. The formulation of a mathematical model of this situation can then be undertaken by superimposing a Cartesian two co-ordinate system on the area, so that each point, i, in the area can be located by reference to two numbers, x_i along the horizontal axis, and y_i along the vertical axis (Fig. 8.1). As distribution costs are proportional to demand and distance over which they have to travel, the cost, C_i of getting the amount of goods required at point i, from the distributing

centre to its destination, equals the product of the distance times the amount times a factor of proportionality, or symbolically : $C_i = k.d_i.D_i$, where D_i is the demand at point i, and d_i the distance between point i and the centre.

Now the distance of two points in a cartesian two co-ordinate system is :

$$d(i,j) = \{(x_j - x_i)^2 + (y_j - y_i)^2\}^{1/2},$$

which is obvious from the Pythagorean theorem.

Giving the unknown distribution point the co-ordinates (x,y) the formula for the distribution cost to point i can be written as :

$$C_i = k.D_i\{(x_i - x)^2 + (y_i - y)^2\}^{1/2}. \quad . \quad (1)$$

The optimality criterion is smallest total distribution cost, hence the sum of the distribution costs to all demands points in the area has to be minimised, that is

$$\text{Minimise} \quad C = k \sum_{i=1}^{n} D_i\{(x_i - x)^2 + (y_i - y)^2\}^{1/2} \quad . \quad (2)$$

given there are n demand points in the area.

Expression (2) is the mathematical model of the problem.

A value for x and y has now to be found which, when substituted into (2), will minimise that expression. The solution is found by the use of differential calculus and turns out to be :

$$x = \frac{\sum_{i=1}^{n} D_i x_i}{\sum_{i=1}^{n} D_i} \quad \text{and} \quad y = \frac{\sum_{i=1}^{n} D_i y_i}{\sum_{i=1}^{n} D_i}$$

that is the co-ordinates of the optimally located distribution centre are the weighted arithmetic means of the demand points ; the demand at a point being its weight.

As has already been indicated, this example was intentionally simplified in order to demonstrate the O.R. approach. It serves, however, also as a pointer to two other important considerations :

(i) the interconnection of questions of distribution with other activities of an enterprise, and
(ii) the danger of over-simplification in abstracting a model from a given situation.

The example was based on the assumption of known demand at all the points ; but, unless a good estimate of these demands is available, the results derived from the model become misleading, even though the model is a correct representation of the actual situation ; or, in professional language, the D_i are the parameters of the model, and the correct estimation of their value is another—and in this case prior—task of the O.R. worker.

Alternative means of distribution were not considered and distribution costs were assumed to be proportional to amount and distance carried. Although transport charges are often based on the ton/mile, a carrier frequently offers quantity rebates. Both alternative means of transport, and non-proportional costs if applicable, have to be built into the model. Using company-owned transport the question of optimal routing forms another problem that would have to be considered concurrently.

Finally, even given that a firm's activities expand into a new market, it is by no means *a priori* certain that the best (i.e. cheapest) means of satisfying the new demand is by setting up a new distribution centre in the area. One should consider the new market together with the old and try to optimise the overall distribution pattern.

It may now be appropriate to describe an actual case in which some of the above-mentioned points have been taken into consideration.

The background to the case is as follows : An organisation had acquired over time a number of factories in various areas. These factories, all of different production capacities, served also as distribution centres for the area, using the company's own transport fleet. The product was a foodstuff, and delivery was direct to company retail outlets. The demand for the product did not show any significant seasonal variation, but the geographical demand pattern for the country had changed, and total demand stood below total production capacity. Some reorganisation had, therefore, become necessary. The O.R. Department, entrusted with the planning of the reorganisation, was asked to consider not only the closing down of some factories, but also, whether in view of the changed demand pattern, it might not be worthwhile, either

to erect a factory or factories at new sites, or subcontract to outside firms in some areas. As a criterion of optimality, minimum cost of satisfying the forecast long-term demand was given.

There was one simplifying feature, namely, that for technical reasons the preferred means of transport were the company's lorries, so that, at least in the first instance, one would not have to consider alternative means of transport.

Although the whole problem does fall within our definition of distribution problems, being concerned with the best way of overcoming the spatial separation between sources of supply and points of demand, it is clear from the start that questions of production must to some extent form part of the study.

It was decided to follow the following plan of campaign : Formulate a model, based on available data (mainly accounting documents), then use this model to obtain a first, approximate, answer and indications for further detailed studies, in order to improve the model, which thus should yield improved results.

It was in the nature of the problem that this first model would be on the lines of a transportation type linear programme, but taking into account production costs as well as transport costs. In consequence the first difficulty was how to deal with the large number of demand points (the destinations of the transportation matrix (see Chapter VII on Linear Programming)). From accounts data it appeared that delivery costs within an area under existing arrangements were a fixed percentage of the value, at sales price, of the goods delivered ; that percentage varying with the area. It was then reasoned that, given an area's factory was closed, its demand could be satisfied from another factory by delivering the required total amount to the site of the closed factory, and using the area's transport organisation for retail delivery within the area as before. This sounds a far-fetched procedure, but has the advantage that it reduces the number of demand points in the model to the number of existing factories. In practice, the demand of an area, whose factory was to be closed, would be split up among the remaining factories, and new routing schedules would be arranged from those factories to retail outlets. One can, however, say that delivery cost under this more

practical arrangement would certainly not be higher than costs under the system of bulk delivery to former factory site and local delivery from there, and that both arrangements probably have a similar cost structure. For the purpose of formulating the model it was assumed that the bulk delivery arrangement would become operative ; in this way the difficulty of having some 700 demand points was overcome.

The delivery cost from any factory to any area, and for any amount of product, can now be written formally using the notation as below :

$0 \leqslant x_{ij}$ = the amount of product delivered from factory in ith area to retail outlets in jth area
γ_j = the proportion of the sales value of goods delivered that accounts for distribution cost within the jth area
p = the sales price
τ_{ij} = the unit transport cost from site of factory in ith area to site of factory in jth area.

This gives for delivery cost T_{ij} :

$$T_{ij} = \tau_{ij}x_{ij} + \gamma_j p x_{ij} = t_{ij}x_{ij}$$

where $$t_{ij} = \tau_{ij} + \gamma_j p \quad . \quad . \quad . \quad . \quad (3)$$

Next find an expression for production costs.

Let $$X_i = \sum_{j=1}^{n} x_{ij} \qquad i, j = 1, 2, \ldots n$$

that is X_i is the total production at the factory in ith area, and there are *in toto* n factories and areas

K_i = production capacity at the factory in ith area
α_i, β_i = positive constants, typical for factory in ith area
C_i = production cost at factory in ith area.

From a statistical study of accounting data the following relationship between production cost and amount produced was derived

$$\frac{C_i}{pX_i} = -\frac{-\alpha_i X_i}{K_i} + \beta_i \quad . \quad . \quad . \quad (4)$$

or, in words, the ratio of cost of production to sales value of production decreases as full capacity working is approached ;

this relation was found to be valid for all factories, the constants α, β were characteristic of an individual factory, and effectively, pointers to its profitability.

From (4) production costs can be derived as an explicit function of amount produced by setting

$$\alpha_i p/K_i = a_i \quad \text{and} \quad \beta_i p = b_i$$

and writing $$C_i = -a_i X_i^2 + b_i X_i \quad . \quad . \quad . \quad (4')$$

Consider first the case of finding an optimal production and distribution pattern for existing factories only. The object is to minimise the total production and delivery cost, say, *TC*. The formal expression for *TC* must be the sum over all factories and area of expressions (3) and (4), i.e.

$$TC = \sum_{i=1}^{n} a_i \left(\sum_{j=1}^{n} x_{ij} \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} (b_i + t_{ij}) x_{ij} \quad . \quad (5)$$

and in analogy to the transportation model of linear programming we can now write :

Maximise (5), subject to : $$\sum_{i=1}^{n} x_{ij} = D_j$$

$$\sum_{j=1}^{n} x_{ij} \leqslant K_i \quad . \quad . \quad . \quad (6)$$

where D_j is the demand in the jth area.

The first expression under (6) stands for a set of n equations, one equation for each area, and states that the sum of production in each factory, destined for a given area, must equal the demand in that area. The second expression stands for a set of n inequalities, one for each factory, and states that the sum of the amounts produced for all areas in any given factory must be less than, or equal to, the production capacity of that factory.

Leaving aside, for the moment, the question of solvability of this " quadratic " programme, a solution to (5) and (6), that is, a set of values for the x_{ij}, which satisfies the equations and inequalities of (6) and at the same time minimises the value of (5), would indicate which, if any, of the factories ought to be closed, the amounts of goods each factory has to produce and from which factory or factories each area has to be supplied.

Next comes the question of possible new factories, and/or subcontracting. For the former one had to determine possible location, an expression for production cost, and production capacity. Delivery cost would, of course, be of the same form as expression (3).

Location was found by a calculation, similar to that given in the theoretical example on page 183, resulting in three possible sites for new factories. Production cost and capacity were obtained from estimates of the technical department, the former being of the same form as (4′).

Possible new factories can, therefore, be fitted into the above model simply by treating them in the same way as existing ones. Giving the subscripts $n + 1$, $n + 2$, $n + 3$, to the three possible new factories, the only change in the model as stated by (5) and (6) is in the upper summation limits of the set of equations of (6) from n to $n + 3$.

As for subcontracting, quotations were obtained from various merchants for supplying given areas, inclusive delivery to retail outlets. These quotations fell into two groups, some quoting a single price per unit, irrespective of quantity delivered, some offering quantity discounts. Expression (4′) can be used to symbolise the cost of having a given area's demand met by subcontractors in the following way :

(i) for the case of a single price quotation : set $a_i = 0$, and b_i equal to the quoted price. The subscript i refers now to the ith subcontractor.

(ii) for the case of quotations with quantity discounts ; Drawing a graph of cost against quantity one obtains a broken line, as in Fig. 8.2. By suitable selection of constants a_i and b_i, a curve can be fitted to this graph, the algebraic formula for which is (4′). Expression (4′) therefore can be used as a sufficiently close approximation to the quoted costs.

It is now possible to extend the model given by (5) and (6) to cover the case of subcontractors. Again, all that is needed is to extend the upper limit of summation over the i subscripts to $n + 3 + k$, if there are k possible subcontractors, and setting $t_{ij} = 0$, when i refers to a subcontractor.

Expressions (5) and (6), therefore, give the basic model of

the problem. A first solution was obtained, using the sales department's estimate of future average yearly demand, the known capacities and cost functions of existing factories, and the technical department's estimate of capacity and cost function for new factories as parameters.

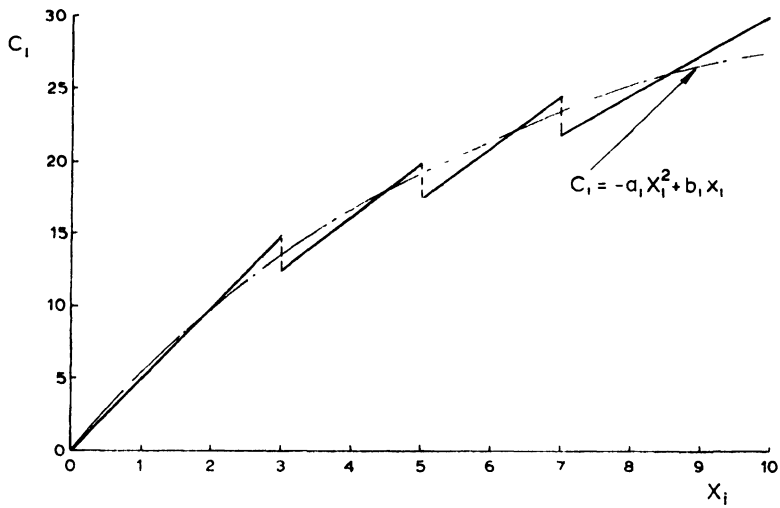


FIG. 8.2

As the model is quadratic, not linear, the computation proceeded in the following way : The quadratic part of (5) was at first disregarded and the following transportation linear programme solved

Minimise

$$\sum_{i=1}^{r}\sum_{j=1}^{n}(b_i + t_{ij})x_i \qquad \left.\begin{matrix} i = 1, 2 \ldots n + 3 + k \\ j = 1, 2, \ldots n \end{matrix}\right\} . \qquad (7')$$

subject to (6).

This yielded a set of values for x_{ij}, say, x'_{ij}, and for each factory an amount of unused capacity, say $S_i \geqslant 0$. The following transformation was then made :

$$y_{ij} = \frac{x_{ij}}{X_i} ; \quad S_{ij} = S_i y_{ij} ; \quad X_i = \sum_j x_{ij} \qquad (7)$$

and the linear programme

$$\text{Minimise } Z = \sum_{i=1}^{r} - aX_i + \sum_{i=1}^{r} \sum_{j=1}^{r} (b_i + t_{ij})y_{ij} \quad . \quad . \quad (8)$$

subject to

$$\left.\begin{aligned} &\sum_{j=1}^{n} y_{ij} = 1 \\ &\sum_{i=1}^{r} (K_i y_{ij} - S_{ij}) = D_j \\ &\sum_{j=1}^{n} S_{ij} + X_i = K_i \end{aligned}\right\} \quad . \quad . \quad . \quad (9)$$

solved, using as a first feasible solution values for y_{ij}, S_{ij} and X_i obtained from the values x'_{ij} by transformation (7). The final, optimal solution to (8) and (9) gives the optimal solution for (5) and (6) by back substitution into (7).

The result of this computation indicated that a small number of factories could satisfy the expected future demand at a cost relatively and absolutely smaller than that of the present system.

The point has now been reached where perhaps the most important part of the O.R. work on this problem starts ; work that, incidentally, is still in progress. All the assumptions underlying the model must now be investigated and proven, and the model must finally be tested to see how far a solution remains optimal when demand varies within certain limits. In the space of this short chapter only the O.R. studies that stem from this first model can be enumerated.

(i) a study of the production process, with the twofold aim of confirming the cost parameters, and possibly improving them.

(ii) a study of methods for forecasting future demand and the limits of its fluctuation.

(iii) given a solution, an investigation of the new routing schedules arising from it and testing whether the assumption that sending the total area demand to a given point in the area for local distribution is a sufficiently close approximation to the distribution arrangement that would actually be used.

The case described is typical, in that it shows the central position of mathematical programming in problems of distribution and at the same time underlines that this alone is not

sufficient to obtain a satisfactory result. Other tools of O.R., such as statistics, Monte Carlo simulations, etc., have to be used in conjunction with the programming model, before one can say that an investigation is successfully brought to an end.

Up to now only problems arising from the spatial separation of demand from supply have been considered and those due to temporal separation have been ignored. These normally appear when pronounced seasonal demand fluctuations exist and capacity to produce supply during a peak demand period is less than the demand in that period. In such circumstances one must, as a rule, produce for the peak demand in preceding

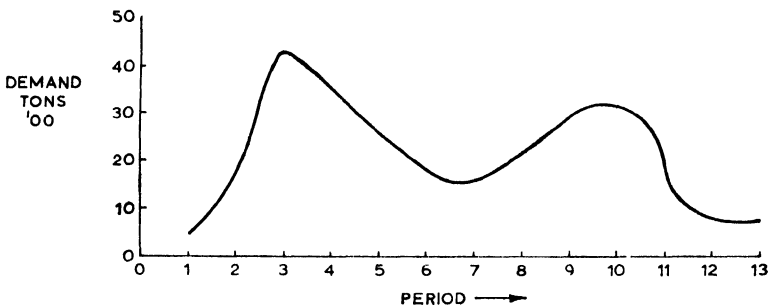


FIG. 8.3

periods and store the product till required. One can, however, also try to even out demand variations by offering rebates during off-peak periods.

Again, an actual case will be used as the basis for developing a model. The case chosen involves a " two-sided " distribution problem ; that is, the planner has to consider not only the temporal distribution of finished product, so as to satisfy customer demand, but also the raw material buying pattern, arising from that distribution, the raw material being a commodity dealt with on the open market and therefore having a fluctuating price. The model whose development will be described is again a mathematical programme, and again more than one of the company's departments were involved in the operation.

This then is the background of the case. A business manufactures a chemical used in agriculture. The demand pattern for a typical year is shown in Fig. 8.3. The company-owned

store attached to the factory is used both for storing raw material and finished product. Both demand peaks are in excess of monthly production capacity, and at least one of them is so high that some additional storage space will have to be hired some months beforehand. Some customers can be induced by rebates to buy the product before they need it and store it at their own expense. Finally, the factory can operate on a one-, two- or three-shift basis. Therefore, at the beginning of the year's operation a number of interrelated decisions have to be made, namely :

(i) a raw material buying schedule
(ii) a production schedule
(iii) a rebate scheme, and/or
(iv) a hiring schedule for additional storage.

The basic information available for these decisions are two forecasts, one of demand, one of raw material prices, for the year. In tracing the development of the model from which all these decisions will be derived, these forecasts will be taken as given.

A linear programming model for production scheduling under fluctuating demand conditions has been given by E. H. Bowman[1] and the formulation of the model for this case was based on Bowman's original idea.

Dealing first with the production schedule, introduce the following notation :

x_{ijk} = the amount produced in the kth shift of the ith month, for sale in the jth month ; $i, j = 1, 2, \ldots n$; $k = 1, 2, 3$
C_{ik} = the production capacity of kth shift in ith month
D_j = the demand in jth month
α_{ik} = unit production cost in kth shift of ith month
β = unit storage cost per month
η = sales price of product.

For the variable x_{ijk} it is obvious that $j \geqslant i$, as one cannot produce during one period for sale in a preceding period ; equally $x_{ijk} \geqslant 0$, as one cannot produce a negative amount of goods.

Given then that in some months demand exceeds production

capacity, so that in order to meet that demand some of the previous months' production must be held in store, a production schedule is required, which either minimises total costs or maximises profits. For reasons that will become clear later, maximisation of profits was chosen as optimality criterion. Therefore, assuming that there is no stock in hand at the beginning of the planning period, the model is :

$$\text{Maximise} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{z} \{\eta - [\alpha_{ik} + (j - i)\beta]\} x_{ijk} \quad . \quad (10)$$

$$\text{Subject to} \quad \left.\begin{aligned} \sum_{i=1}^{j} \sum_{k=1}^{z} x_{ijk} &= D_j \\ \sum_{j=1}^{n} x_{ijk} &= C_{ik} \end{aligned}\right\} \quad . \quad . \quad . \quad (11)$$

This is a transportation-type linear programme. If there were stock in hand at the beginning of the planning period it could be taken into the model by counting subscripts i and k from 0 and designating by x_{0j0} the amount from beginning stock destined for sale in jth period, and by C_{00} the total amount of beginning stock. This will add one inequality to the set under (11). In a similar way production for stock can be taken care of by increasing the j subscripts by one, and designating by $x_{i,n+1,k}$ the amount produced for stock in the kth shift of ith month and by D_{n+1}, the amount of stock that ought to be in hand at the end of the planning period. This will add one equation to the set under (11), and, of course, the summations will all be over the extended range of subscripts.

As initial and final inventories thus fit easily into the model, in order to simplify matters it will be assumed that in this case there were no initial or end stocks.

A solution to the linear programme given by (10) and (11) will indicate amounts to be held in store over certain periods. It has not taken into account that storage space is restricted. If, however, the paramount consideration is to meet demand, this restriction need not be considered because, provided one is willing to pay the price, one can always obtain as much storage space as is needed. Nevertheless, a storage hiring schedule is needed that is optimal in the sense of profit maximisation. From the solution to (10) and (11) one can derive

such a schedule by ranking available storage space in order of storage cost (cheapest first), and distributing the amount to be stored per period, to the stores in that order. That is, if in month i the amount to be held in store is x_i (this is obtained from the solution, thus $x_i = \sum_{r=1}^{i} \sum_{j=i+1}^{n} \sum_{k=1}^{z} x_{rjk}$) and unit storage cost of the rth ranking store is b_r and its capacity S_r, y_r will designate the amount to be stored in this store. The storage schedule is then a set of values y_m thus : $\{y_1, y_2, \ldots, y_l\}$

where $y_1 = S_1 \quad y_2 = S_2, \ldots y_{l-1} = S_{l-1}, \quad y_l \leqslant S_l$

and $$\sum_{m=1}^{l} y_m = x_1$$

The cost of storing the amount x_i for one month is $\sum_{m=1}^{l} b_m y_m$ and since $b_1 < b_2 < \ldots < b_l$ this cost is the minimum achievable in these circumstances.

In this case the lowest storage cost was, understandably, that of the company-owned store, and the β of expression (10) refers to that cost. The true profit (i.e. excess of income from sales over total costs) must, therefore, be computed by subtracting from the value of the solution the cost of storage in hired stores, if amounts to be stored exceed the company's store capacity.

The model given by (10), (11) and (12) describes the situation as far as demand for the finished product is concerned. There is, however, the question of raw material, which occupies storage space and ought to be bought at the most favourable time, and has therefore to be considered simultaneously with the production and storage schedule.

Call :

ν_i — the amount of raw material bought in the ith month
δ_i = its price
ω = a conversion factor ; so that the amount of raw material needed to produce x_{ijk} of finished goods can be written as ωx_{ijk}.

As the raw material is a commodity, it has two forecast prices, for spot buying and forward buying respectively ; δ_i is the lesser of the two prices for raw material to be delivered in month i.

The amount of raw material available at any period must at least be equal to total production during this period. This can be expressed formally by a set of n inequalities, thus :

$$\sum_{i=1}^{l} v_i \quad \omega \sum_{i=1}^{l} \sum_{j=l}^{n} \sum_{k=1}^{3} x_{ijk} \geqslant 0 \qquad . \qquad . \quad (13)$$

or more explicitly :

$$\left.\begin{array}{l} v_1 - \omega \sum_{j=1}^{n} \sum_{k=1}^{3} x_{1jk} \geqslant 0 \\ \vdots \\ v_1 + \ldots + v_l - \omega\left(\sum_{j=1}^{n} \sum_{k=1}^{3} x_{1jk} + \ldots + \sum_{j=l}^{n} \sum_{k=1}^{3} x_{ljk} \right) \geqslant 0 \end{array}\right\} \quad (14)$$

The left side of inequalities (14) gives also the amount of raw material that has to be stored during the 1st, 2nd, . . . lth month, at a cost of β times that amount.

The model given by (10) and (11) can now be expanded to cover also the raw material buying.

In order to demonstrate how this is done in detail the formulation will be set down, assuming that the planning period consists of four months only and that only single-shift working is possible. Subscript k can therefore be omitted.

The total storage cost of raw material is then :

$$\begin{aligned} &\beta\{(v_1 - \omega[x_{11} + x_{12} + x_{13} + x_{14}])\} \\ &\quad + [v_1 + v_2 - \omega(x_{11} + x_{12} + x_{13} + x_{14} + x_{22} + x_{23} + x_{24})] \\ &\quad + [v_1 + v_2 + v_3 \quad \omega(x_{11} + x_{12} + x_{13} + x_{14} \\ &\qquad + x_{22} + x_{23} + x_{24} + x_{33} + x_{34})] \\ &\quad + [v_1 + v_2 + v_3 + v_4 - \omega(x_{11} + x_{12} + x_{13} + x_{14} \\ &\qquad + x_{22} + x_{23} + x_{24} + x_{33} + x_{34} + x_{44})] \end{aligned}$$

$$\begin{aligned} &- 4\beta v_1 + 3\beta v_2 + 2\beta v_3 + \beta v_4 - 4\beta\omega(x_{11} + x_{12} + x_{13} + x_{14}) \\ &\qquad - 3\beta\omega(x_{22} + x_{23} + x_{24}) - 2\beta\omega(x_{33} + x_{34}) - \beta\omega x_{44} \end{aligned}$$

or, in general, during the nth month, multiply x_{ij} by $-\beta\omega(n + 1 - i)$ and v_i by $\beta(n + 1 - i)$.

The expected sales income realised from x_{ij} is ηx_{ij} ; this costs $\alpha_i x_{ij}$ to produce and $(j - i)\beta x_{ij}$ to store. A table (Table 8.1) can then be drawn up showing how to calculate the total profit realised, given values for the x_{ij} and v_i, by multiplying each x_{ij} or v_i by all the factors shown below it, and summing all these products. From Table 8.1 the objec-

TABLE 8.1. Cost factors

	x_{11}	x_{12}	x_{13}	x_{14}	x_{22}	x_{23}	x_{24}	x_{33}	x_{34}	x_{44}	v_1	v_2	v_3	v_4
	η	η	η	η	η	η	η	η	η	η	η	η	η	η
	$-\alpha_1$	$-\alpha_1$	$-\alpha_1$	$-\alpha_1$	$-\alpha_2$	$-\alpha_2$	$-\alpha_2$	$-\alpha_3$	$-\alpha_3$	$-\alpha_4$				
	$4\omega\beta$	$(4\omega-1)\beta$	$(4\omega-2)\beta$	$(4\omega-3)\beta$	$3\omega\beta$	$(3\omega-1)\beta$	$(3\omega-2)\beta$	$2\omega\beta$	$(2\omega-1)\beta$	$\omega\beta$	-4β	-3β	-2β	$-\beta$
											$-\delta_1$	$-\delta_2$	$-\delta_3$	$-\delta_4$
Total	a_{11}	a_{12}	a_{13}	a_{14}	a_{22}	a_{23}	a_{24}	a_{33}	a_{34}	a_{44}	c_1	c_2	c_3	c_4

tive function (that is the expression that is to be maximised) can be written as below :

Maximise :

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\{\eta_j - \alpha_i + \beta[i - j + \omega(n+1-i)]\}\chi_{ij} - [\delta_i + (n+1-i)\beta]\nu_i \quad . \quad (15)$$

and the conditions of this maximisation as :

$$\left.\begin{aligned} &\sum_{i=1}^{I} x_{ij} = D_j \\ &\sum_{j=1}^{n} x_{ij} \leq C_i \\ &\sum_{i=1}^{I} \nu_i - \omega \sum_{i=1}^{I}\sum_{j=1}^{n} x_{ij} \geq 0 \end{aligned}\right\} \quad . \quad . \quad (16)$$

The complete linear programming tableau is given in Table 8.2. The expansion to n months and k shifts is obvious.

Except for the consideration of rebate schemes, the model for the case is now complete. Formally, a rebate scheme will consist of a list of sales prices, each price being valid for a given month—so that we can write : $\eta_i : i = 1,2, \ldots, n$. The effect of this will be a change in the D_j. The model cannot easily be expanded to cover this change, as it concerns the parameters, not the form of the model. Furthermore, demand and raw material price values are notoriously difficult to forecast in this case, the one being influenced by the so unsteady weather, the other by political factors. What can be done in a case like this is to recompute a solution for the model, using different sets of demand and price parameters, thus obtaining a list of optimal solutions for varying circumstances ; in effect, one uses a simulation procedure. Maximisation of profits has been chosen as optimality criterion in preference to minimisation of cost—just because of the need to consider a changing price structure, and lowest total costs in these circumstances do not necessarily mean highest profits. A direct computational method to obtain such a list of optimal solutions for sets of possible prices is given by S. I. Gass and T. L. Saaty.[2] In the case here described, because the change affected the demand values also, a number of solutions were computed for sets of demand values, forecast on the basis of

TABLE 8.2. Linear Programming Tableau

Unknowns	x_{11}	x_{12}	x_{13}	x_{14}	x_{22}	x_{23}	x_{24}	x_{33}	x_{34}	x_{44}	v_1	v_2	v_3	v_4	
Values	a_{11}	a_{12}	a_{13}	a_{14}	a_{22}	a_{23}	a_{24}	a_{33}	a_{34}	a_{44}	c_1	c_2	c_3	c_4	
	1	1	1	1											$\leqslant C_1$
					1	1	1								$\leqslant C_2$
								1	1						$\leqslant C_3$
										1					$\leqslant C_4$
	1														$=D_1$
		1			1										$=D_2$
			1			1		1							$=D_3$
				1			1		1	1					$=D_4$
	$-\omega$	$-\omega$	$-\omega$	$-\omega$							1				$\geqslant 0$
	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$				1	1			$\geqslant 0$
	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$		1	1	1		$\geqslant 0$
	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	$-\omega$	1	1	1	1	$\geqslant 0$

given rebate schemes, and each recomputed for various sets of raw material prices.

Although the basic models used in the two cases described are of the same class, i.e. mathematical programmes, the way in which they are used is different. In the first case the model seemed to yield a reasonable answer, and research was directed to prove the assumptions and values on which the model was based. In the second case there was a situation of uncertainty as regards the parameters, and direct statistical methods are not yet sufficiently developed to deal with a forecasting problem of this complexity ; therefore, the model was used as the basis for a simulation procedure to obtain a picture of how varying conditions influence the optimality of a set of decisions.

The models described in this chapter are by no means the only ones suitable for distribution problems. In the technical literature a number of others are described, theoretically or with case histories. The purpose of the chapter will, however, have been accomplished if it has shown the reader that it is possible to describe distribution problems formally in the language of mathematics, and that, once this has been done, the formulation or model will itself indicate the investigations or actions necessary to obtain reasonable answers to the given problem.

REFERENCES

[1] BOWMAN, E. H. " Production Scheduling by the Transportation Method of Linear Programming ", *Opns Res.*, **4**, No. 1, Feb. 1956, pp. 101–3.

[2] GASS, S. I., and SAATY, T. L. " Parametric Objective Function ", Parts 1 and 2, *J. Opns Res. Soc. Am.*, **2**, No. 3, Aug. 1954, pp. 316–19 ; **3**, No. 4, Nov. 1955, pp. 395–401.

*

Chapter IX

LABOUR WASTAGE AND TURNOVER

CHAPTER XIII discusses some of the human problems which may be encountered during an Operational Research investigation. Some were those which may be directly provoked by the introduction of O.R. into a firm, such as the natural antagonism which the O.R. man, by his mere presence as an investigator, may arouse in the firm, or the difficulties he may encounter at the stage of implementation. The other group consisted of problems which already existed and which either contributed to the difficulties of the investigation (such as a rivalry between two men in executive position) or may indeed have been the indirect cause of the O.R. investigation being initiated in the first place (such as poor morale in a department which was reflected in, say, low utilisation of manpower in that section of the firm).

The existence and extent of dissatisfaction in a firm are indicated by the rate of labour turnover* and the labour wastage.* Many studies have been made of these subjects, and of particular interest among these are a number which tend to support the idea that most of such human problems are avoidable and traceable to such factors as a poor communication system in a firm.[1]

Undoubtedly such factors are important and are in some way related to labour turnover. However, this approach will not be developed further in this chapter, which will be confined to the methods of study of the fundamental pattern and cost of labour turnover and wastage on the assumption that there

* These terms are not synonymous. Labour wastage is the ratio of the number of leavers in a given period to the total manpower over that period ; labour turnover refers to the process of continual change in manpower (i.e. includes both wastage and recruitment). The general method of calculation of wastage, which is the one on which figures used here are based, is :

% Wastage = No. of leavers in the year $\times$ 100/½(no. of men at beginning of year + no. of men at end of the year)

would still be a labour turnover problem even if employer/employee relations, communications, amenities, etc. were all that could be wished for.

Operational Research is essentially a quantitative method of reasoning, and when applied to labour turnover attempts to measure the factors affecting labour turnover, the cost incurred and the monetary value of, say, a 5% reduction in the labour turnover of a factory. It may well be that the measurable saving achieved would be substantially less than the cost of the improvement. In such a case management would have to judge whether those savings which result from a more stable labour force and are not measurable, are sufficient to make the expenditure worthwhile.

In the analysis of costs it is the marginal cost of labour turnover which is of real importance and it is possible through O.R. to arrive at an estimate of this cost. Before doing this, however, it is necessary to consider the pattern of labour turnover, because the various categories of worker (the definitions being those of expected length of service and not necessarily of type of job) cost different amounts to replace.

The basic approach used in O.R. to the problem of labour turnover and wastage is then to regard it as an example of replacement theory. From this it follows that the first step is to break down the numbers of people involved into various categories, each of which has its own loss rate.

The wastage pattern of a firm can be examined by various methods of analysis, but the most useful have proved to be the following and it is with these that subsequent discussion will be concerned: analysis can be of the length of service for men grouped by age, occupation (i.e. job and status within the factory), place of residence (in terms of distance from the factory) and last job (by type and length of duration).

In the discussion which follows wastage due to factors outside the men's control, such as death, redundancy and retirement (due to age, disability or illness), are not considered. Except where there is a sudden cut in total manpower, wastage in these categories is small and can be accurately forecast.

Several studies carried out in various industries have shown that the length of service distribution of any group of new entrants is log-normal. In other words a large proportion of

any group leaves early in service (early peak in curve), but the rate of leaving then decreases rapidly, with a few staying on relatively permanently; an example is given in Fig. 9.1. Amongst other things this means that if a factory increases its manpower rapidly, perhaps because of production expansion, even with recruitment of men of similar wastage characteristics to those previously recruited, it can expect its total wastage to increase rapidly, perhaps to twice its normal value, and then after three years or so to regain its normal level if total manpower remains fairly constant. The log-normal

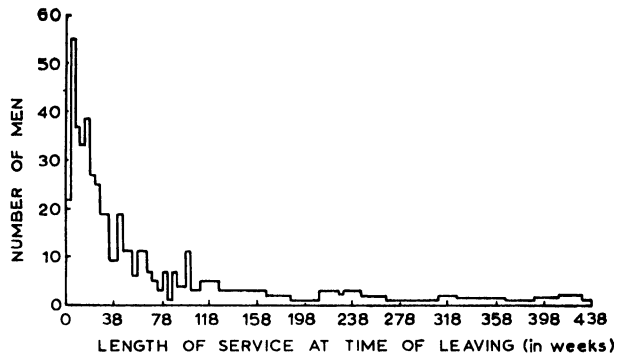


FIG. 9.1

distribution also has the useful property that when plotted on log-probability paper a straight line results. Thus forecasting wastage of a group of recruits is relatively simple after the first few have left. It also means that the wastage of different groups of recruits can be compared by the percentages who have left within, say, their first year of service.

By far the most important conclusion resulting from operational research studies of labour wastage, is that the distance a man lives from a factory has a big effect on both the likelihood of his wanting to work at the factory and his expected length of service if he does work there. This can be summarised by the example given in Fig. 9.2 of the supply of labour and wastage at various distances for factories employing about 1,000 men in rural areas with no competitors of similar labour requirements within a mile. Where competition is stronger, such as when another large factory is in the same

town, the percentages of populations attracted are less but the wastages are only slightly increased.

Thus a factory can be regarded as exerting an attractive force which rapidly dissipates with distance outwards. In some respects these attractive forces can be considered to be similar to those of magnets. Thus in this " model " a factory can be considered to be a magnetic pole. Competitors can be considered as poles with different pole strengths situated at

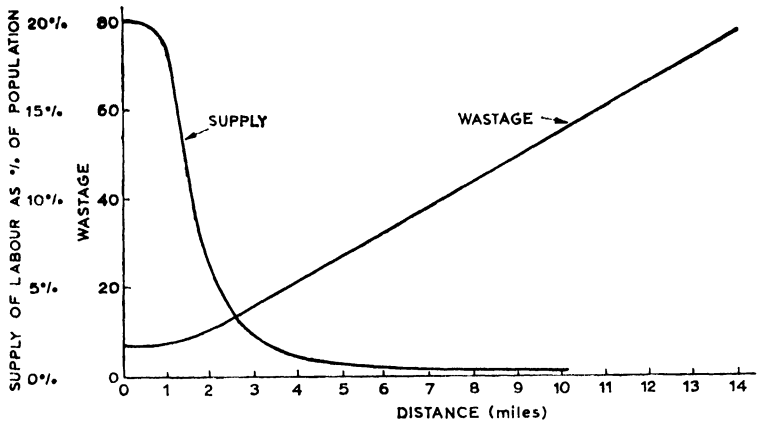


FIG. 9.2

various distances from each other. Populations can be considered as piles of iron filings which for some reason divide between the various poles according to the relative strengths of the various fields at the point considered. From time to time the particular iron filings that go to the various poles will change and the frequency of changing (i.e. wastage) will depend again on the relative field strengths (e.g. if only two fields exist and one is 100 times as strong as the other the amount of changing will be low). The flaws in this model compared to actual magnetic fields can easily be seen, but this does illustrate the type of model which appears to exist.

The supply of men to the factory will therefore depend mainly on the location and size of populations and to a lesser extent on the location and relative competitive strength (cf. relative attractive strength of magnets) of other factories.

The wastage of men working at a factory will depend mainly on how large the supply of labour is relative to requirements but to a lesser extent on where the labour is located. Thus it has been found that where there is an ample supply of labour, management tends to select people living close to the factory and the wastage is generally less than 10%; where supply and requirements are roughly equal, the wastage may be expected to be about 15%; where supply is inadequate and special measures have to be taken to get the manpower required, wastage is at its highest, but is usually within the range of 20% to 40%. This is not an altogether surprising finding, since it is obvious that when jobs are plentiful a man can afford to be more choosy than when they are hard to find.

This model concept of the supply of manpower to a factory can be used in conjunction with Fig. 9.2 to estimate the supply of manpower for a new factory or when manpower requirements are likely to increase. Thus, for example, a factory in a rural area requiring a labour force of 1,000 with a total population three times that size within 2 miles will be able to recruit about half of its men from that source and with the normal distribution of populations should be able to get the others from other nearby towns and villages.

However, if the nearest population were, say, 5 miles away only about 0·3% would be available to the factory under normal circumstances, and if competition near that town was exceptionally strong (e.g. say another large factory located in the town) the proportion attracted would be considerably less. Thus a population of some 300,000 at 5 miles would normally be necessary if no population were available nearer than this. It will be noted from Fig. 9.2 that the supply of manpower normally reduces rapidly at about 2 miles. Very few factories can therefore maintain their manpower without about half of their labour travelling less than 2 miles.

In times of decreasing employment in local factories the supply of manpower available to a factory increases and hence more men within the lower wastage groups are recruited, thus slightly reducing the total wastage. Conversely, when unemployment rates decrease, the labour one is able to recruit has on average a higher wastage.

Although much of what has been said above has long

been realised, it is only recently that quantitative evidence has been collected and analysed to show the pattern which exists.

Knowledge of the effect of distance on labour wastage will be of value when a firm finds that its recruitment methods are attracting an inadequate number of men who live within daily travelling distance of the factory. Possible courses of action are then either to do nothing, which will result in a wastage of about 15% to 20% and a fall in total manpower; or extra efforts may be made to recruit men locally, by such things as greater recruitment drives among school-leavers, more advertising, or the provision of special buses—by this method it will usually be possible to maintain the required level of manpower, but the wastage incurred will be about 20%. In some industries, such as mining (where the place of work is determined by the minerals rather than by the supply of manpower), there may be a third possibility, namely, to recruit men from quite a distance away. Wastage then depends largely on whether or not housing is provided ; if it is not, then the wastage for such men may quite easily be over 200%. If housing is provided, the length of service characteristics of the men change from one log-normal distribution to another, corresponding roughly to that for local men housed at the same distance from the factory. The policy to be followed will of course depend on the individual firm and the urgency of its requirements, but this sort of information can help management to decide what recruiting policy is best suited to their needs and to the amount of money they are prepared to spend.

At this point it may be as well to mention briefly the findings on the value, in terms of reduction of labour wastage, of the provision of special transport where a factory is situated in an area where normal transport is not adequate. In one district where a factory wished to attract additional men from a new housing estate, a special bus service was introduced by the firm for a trial period of nine months. The object was to show that the local manpower supply had not been fully tapped and to measure how many more men were recruited and how the labour turnover costs were affected.

Turnover figures and costs for the year preceding the intro-

duction of the bus and the year following it were compared. The figures showed that recruitment from the new estate rose after the bus service began and that the higher level was maintained throughout the year. Wastage figures for men from the estate who already worked at the factory did not alter appreciably, but for workers recruited since the experiment began the wastage level was slightly higher. However, without the bus service it would have been necessary to recruit men from much farther away, and their turnover cost would have been considerably more than that of the recruits from the new estate, and the expenses incurred in the running of the bus service.

All in all then, the experiment showed that the bus service was more than worth while and it was accordingly decided to make it permanent. As one might expect it has been found elsewhere that where transport is not available the supply of manpower may be well below the level recorded at other places where transport facilities are good. For example, at 3 miles, good transport has been shown to double the number of men available.

The geographical position of a firm in relation to sources of manpower is one major factor which influences labour turnover and wastage. The situation can be affected to some extent by transport and housing facilities, but in general it is an inescapable fact that the farther away a factory is situated from its source of manpower the more difficult it will be to recruit men in the first place, and to persuade them to stay at the job.

The other really important factor which significantly affects labour wastage is the age of the men joining a firm. Analysis of numbers of men leaving a factory by age-groups, as in Table 9.1, shows that men in the 14 to 19 age-group stay longer than those aged between 20 to 35. 15-year-old entrants have a wastage of about 5%, but this figure increases with age of entry up to about 25 years old and decreases thereafter up to retirement.

This overall pattern has been found to hold true in most other factories and would therefore seem to be of fairly general significance. When the 15 to 19 age-group is broken down further, it is found that the low wastage in this group is mainly

due to the number of school-leavers ; by the age of 17, recruits have as high a wastage as the higher age-groups.

This again will have an effect on recruiting campaigns and selection of new employees when there is an abundance of applicants. It would obviously be ridiculous to state that personnel departments should always prefer a 15-year-old applicant to one of 25. Clearly there are other considerations which have to be borne in mind, such as experience, technical ability and the like, but the effect of age should not be forgotten.

TABLE 9.1. *Length of service by age of joining—averaged over 8 years of recruitment*

Length of service (weeks)	*Age on joining (years)*							
	14–19	*20–24*	*25–29*	*30–34*	*35–39*	*40–44*	*45 and over*	*Total*
0–26	92	254	201	120	77	43	35	822
26–52	45	77	61	43	32	13	24	295
53 and over	208	168	205	138	156	124	165	1164
Total	345	499	467	301	265	180	224	2281

Another feature which influences labour wastage is the position a man occupies in a firm. Studies carried out so far show that wastage is generally lower amongst men who have a skilled job and higher among the unskilled. This links up with the effect of both length of service and age, and is what one might expect, seeing that experience will come with age. Clearly men who have acquired a certain degree of knowledge of one industry will be less likely to leave to take a job demanding a different kind of skill which they will have to learn from the beginning. On the other hand, men whose basic work can be learned in a day or two will have a different outlook. This again is open to analysis by the social scientist, who may argue that acquisition of specialised knowledge will help to provide a man with a sense of security in his work, whereas the man who feels that any fool could do his job is likely to feel insecure and more ready to be attracted by, say, higher wages offered elsewhere.

Having discussed some of the factors which influence labour

turnover, and wastage, it is important to stress that nonsensical conclusions can be reached if any one factor is interpreted without reference to the others.

So far no mention has been made of the other main aspect of labour turnover and wastage which is important to management, namely, the cost. Having an idea of the pattern of wastage is helpful, but the ultimate question which has to be answered is : " What does this mean in terms of money ? " Many studies have been made of the cost of labour turnover in various industries.[2] Inevitably the items of cost which are significant vary according to the type of industry, but some of the factors most frequently included in the estimate are : advertising costs of recruiting, training costs, costs due to productivity differences between inexperienced and experienced workers, cost of damage to product or machines due to lack of experience of workers, cost of idle machine-time due to vacancies caused by losses of men working on particular machines. Clearly the possible range of cost of a man leaving is very great. Studies suggest that it can be anything from £8 to £1,000 depending of course on the industry and on the degree of skill of the leaver.

In many studies of the cost of labour turnover the various cost items given above are totalled and divided by the number of leavers and the turnover cost is expressed as the *average* cost per man leaving. While this figure may be interesting it is not one on which decisions can be made, because some labour turnover must always be present, the cost of some of the items may not be affected by the amount of turnover and because the cost per leaver cannot be considered without reference to the length of service (for example a factory could have at the lower limit a turnover rate of, say, 5 %, with no one leaving voluntarily and with a turnover cost per leaver of, say, £1,000 —it may be that this cost could not be reduced). What are really needed then are marginal costs of turnover for different groups of men expressed as costs per man-year of service obtained. For any given job, say for fitters, recruitment policy can then be based on the minimum cost per year of service obtained after allowing for the time of training and any differences in performance between a man who leaves after many years of service and one who has just completed his training.

For example, consider a job for which two types of men recruited from beyond daily travelling distance are to be considered. Without permanent housing the men stay an average of 6 months, with housing they stay 10 years. Training takes 3 months and, together with the other turnover costs, losing a man costs a total of £100. Provision of housing loses the company £50 per year. Ignoring the inaccuracy resulting from the fact that all men do not stay exactly the same time the two costs per man-year of service excluding training can be summarised as follows :

Housed men $\frac{£100}{9\frac{3}{4}}$ + £50 per man-year = about £60 per man-year

Men not housed $\frac{£100}{\frac{1}{4}}$ = £400 per man-year.

Thus it follows that transferred men should be housed. Similarly, cost decisions have to be made on whether 15-year-olds should be recruited, whether advertising or buses to recruit more people of certain categories are worth while, etc. In each case the saving in turnover cost per man-year has to be compared with the cost of changing the pattern of recruitment.

In analyses of this type it must not be forgotten that in some cases even the cheapest means of maintaining the marginally high cost groups in factory manpower may be such that it would be better to reduce the planned level of production rather than to accept the high turnover cost.

REFERENCES

[1] See for example

GUEST, R. H. (1955). "A Neglected Factor in Labour Turnover", *Occupational Psychology*, Vol 29, No. 4, pp. 217.

HORNE, R. S. (1949). "This Firm Keeps its Labour", *Bulletin of Industrial Psychology and Personnel Practice*, Vol. V, No. 4, pp. 15–21.

JAQUES, E., RICE, A. K., and HILL, J. M. (1951). "The Social and Psychological Impact of a Change in Method of Wage Payment—The Glacier Project V", *Human Relations*, Vol. IV, pp. 315–40.

POIDEVIN, B. L. (1956). "A Study of Factors Influencing Labour Turnover", *Personnel Practice Bulletin*, Vol. XII, No. 4, pp. 11–18.

[2] See for example

COOK, P. H. (1950). "Financial Aspects of Labour Turnover", *British Management Review*, Vol. IX, No. 2, July, pp. 51–9.

"Labour Turnover", *Personnel Management Series I*, British Institute of Management.

BYRT, W. J. (1955). "Recent Research into the Cost of Labour Turnover", *Personnel Practice Bulletin*, Vol. XI, No. 3, Sept., pp. 29–38.

Chapter X

MILITARY APPLICATIONS

Historical Introduction

THERE are three main reasons for the presence of a chapter on military Operational Research in a book devoted to industrial and commercial applications. The first is for historical purposes, for many developments of technique stemmed from military problems. The second reason is, however, more cogent. Industrial applications have (in general) followed military applications. There has been a strong link in type of approach to problems and in mathematical techniques. This link can be demonstrated and from its existence it can be argued that industries and businesses not yet using Operational Research must have problems in common with other concerns which now use Operational Research.

The third reason (in many ways similar to the second) may be the most important of all. Military Operational Research is now concerned with fields where industrial Operational Research has hardly penetrated. Again, arguing by analogy, it is likely that industrial Operational Research scientists will follow the paths already being explored by their military counterparts.

No doubt work of a nature now described as Operational Research has been done since the earliest days of the human race. It is unlikely that information about the size and duties of the Great Pyramid O.R. section will ever come to light, or that British history will be enlivened by discovery of O.R. studies on the most economic method of constructing Stonehenge. Diverting thoughts on studies of the relative merits of different types of castle (or the optimum size of cathedrals) must alas be viewed as fiction. It is a pity that it is most unlikely that the weapon evaluation group which discussed the relative merits of arquebus, crossbow and longbow ever published its findings.

Some published information on early applications of

scientific ideas to warfare does exist. Possibly the best-known example in Graeco-Roman times is that of Archimedes with his super catapults.[1] A later and even better known example is the work of Carnot, Monge and, to a lesser extent, Fourier,[2] although it can be said that these were eminent scientists employed in military tasks rather than men applying scientific method to military problems.

The earliest well-known application of Operational Research to warlike problems is the work of Lanchester.[2] Lanchester developed a mathematical theory of the strategy of concentration of forces. He showed the effects of differing assumptions on fire power and attrition laws in a manner which permitted prediction of results of warfare. These results seem to be more useful in naval and aerial battles than land battles, where the effects of terrain may be considerable. This work is still of interest. A recent paper[3] has extended Lanchester's work, and at the same time shown the importance of mathematical methods in studying warfare.

In 1937 the Royal Air Force received a new equipment (radio direction finder—R.D.F.) now called (generally) radar. This posed many operational problems, and a research section was formed to deal with them. In order to distinguish it from other research sections the word " operational " was added to its title. This section was the first ever to devote itself entirely to studies of operations. By the outbreak of World War II in 1939 the section was well established. By 1940 other activities had come into being. Some of their work has been described by various workers at the British Association 1947 meeting at Dundee.[4]

In general, military Operational Research was at this time concerned mainly with weapons, equipment and tactical problems. The range of work varied from analysis of data (particularly for explanations of abnormalities from which ultimate advantage might be secured) in connection with the newly developed radar equipment and studies of the distribution of hits and penetrations in our own and the enemy's tank casualties and the influence of this on tactics, to the planning of bomber attacks over Germany and a study of the optimum size for a merchant convoy to be at its most effective—in terms both of minimum losses from submarine action and of minimum

escort requirement. In all instances, emphasis was on repetitive situations where sufficient data could be assembled to permit analysis.

Probability

An important concept introduced at about this time was that of probability. In any military operation there is a large element of uncertainty. When a weapon is fired the projectile may or may not hit its target. If it does, the damage may be unimportant, serious (i.e. prematurely reducing the effectiveness of the victim) or lethal. Clearly there will be a continuous damage spectrum rather than distinct categories. Studies of weapon effectiveness must make allowances for these situations. Statistical methods were particularly successful in dealing with aircraft problems, both of the fighter and bomber forces. A ready extension to anti-aircraft gunnery followed.

In a quite different field, a fighting vehicle may be virtually immune against one type of attack, but not immune against all possible attacks. It may be defenceless against some attacks, but have an ascertainable chance against still another type. The circumstances of usefulness depend on the relative frequency of the different types of attack. Similarly, a track vehicle can usually negotiate some obstacles, but there are others which usually defeat or bog-down the machine. All this leads to studies of the circumstances in which a particular armoured tracked vehicle is likely to be useful, and of circumstances in which it is quite certainly not useful.

These ideas of probability were extended to many other fields. Work was done on supply problems in the Royal Air Force. The mathematical ideas evolved were published in 1946.[5] This paper has been followed by a vast literature in the fields of queueing theory and stock problems generally. Some recent advances were recorded in Chapter IV, above. This is an excellent instance of a problem of equal interest to industry and the military as opposed to techniques or types of thinking which can be applied to problems of greatly differing appearance but with underlying common features.

It is not suggested that this military interest in provisioning was the first recorded. Pioneer work in queueing as applied to telephony dates from the early twentieth century.[6] Even

so, industrial interest in mathematical treatment of stock problems has followed military interest.

Queueing theory has applications in telephone systems to provision of trunk lines, layouts and exchange equipment. This work is linked with problems of what sort of equipment is needed. Considerations of objectives and relative merits of different types of equipment are just as essential to a satisfactory system as good design. In the earliest stages of new equipment, design of something which will work must take first place. In later stages, however, a different concept will apply.

This may be illustrated by two military examples. The first is the simple case of, say, a new field gun. Guns have been made for some hundreds of years, and so, although designing a new gun may be tedious, it is not very difficult once a technical specification is written. Writing a technical specification is also easy—if the new gun is merely a scaled-up or slightly improved version of the old one. If, on the other hand, the gun is to be a real advance there is great difficulty in writing the specification. The soldier can say what is the purpose of the weapon, but can seldom write a specification for a designer. A designer can seldom know whether what he produces is what is wanted. The Operational Research scientist is the link between soldier and designer. He has adequate numerical data on the performance of many weapons under conditions of war. He knows how range performance must be downgraded to give battle performance. He can take the soldier's requirements for a new weapon, and, in conjunction with the designer, get out a series of technical specifications and performances for weapons which only exist in sketch forms. From these the soldier may take his choice. He may prefer lightness to long life. He may prefer an expensive weapon with cheap ammunition rather than the converse. He may sacrifice range for lethality. The Operational Research contribution lies in exposing the merits and demerits of the various options. Much work of this nature was carried out in both Britain and the United States in the years immediately following World War II.

The second example is a natural development of this work to systems of weapons. The concept here is that (taking air defence as an example) no single weapon can hope to contend

with every form of attack. Weapons can, however, be developed to meet specific threats. A gun which will deal with high-flying aircraft has a limit of height below which it is ineffective. A weapon for dealing with hedge-hopping aircraft is obviously unlikely to deal with high-altitude targets. Carrying this argument further brings guided missiles and piloted aircraft into the picture. Each of these weapons has its strength and weakness against specified attacks.

A great deal of work has been devoted to studies of a unified system of defence against all likely attacks. Obviously this is not work which can ever be published in detail for the benefit of the public. Some account has been given of the work in a recent book on Systems Engineering.[7] In this field of design of large-scale systems the analogy between, say, telephone systems and air defence systems is clear. It is not difficult to extend the ideas on weapons for a particular purpose to industrial plant. In analysis of requirements for weapons a factor of great importance is type of mechanism. Take for instance automatic or semi-automatic guns. At the pistol size there are two main types—revolver and automatic. In each case firing one round brings the next into position. But a revolver is limited to about six low-velocity rounds, whereas an automatic may fire up to a dozen or more. In the automatic, the firing round brings the next into position, whereas the trigger-finger operates the revolver. The automatic mechanism may be a plain blow-back breech. It may on the other hand lock. It is clear that blow-back mechanisms have to be restricted to low chamber pressures, and the higher the pressure the stronger the locking device.

Analysis of plant and machine tool designs and performances will have similar characteristics. It is known (although information does not seem to have been published in detail) that considerations of economic plant size are as important as market forecasts in settling specifications.

Industrial Analogies

An analogue of this type of analysis must exist in the field of batch size. In general the more complex a plant the bigger its economic batch size. Industrial Operational Research is following military work in the field of analyses of requirements,

but does not yet seem to have invaded the field of commercial policy. Much work has been done on batch size and stocking policies, but little on elimination of the undesirable. Yet it is commonplace to say that small lots, if they have to be made, should be made in plants designed to make small lots economically.

This argument that " since some businesses have problems in common with the services others will have similar problems " can be carried too far. Large businesses will always benefit from Operational Research (even if they call it something else) provided that they are willing to benefit. Small businesses will also benefit, but not perhaps so much owing to scale effects.

There are fields, however, where military Operational Research appears to have no civil counterpart. It is a normal part of the duties of a War Department continually to revise its policies and plan well ahead. [Gibes about preparing for the next war as if it was to be like the last not admitted !] More and more of this long-range policy work is being submitted to scientific scrutiny. A large proportion of military Operational Research effort goes into forward thinking, particularly in devising mathematical models and trying out techniques for future use.

It may be that there is no real industrial equivalent, for some at least of this work might be expected to be done in universities. It is unfortunate that, so far, universities as a whole have not devoted a great deal of effort to the techniques of Operational Research, although there are, fortunately, one or two outstanding exceptions. It is to be greatly hoped that there will be an upsurge of interest in management science in universities, at least comparable with the interest now being shown in technical colleges.

The industrial equivalent of long-range military studies will to a great extent depend on the nature of the industry, and it may well be that the best comparison would lie in considering and erecting a new plant or factory. This is a line of thought rather neglected in industrial circles. Operational Research has a great contribution to make in improving the efficiency of existing product manufacture, and in dealing with products whose manufacture should cease in the very near future. There seems to be no particular reason, however, why it should

not be used to cover the whole life of the product, even before the product is in production.

In the early days of military Operational Research, as in the early days of industrial research, the final decision or choice of a course of action was left to the customer. In a sense this principle must always be followed. It is discussed at some length in Chapter XVI, where it is made clear that decisions must be made by responsible authorities. However, there is a tendency for the question "What would you do in my position?" to be posed with increasing frequency. This may arise from the ever-increasing complexity of problems studied, but to some extent it is a manifestation of confidence.

Operational Research is well established in the military field. Time has only confirmed the virtue of using Operational Research and so the interval between presentation of results and their acceptance is growing less. It is very natural to hesitate over accepting new methods, especially when (as sometimes happens) they lead to reversals of popular belief.

Once proper confidence in the O.R. scientist and his methods has grown, acceptance is easy. It is in these circumstances that the O.R. scientist is expected to point out his preferred solution.

Organisation

There is another field where comparisons between military and industrial Operational Research are of great interest and are in use, and that is in organisation. Some reference has already been made in this book to the position of Operational Research in a business. Chapter XVI discusses the relation of Operational Research with management and with other management aids and disciplines. An outline[3] of British military O.R. organisation was presented at a N.A.T.O. meeting in 1957.

It is a trite observation that a thing that is not used might as well not exist, and it is vitally important when an Operational Research study has been brought to a state where decision for action may be taken, that the appropriate authority can be readily informed. Lessons may be learnt from the military in this respect. In general, military O.R. personnel, although perhaps not well remunerated by industrial standards,

have at least the status needed for ensuring that their recommendations can be carried out. If a comparison is drawn between a large industrial concern and the Army or the Air Force, it is soon apparent that exact comparisons are not possible. Even so, it may not be too inappropriate to describe the War Office Director as an equivalent to a director in a large firm, or to describe military members of, say, the Army Council, as managing directors. On this basis both the Army and the Air Force have directors in charge of Operational Research in the persons of their respective Scientific Advisers.

Operational Research establishments or sections are headed by officers of various levels, equivalent to brigadier, colonel or lieutenant-colonel. There is no difficulty whatever for an officer at this level to discuss problems freely with Directors or Commanders. It is quite normal for the head of an Operational Research service unit overseas to have direct access to the Commander-in-Chief, and perfectly normal to have day-to-day dealings with the Chief of Staff, or such heads of staff branches as may be relevant. There are still very few industrial Operational Research workers with direct access to members of their company's main board, and still fewer who have access to the chairman of the company or the managing director.

This may be a consequence of Operational Research tending to attract the younger man, but there still seems to be a belief in many firms that Operational Research can flourish at the end of a long chain of command. It is true that even in these circumstances the Operational Research scientist can make a useful contribution, but unnecessary handicaps can only be very wasteful.

To a large extent the Operational Research worker is dependent on two things : a repetitive situation and good data.

Repetitive situations spring to mind readily in both military and commercial/industrial situations. Some data are available in both cases, but usually of a kind not really relevant to the O.R. scientists' needs. Morse[9] argues that military Operational Research is easier than industrial because data are better. This American view is of interest, but may seem strange in Britain. It is true that military organisations are now used to special collection of data for the benefit of O.R., and that

industry may have a good deal to learn in this respect. On the other hand, industrial decisions frequently depend on economic criteria. It is seldom wrong to enquire if a course of action is profitable in industry, and, in general, industry is better equipped to discuss costs and profits than government service. This is not a reflection on government methods, but a consequence of a situation where spending money to save money is an impossibility. In industry additional finance may be difficult, but can usually be obtained if a good case can be made out. Additional finance for a spending Ministry can only be obtained at the cost of a cut elsewhere, additional taxation or inflation. None of these are really practicable alternatives. In theory, no doubt, a spending department's budget should allow for some spending to make savings. Restrictions, however, must always exist.

This leads to a quite different conclusion from Morse : that industrial O.R. is generally easier than military, and since military O.R. has been—and is—successful, how much more successful industrial O.R. must be.

REFERENCES

[1] BELL, E. T. (1953). *Men of Mathematics* (Pelican Books), Ch. 2, 12 and 14.

[2] LANCHESTER, F. W. (1916). *Aircraft in Warfare : The Dawn of the Fourth Arm* (London, Constable) ; *or*
MORSE, P. M., and KIMBALL, G. E. (1954). *Methods of Operations Research* (London, Chapman, & Hall), pp. 63–77.

[3] WEISS, H. R. (1956). " Lanchester-Type Models of Warfare ", *Proceedings of the First International Conference on Operational Research—Oxford 1957* (London, English Universities Press), pp. 82–9.

[4] BLACKETT, P. M. S. (1948). " Operational Research in War and Peace " *and* " Operational Research ", *Advanc. Sci.*, **V16**, 320–32 ; **V17**, 26–38 ; *also*
CROWTHER, J. G., and WHIDDINGTON, R. (1947). *Science at War* (London, H.M.S.O.).

[5] PITT, H. R. (1946). " A Theorem in Random Functions with Applications to a Theory of Provisioning ", *J. Lond. math. Soc.*, **21**, 16–22.

[6] BROCKMEYER, E., HALSTROM, H. L., and JENSEN, A. (1948). " The Life and Works of A. K. Erlang ", *Trans. Dan. Acad. tech. Sci.*, **2**.

[7] GOODE, H. H., and MACHOL, R. E. (1957). *Systems Engineering : An Introduction to the Design of large-scale Systems* (London and New York, McGraw-Hill).

[8] GADSBY, G. N. (1957). " The Service Pattern ", *The Manager*, **25**, 706–8, 714.

[9] MORSE, P. M. (1957). " The Relationship between Military and Industrial Operations Research ", Chapter 3 of *Operations Research Applied* (American Management Association Special Report 17).

Chapter XI

DATA HANDLING

WHATEVER the problem, whatever the way of tackling it, one of the important functions of Operational Research is the provision of quantitative information to management. Metrology, in the special sense of determining the appropriate unit or scale of measurement for the system under investigation, is the foundation of nearly all the quantitative work involved—including the construction of a model to replicate the behaviour of the system. Collection, analysis and presentation of numerical information may account for up to 90% of the work involved in an investigation ; if this is to be kept manageable, it must clearly be controlled, and it is the purpose of this chapter to outline some of the important principles governing the proper handling of data.

There is an oft-quoted phrase that " You cannot get more out of the data than is already in them." True enough, but sometimes a dangerous precept since its very triteness may blind you to its converse -if you mishandle the data they will yield less, or less accurate, information than they would with proper treatment. Besides, it is not always possible to tell just what information is hidden in the data that might be brought out with proper handling. The following are typical cases : non-linear correlation, the occurrence of interactions between two or more variables, the presence of an unknown parameter not yet allowed for in the model. One must be alert to such possibilities and ready to extend the analysis accordingly.

The main aspects of data handling are Collection, which includes choice of data, sources of data, and methods of collection ; secondly, Analysis, which may be subdivided into objective (which determines the methods and accuracy required), choice of criteria, and techniques of numerical analysis ; finally, Presentation.

COLLECTION OF DATA

Choice of Data

The phrase " choice of data " might be taken to imply both that the relevant problems had been established, and that there were alternative sources or sets of data available. Ideally this should be true, but practical O.R. investigations usually work against the calendar, and much valuable time is lost if the collection of data cannot be started at a very early stage ; indeed, it may well be the rough analysis of the first data that points the way towards particular problems to be solved later.

The collection of data is frequently begun before one can say exactly how they are to be used. How, then, can this collection be planned in the initial stages when so little is as yet known ? The answer lies in the thoroughness with which the team-leader drafts his initial Table of Contents for the report that he will ultimately issue. Preparing a draft Table of Contents is important at the start of any investigation, for several reasons : it helps initially in considering the background or context of the study, and makes it easier to delimit the probable scope ; it makes co-operation easier to obtain by indicating to others the objectives and the logical steps by which it is proposed to reach them ; the early examination of these objectives is an aid to selecting personnel for the team, as it indicates what technical know-how may be required ; and finally, the draft Table of Contents is a helpful guide to the probable duration and cost of the work.

If this document has been drawn up carefully, then the main fields will be evident—and their logical sequence. From this it is fairly easy to deduce the types of problem that are likely to be encountered, and hence the sort of data that will be required for analysis and so have to be collected. A brief study of this is usually enough to indicate likely sources, and these can be explored to see what already exists (and its accuracy) and what may have to be specially recorded. In this way it is possible to choose much of the data for collection at quite an early stage—before one can specify exactly how it is to be used—and this collection can continue while qualitative

information is still being sought and detailed problems defined.

It should also be borne in mind that it is usually more economic (both in time and cash) to collect in the first instance more data than are subsequently required than to mount a second study to make good deficiencies in the first collection. Of course, once a problem has been specified the necessary statistics and other information can be specified too ; if they already exist they can be collected, but if not then arrangements can be made to obtain them. Even in this case, however, it is worth a moment's thought whether the recording of one or two additional parameters at the same time might provide the necessary information for some other problem that may have to be considered later.

Sources of Data

There are three main sources of data : Past and Current Records ; Special Readings ; Simulation. The important points in considering each of these are the present availability of information, its relevance and accuracy, and the need to check or confirm all data on which conclusions may be based.

Past and Current Records. These records are the first and main source of data, with the others used to supplement, to confirm, or to correct them as necessary. There are Production and Operating statistics ; Cost Office and Accounts Department statements ; Sales returns ; Work Study data ; Engineers' records ; and so on, a vast amount of information for the sifting.

Now, these records were made in the first instance for a different purpose and what they purport to show may not be exactly what they do show, or perhaps what the O.R. team really want to know. Thus, production figures obtained from, say, the Cost Office and from Despatch may differ if the former were prepared for bonus purposes and count all pieces processed, whereas the latter discount pieces rejected at final inspection : similarly, sales returns may differ from quantities despatched if the latter include goods sent to stores or stock depots and not yet passed on to a customer. In just the same way, single series of readings must be checked for internal consistency, especially if compiled by more than one operator,

and for method—for example, was the quantity of scrap actually weighed or was it just estimated from the size of the heap ?

There is another important point to bear in mind, namely, the danger of spurious correlations. A simple case may be that a rise in temperature was often found to be associated with a fall in yield : but it does not follow that raising the temperature would necessarily decrease the yield under controlled conditions. Historically they may both have been effects of a change in some third, unrecorded, parameter—say, the presence of an impurity.

A case example of this was analysis of the coke consumption of a blast furnace per ton of iron produced, against the basicity ratio of the slag ; taken over a period of 210 weeks the two appeared to be fairly strongly correlated.[1] A superficial inspection of the data indicated that the high basicity values occurred at the end of the period, and four years is long enough for time trends to appear in most processes—certainly in a blast furnace, which is traditionally regarded as a temperamental piece of plant. Was there a true correlation or was it merely a time trend ?

To resolve this question the data were divided into six-week periods, and it was shown that these periods were short enough for values to be random and therefore free from trends. Within each six-week period the changes in coke rate and in basicity from their mean values in the period were measured ; on plotting these changes the apparent correlation disappeared. It is believed that the original data contained a time trend, for reasons that are still unknown, and that this had produced a spurious correlation.

Special Readings. Where gaps exist in the records, or where greater accuracy or confirmation is required, it may be necessary to institute special readings during the course of the investigation. These might be observations made by the team themselves, or records made by continuous reading instruments installed for this purpose, or simply additional bookings made by operators.

Such readings may not require any changes in the production line, but merely the booking of information that was always available though not previously required ; or they may involve

merely a minor addition to the production sequence, such as installing a weighbridge in the line : or they may, with the manager's approval, mean a temporary alteration to practices to give particular information, especially of critical values : or they may even be designed experiments, usually in co-operation with one of the specialist departments.

Before instituting new routines, however, it is worth taking a second look to see if the information is already available in some different form, and could be obtained by analysing or manipulating data already recorded. For example, when punched cards are used for production or sales analysis it is quite common for more information to be put on the cards than is actually used in preparing the published statistics—but it is there for the O.R. team if required. Again, if production weights at different stages are required, it may be possible to calculate them from known weights of scrap, without having to institute special weighings.

Simulation. Simulation, though a true source of data for prediction, cannot normally be used until the study is sufficiently far advanced for a mathematical model to be made. Then, by programming the model for a computer or, in simpler cases, by manual operation, a series of data can be obtained representing the behaviour of the system under different defined conditions. The advantages of simulation are that it does not disturb either current production or operatives, and statistics relating to months of operation can sometimes be obtained in a few hours or minutes.

Others. Besides these main sources, there are certain others to be taken into account. There are opinions held by the manager and his staff, based on their experience of the plant. Unless supported by documents these can only be accepted as indicative, but they are important and may be the only confirmation available to support another source of data. On the other hand, if they were at variance with that other source it would be even more important to confirm one or the other and, if the management's views were in error, to inform them as soon as possible so as to avoid the possibility of conclusions reached at a later stage being rejected on the grounds that the data are unacceptable.

There may also be data available in text-books, e.g. properties

of materials ; there are makers' specification for plant or machinery ; and there may be safety regulations or other agreements that set limitations to performance.

Methods of Collection

Method of Collection may refer to the physical methods of making the recording, or to a principle such as whether to collect in full or by sampling.

Data recording may be fully automatic, as in the case of a continuous reading instrument ; it may be semi-automatic in the sense of the actual recording being mechanical, or photographic, with the instrument controlled by an observer ; or it may be manual. Of these, the semi-automatic and manual are the most likely to be used, the former being particularly important when the data are related to a time base. For economy in taking (and analysing) observations, it is usual to concentrate only on recordings of interest or significance, and mechanical methods of recording—e.g. punching coded information direct on to teleprinter tape—allow more variables to be followed simultaneously. Besides, teleprinter tape, film and the like, carry the data in a form ready prepared for analysis on the appropriate machine.

Methods of collecting data, in the sense of sampling as opposed to full recording, depend not only on the accuracy required, but also on the availability of information. Sometimes the full information cannot be obtained, whereas a representative sample can ; sometimes the full information may be obtainable, but not within the time allowed—e.g. the full chain of events (pattern of production, etc.) is not expected to occur soon enough.

Market surveys, or consumer research, use sampling techniques ; so does Activity Sampling, as a source of Work Study data (which can be very helpful in O.R. as well). Another occasion when statistical sampling is useful arises when, at a comparatively late stage of an investigation, it is found that there are gaps—that important information is missing and there is not enough time to go back and mount another full survey. When it is not possible or reasonable to obtain every reading—say for reasons of time, or causing interference with production—one can combine detailed recordings of selected

parts with sampling of the rest of the field. In this way the accuracy of the statistical approach is verified, whilst conclusions on predictions that depend on the particular conditions selected are seen to be soundly based.

ANALYSIS OF DATA

Objectives

The first, preliminary, analysis of data may come quite early on when the team is still diagnosing the real problems. They may be fortunate and have hypotheses to test ; or they may, frankly, be looking for a lead to the next step. In the latter case the techniques may be very simple—little more than drawing graphs to search for trends, distributions or correlations. It will sometimes happen that the results of these preliminary studies are sufficient to suggest to management the appropriate action to take, so that further, detailed study is not required. Some case examples are given later, in the section on Numerical Analysis.

In general, preliminary analyses, together with qualitative information that has been collected at the same time, provide information on the behaviour of the system under different conditions and so help to determine what problems have to be studied. And this study, of course, extends to the examination of operating or economic conditions that may be, or ought to be, imposed in the future.

It is at this rather later stage that the formulation of objectives becomes very important, both to save time and to get the most out of the data. The particular objective indicates what data should be used, the accuracy required, the criteria of choice or decision, and often, too, the appropriate technique of analysis. Most objectives are special to the investigation, but there are two that are general to O.R., namely, construction of a mathematical model to represent the behaviour of the real system, and determination of the appropriate units of measurement in which to express this behaviour.

Establishing the units of measurement—which often requires invention—and the form of the model, shows how the data are to be treated. Treating data to yield particular values, or series of values, is Numerical Analysis, a branch of mathe-

matics in its own right : it should be distinguished from Data Processing, which refers rather to the equipment and its method of use than to the mathematical principles applied. Numerical Analysis is a wide subject that would justify a text-book on its own—and there are several in existence—but an outline of the field and mention of some of the more important methods is given later in this chapter.

Choice of Criteria

A criterion is a means, or standard, of judging ; and judgment is an executive function for which O.R. may be required to provide data or predictions. These data must be presented in such a way as to indicate the likely outcome of different decisions that the executive may make. So criteria must be established, first to measure the expected outcome of different decisions ; and secondly, to show, under different conditions, how to choose the best line of action to lead to a stated objective.

The most familiar criteria are the various measures of efficiency or achievement ; examples are profits, profitability, turnover, shortness of delivery dates, balance of plant, etc. In the context of O.R., however, these are usually given as management objectives. Their first impact on the investigation may be in establishing the units of measurement (mentioned in the previous section) that are fundamental to the system and reflect its progress ; another impact may be in their self-consistency—it is not unknown for management objectives to be inconsistent. On occasion, too, there may be some doubt as to whether the right criteria have been chosen ; these may follow from company policy and be inviolable, or they may turn out to be no more than suggestions that the manager would be willing to change if the advantage of alternatives could be demonstrated. Thus, if the given problem were to minimise maintenance downtime consistent with a stated minimum yield per unit of raw material, the team might come back with a suggestion that it would be more economical to waive the limitation on yield, and to install new plant for waste recovery.

There is another type of criterion, however, that lies at the heart of O.R., the criterion of choice or decision, the necessary and sufficient conditions for a particular line of action to be

optimum. Given the economic objectives, under what circumstances should the manager take action A, and when action B? And what are the critical conditions that determine the right moment at which to change? Fundamentally this is a problem of determining, for different circumstances, those parameters to changes in which the system is most sensitive—i.e. which should be treated as independent variables. Many of the techniques described elsewhere in this book lead to functions or relations to be used in decision making. In some cases there will only be one parameter that it is convenient to use as independent variable; in other cases there may be more than one.

When more than one parameter can be, or has to be, used as an independent variable there may be more than one criterion of choice. (It should be noted that this may have arisen because the appropriate units for that particular system were not discovered at an earlier stage. A suitable change of units—i.e. transformation of co-ordinates—may resolve the difficulty.) How, then, does one decide which criterion to use? If it is possible to give a general answer it is this: that the criterion to use, in any set of conditions, is that which minimises the information needed by the manager in making his decisions. The near optimal decision that can be taken at once may be far more valuable than the absolute optimum that takes time to compute.

Techniques of Numerical Analysis

In the section of this chapter on Objectives, reference was made to some of the simpler techniques that are used in the preliminary analysis of data. Of course, these techniques may also be used at later stages, and three of them are given here with case examples; others, requiring more familiarity with mathematics, are also outlined; references are given where the formulae and explanations can be found. Every analyst has his own pet methods, of course, and though the formulae quoted in this section are good all-round methods that will meet the majority of cases, the individual may prefer to use others that he can find in the reference books mentioned.

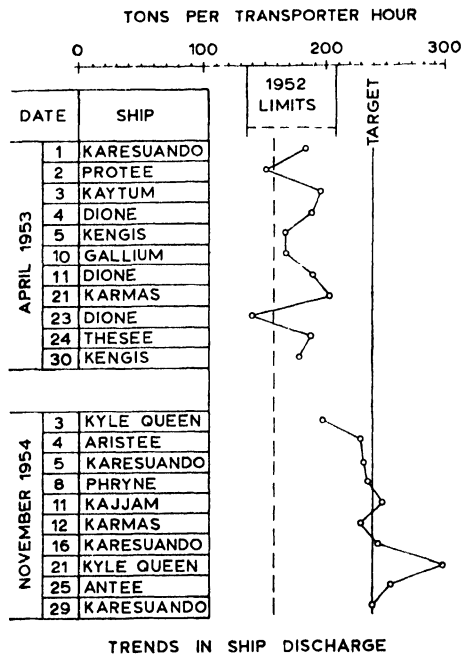
Analysis by Classification—one of the oldest and most powerful ways of examining large amounts of data—was used

in a study of turn-round times and delays experienced by lorries distributing part of a finished product. More than 1,300 lorries per week entered the works and considerable congestion occurred in the service bays. From the records of the Works Traffic Department, turn-round times for a sample period of three weeks were examined on frequency diagrams, first of all for uniformity of working round the clock, and then for each of the operations—such as reporting on arrival, weighing empty, travelling to bay, loading, re-weighing, collecting delivery note—and for each internal district within the works. The expected distributions were in fact found for the components and districts, with the exception of one bay for which the total turn-round times extended in a tail some three and a half hours longer than that of the other bays. The management then took suitable action in this bay, which was reorganised in such a way as to eliminate the congestion causing the tail.

For this type of comparison to be dependable, the situation must be static and the experience representative. Moreover, there may be unnoticed differential weightings between two classes which bias the comparison—such as, in the example given, different proportions of two kinds of lorries. This ties up with the third complication which has to be allowed for, namely, the possibility that there may be differences between classes other than the obvious ones associated with the basis for classification.

Analysis of Data for Trends is often necessary with the intention, say, of emphasising departures from the normal control limits. An analysis of this kind was made on discharge rates of ships (each of which is generally regarded as being unique and therefore, at first sight, not an obvious application for a control chart), and frequency histograms of the rates of discharge in tons per crane hour were made for all ships visiting a works' wharf in 1952. The remarkable observation was made that, in spite of several major differences in design and size, and in the density and physical characteristics of the ore carried, if the ships were divided into two classes only—the "good" and the "bad"—then symmetrical, and fairly small, distributions of discharge rates were obtained. 95% limits were obtained from these distributions and two control charts initiated.

For the early part of 1953 it was confirmed that the current discharge rates fell within these limits, and a master board (Fig. 11.1) was erected on the wharf showing the discharge rates for different ships ; instead of 95% limits, however, a red line was drawn—between the limits—to indicate that the results were expected to be above this line. An optimum target was also established, which was well above the then existing figures.



TRENDS IN SHIP DISCHARGE

FIG. 11.1

With the co-operation of the management and the men, and the introduction of new methods and machinery, discharge rates rose steadily. It is believed that the wall chart helped to confirm the pride of the skilled operators in the job they were doing, and it certainly indicated to the management at a glance the previous history for a year.

In passing, it might be observed that non-linear scales can sometimes help to clarify lessons not at all obvious on a linear scale. Accuracy of fit is not critical in the above illustration. If, on the other hand, changes in, say, repair costs to ships

over a number of years were being examined, then the trend might be small but important and an increased accuracy of fit is necessary. Another relevant point is that the object of searching for a trend may be to eliminate it from the data, and one example of this has already been quoted—the comparison of coke consumption in a blast furnace, per ton of iron produced, with the basicity ratio of the slag.

Analysis for Cycles may prove a useful approach to problems in which some sort of recurrent pattern of events is suspected. One problem studied in this way was that of derailments, to see if there was any well-defined seasonal variation—this was indeed found, with a maximum in the spring quarter. Another problem was that of eccentricity in welding rods : although some of this was expected to be random, it was thought that there might be cyclical variations related to the machine cycle. The initial measurements were graphed, the slight cyclic variation was confirmed and its periodicity and amplitude under different operating conditions estimated. A second series of readings was then taken at intervals in phase with the expected periodicity : comparison of these with changes in the machine parameters suggested which of them were important in causing the cyclic variation, and led to suggestions for its reduction and virtual elimination.

If possible, cyclic variations should be treated graphically as in the above example. It is usually far easier to fit an estimated curve to the points as plotted than to proceed to Fourier Analysis, or the use of Periodograms, etc.[2] As a rule, cycles have simple patterns and there is only a small number of distinct types, so it is very unsafe to infer a causal relationship between two cycles merely from a similarity in shape. If data are being analysed for cycles of known period (determined graphically, for example), then estimation of the means that form the phase components is a fairly simple matter : in correlating two time series with strongly marked cycles, the precision of the correlation does not increase much with the number of points (since each successive complete period tells the same story all over again), and one might as well just correlate the phase averages.

Interpolation and Extrapolation. When a function has been tabulated for particular observations and other values are

required, they may be obtained by interpolation (values within the range of observations) or extrapolation (values outside the range). The variable for which observations or readings were made is often time, and the series may be used to determine seasonal variations, short- and long-term trends, and finally forecasts—of sales, production, raw material costs, etc.

When the observations are evenly spaced (at equal intervals) the best-known formula is the *Gregory-Newton Formula* ([2], p. 10), based on the finite difference operator. There are several other formulae of value in particular cases, including more rapidly converging formulae based on central differences, but these require specialised knowledge.

If the intervals between the observations are unequal then one can always apply *Newton's Formula of Divided Differences* ([2], p. 24).

In any of these cases the accuracy of the estimated value can be determined by computing the magnitude of a " remainder term " in the series. Derivatives of Tabulated Functions can also be determined.

Difference Equations. O.R. problems, such as Queueing and Replacement, based on a " birth-and-death " process lead to Difference Equations, in which the value of a function $f(t)$ at time t, say, is connected with the values at other discrete times, say, $(t - 1)$ and $(t + 1)$. They are analogous to differential equations ; for example, the solution of a Linear Difference Equation with constant coefficients is the solution of the Complementary Equation plus a Particular Solution. In O.R. work, however, it often happens that the difference equations are more difficult and a knowledge of Transforms (e.g. Laplace Transform [3], Chap. XV) may be required.

The analogue of integration is summation, so that certain summation problems can also be solved in this way.

Numerical Integration may be necessary when only numerical values of a function are known, but not the function itself. The best-known method is *Simpson's Rule* ([3], p. 171), which is exact if the function is a polynomial of the second or third degree, and very accurate if it can be closely represented by such a polynomial in each interval of the range of integration. There are many similar formulae, but a powerful general method for a tabulated function is *Gregory's Formula for*

Numerical Integration ([3], p. 191), and the computation involved is straightforward and not too laborious.

Integration of Differential Equations. In effect, the numerical integration of differential equations is extrapolation of a function involving derivatives. There are two main classes of problem : first, Marching Problems, where the initial conditions are known and are " marched " forward at each stage of extrapolation ; and secondly, Jury Problems, where only some of the conditions apply initially, and the others have to be satisfied at later stages of the integration. This is a similar situation to that encountered in Dynamic Programming.

The simplest method, when approximation to the function by a polynomial of degree 3 is good enough, is *Milne's Method,*[4] but in the general case the most powerful method, convenient for automatic computation, is the use of central divided differences known as the *Deferred Approach to the Limit*[5] in conjunction with interpolation by proportional parts (see next).

Numerical Solution of Equations. The simplest way of solving algebraic or transcendental equations is by means of a graph ; in any case, some such start is required to locate the roots even when greater accuracy is wanted later. The best-known iterative method of solution is the *Newton-Raphson, or Tangent, Method* ([2], p. 84).

A more powerful method is *Inverse Interpolation by Proportional Parts (Aitken)* ([3], pp. 76, 97).

Nomograms. When the same equation (or set of equations) has to be solved frequently with different values of the coefficients, a standard programme can be prepared for an electronic computer. However, if graphical solution is sufficiently accurate it is often simpler (and always cheaper) to prepare a Nomogram.[6]

Least Squares. Least Squares is a method of curve fitting and, for any given curve, can be used to calculate the coefficients that give the best fit to observed or tabulated data. It is very important to understand that the method cannot be used to determine what curve should be fitted (except perhaps by a comparison of residual errors), and this remains at the choice of the user, being estimated from the apparent trend of the data when drawn on a graph.

When a relation of known form connects the variables, the

method of least squares can be applied to series of observations or measurements (liable to experimental error) to determine the best values for the coefficients in the relation. For further information on least squares see [2].

PRESENTATION OF RESULTS

The ultimate success of O.R., unlike that of pure research, lies not in the quality of the results themselves, but in their acceptance by management. Successful presentation of moderately good results can benefit the concern far more than the unsuccessful presentation of a brilliant study.

It is difficult to write on presentation other than by listing maxim and precept, and one craves the reader's indulgence for doing so, but they are so frequently disregarded. The reason for that is thoughtlessness, which is incompatible with good O.R.

The first principle to bear in mind is that the manager does not want the mathematician or scientist to justify his answer on theoretical grounds, he wants him to explain it, briefly and clearly, and to support his explanation with hard facts. He is not so much interested in whether the sums are right as in whether they are the right sums. No theory that is not an essential part of the explanation should appear in the text of a report, but should be put in an appendix. On the other hand, no text that is over-stuffed with supporting evidence can be made readable ; so the facts, too, must be relegated to an appendix (or perhaps left in the working papers !), and only a tabular synopsis put in the text to support the explanation. The question of presentation of results will be dealt with in more detail in Chapter XIV.

The second principle is that results and explanations should be put in the manager's own language. The importance of this lies partly in saving his time, which he values, and partly in demonstrating that the team do understand his problems and know what they are talking about (see also Chapter XIV.) A more important reason, however, is that whereas the report may go to a senior manager for acceptance, it will often be passed down to a more junior man for implementation ; without questioning this man's competence (and sometimes one might), nevertheless he may get the recommendations at

second-hand, and they must be written in the first place in such a way that his interpretation will be easy and correct.

The methods used in presenting numerical solutions have to be tailored to the particular problem. It is seldom that O.R. is asked for a specific single answer ; what is usually wanted is an ordered series of answers on which the manager can decide between different lines of action. The range, or limits, within which answers are required should have been agreed with the manager during the course of the investigation. If the problems only admit of discrete solutions these can be listed separately ; if the solutions are continuous, then graphical or tabular methods will probably be required.

Finally, as a postscript to the whole chapter, here are the rules that Sir Arthur Bowley drew up for the Oxford University Institute of Statistics, whilst acting as its Director during the late war :

1. If your figures are wrong to start with, no amount of adding and subtracting and calculating will produce right answers.
2. When you make tables make them clear and make them self-explanatory (e.g. always put in all the units and all the headings).
3. Percentages of percentages are anathema.
4. Don't use five places of decimals when all the figures will stand is one.

REFERENCES

[1] RIDGION, J. M., and WHITEHOUSE, A. M. (1956). " Studies of Blast-Furnace Assessment ", *Journal of the Iron and Steel Institute*, Vol. 184, November, pp. 249–57.

[2] WHITTAKER, SIR EDMUND, and ROBINSON, G. (1951). *The Calculus of Observations* (Blackie & Sons Ltd, London), 4th ed., Chapter IX.

[3] MILNE-THOMSON, L. M. (1933). *Calculus of Finite Differences* (Macmillan & Co.), 1st ed., 1933, Rep. 1951.

[4] MILNE, W. E. (1952). *Numerical Solution of Differential Equations* (Wiley).

[5] RICHARDSON, L. F. (1910). *Phil. Trans. Roy. Soc. A*, **210**, 307 (1927), ibid., **226**, 300.

[6] ALLCOCK, H. J., and JONES, J. R. *The Nomogram* (Pitman), 4th ed., 1952.

In addition to these, there are several modern books on Numerical Analysis that cover most of the problems referred to in the above, in particular

KOPAL, Z. (1955). *Numerical Analysis* (Chapman & Hall).
REDISH, K. A. (1961). *An Introduction to Computional Methods* (English Universities Press Ltd.).

Chapter XII

DYNAMIC PROGRAMMING

Introduction

THIS account is largely based on the work of Richard Bellman. In many papers and in his book *Dynamic Programming*,[1] Bellman has presented this approach to multi-dimensional problems, i.e. problems involving optimisations of functions involving many variables.

Dynamic programming is basically concerned with the theory of multi-stage decision processes. By this rather formidable phrase are understood processes in which a sequence of choices has to be made, each choice being between two or more possibilities. These processes may be of a deterministic nature, i.e. once a decision is made the outcome is uniquely determined, or stochastic in that the outcome is drawn from a set of possible outcomes according to some given frequency distribution.

An example of the deterministic type might be the periodic decision of a haulage contractor as to how many of his fleet of lorries he should replace. An example of the second is the amount of a commodity to be ordered on successive occasions by a retailer when the demand for the commodity is known only in terms of a probability distribution.

These processes may be concerned either with a finite number of stages or a continual process involving an infinite number of successive decisions. Similarly, the processes may be either discrete or continuous in that there may at each stage be either a finite number of choices possible or an infinite number.

Though natural multi-stage decision processes themselves arise sufficiently frequently for a theory dealing with their solution to be valuable, it is also often possible to regard other problems in this light and to obtain solutions more readily as a result of applying the dynamic programming approach. Thus the theory has a far wider area of application and this chapter will attempt to indicate its scope.

A simple problem will be solved by making use of the method of dynamic programming, and then the basic principle of the theory will be presented. Subsequently, more complex problems, which it is hoped will throw more light on the nature and scope of dynamic programming, will be considered, and indications given of its application to various practical problems. The relationship between dynamic and linear programming will also be discussed.

A Multi-Stage Deterministic Problem

To illustrate the principles of the theory and perhaps to clarify the type of problem to which it may be applied, consider the following simplified example :

Suppose a farmer one year harvests z tons of wheat. His next year's crop is obtained from sowing some of the present year's harvest, x tons sown yielding ax tons the following year, where $a > 1$. The wheat not sown is sold, y tons selling for £$f(y)$ and $f(y)$ is monotonic increasing, i.e. the more wheat sold the greater the revenue. If the farmer wishes to sell out after N harvests, what policy of selling and sowing should he adopt in order to maximise his total receipts from selling wheat over the period * (assuming yield and $f(y)$ constant from year to year) ?

First consider the problem in the following way. Suppose at the end of the first year the farmer sells y_1 tons and sows x_1 tons (and naturally $y_1 + x_1 = z$), then at the end of the second year he will have ax_1 tons to either sell or sow. Again, dividing these into x_2 and y_2, say, this gives $x_2 + y_2 = ax_1$. Continuing this process until the final harvest when ax_{N-1} tons are available,

Year	*Sells*	*Sows*	*Available the following year*
1	y_1	x_1	ax_1
2	y_2	x_2	ax_2
.	.	.	.
.	.	.	.
.	.	.	.
$N-1$	y_{N-1}	x_{N-1}	ax_{N-1}
N	y_N	0	—

* This is analogous to the economic problem—what proportion of income should be invested ?

the farmer will obviously sell the whole crop so that $x_N = 0$ and $y_N = ax_{N-1}$. This situation may be represented in the table on the previous page.

This gives the series of equations :

$$\begin{aligned} z &= x_1 + y_1 \\ ax_1 &= x_2 + y_2 \\ ax_2 &= x_3 + y_3 \\ &\vdots \\ ax_{N-2} &= x_{N-1} + y_{N-1} \\ ax_{N-1} &= y_N \\ x_N &= 0 \end{aligned} \qquad (1)$$

Eliminating the x_i from equations (1) above gives the following :

$$\begin{aligned} x_1 &= z - y_1 \\ x_2 &= a(z - y_1) - y_2 \\ x_3 &= a[a(z - y_1) - y_2] - y_3 \\ &\vdots \\ x_j &= za^{j-1} - (y_j + ay_{j-1} + \ldots + a^{j-1}y_1) \\ &\vdots \\ x_{N-1} &= za^{N-2} - (y_{N-1} + ay_{N-2} + \ldots + a^{N-2}y_1) \end{aligned} \qquad (2)$$

from (2) :

$$ax_{N-1} = za^{N-1} - (ay_{N-1} + a^2y_{N-2} + \ldots + a^{N-1}y_1) = y_N$$

$$\therefore\ za^{N-1} = y_N + ay_{N-1} + \ldots + a^{N-1}y_1$$

But the total value of wheat sold over the period is :

$$P = f(y_1) + \ldots + f(y_N)$$

Thus,

$$P = \sum_{i=1}^{N} f(y_i)$$

has to be maximised subject to the constraint

$za^{N-1} = y_N + ay_{N-1} + \ldots + a^{N-1}y_i$, and $y_i \geqslant 0$, all i.

Now, if $f(y)$ is other than a linear function (and if it is either linear or concave upwards the problem is trivial), this problem, involving as it does a maximisation over N dimensional space, is impossible to solve by the techniques of linear programming and is extremely difficult to solve using other classical methods.

Consider now how readily the problem breaks down by adopting an alternative approach. Define $g_k(q)$ as the total income obtained when starting with q tons and adopting an optimal policy over a k year period ; then

$$g_1(q) = f(q)$$

since if the farmer is in business for one year only he obviously sells all his stock at the end of that year. Now consider $g_2(q)$.

If the optimal policy is such that the farmer should sell y tons at the end of the first year then this gives $g_2(q) = f(y)$ + the return from $a(q - y)$ tons over the single remaining year.

But as $g_2(q)$ represents the optimal policy, the remaining $a(q - y)$ tons must clearly be utilised themselves in an optimal manner, i.e.

$$g_2(q) = f(y) + g_1[a(q - y)], \quad \text{if the } y \text{ is optimal.}$$

If the y is now considered as that which is to be determined this gives :

$$g_2(q) = \underset{0<y<q}{\text{Max}} [f(y) + g_1 a(q - y)]$$

a simple maximisation problem involving only one variable. This may be immediately generalised to give—

$$g_m(q) = \text{Max} [f(y) + g_{m-1} a(q - y)] . \quad . \quad (3)$$

But it has already been shown that $g_1(q) = f(q)$, so that it is possible to proceed step by step considering successive single variable optimisations to determine $g_2(q)$, $g_3(q)$, etc. Hence, the computation is very simple using the recurrence relation (3) and the problem may be solved whatever the form of $f(y)$. Though this method does, of course, involve computing in this example N separate functions, each of these, as has been pointed out, is very easily obtained and the total amount of the computation is very much less than that required when adopting other methods.

This provision of a simple algorism for computing the total return in a relatively simple manner is typical of the application of dynamic programming. The approach does, in fact, reduce the dimensions of a problem. The progressive build-up of the final solution in easy stages is characteristic of the basic approach of the method.

Before leaving this problem, it should be noticed that, in addition to determining the maximum return over the period, the best policy to employ at each stage (i.e. how much wheat should be sown) has been determined. It is also worth noting that, in computing the solution in this way, the solution to the same problem for varying numbers of stages and for different amounts of wheat is necessarily obtained, and thereby insight gained into the form of the solution. Thus, the problem has been embedded in a set of similar problems—this is also typical of dynamic programming.

The key to the solution of the above problem was contained in the statement : " clearly the remaining $a(q - y)$ tons must be utilised in an optimal manner ". This observation is an application of Bellman's " Principle of Optimality ".

The Principle of Optimality

Intrinsic in the above problem was the notion that if the farmer wanted to maximise his receipts with k years remaining, then he had to make the best decision (i.e. select the best values of x and y) and following this decision he had the next year to make the (then) best decision with respect to the remaining $k - 1$ year period and the wheat available at that time. Bellman states his " Principle of Optimality " in the following terms : " An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

A proof by contradiction is immediate.

This principle forms the whole basis of the theory of dynamic programming and enables problems of many dimensions (which may be otherwise intractable) to be reduced to problems of greatly reduced dimensions and often, as in the previous example, to one dimension. Consider now a problem in which the outcome of any decision is not uniquely determined, but

depends upon a probability distribution and see how once again the dynamic programming approach is able to shed much light on the problem. The problem to be considered is simpler than the previous one in that there are only two decisions possible at each stage, rather than many possible alternatives as in the wheat problem.

A Multi-stage Stochastic Problem*

Suppose the object is to utilise a bomber so as to inflict maximum damage on some enemy. The bomber has two possible enemy targets, A and B, to attack. A raid on target A will result either in a fraction r_1 of the enemy's resources in A being damaged or in the bomber being shot down (before inflicting damage) and it is known that the probability of the bomber surviving a mission to A is p_1. Target B is similarly associated with a fraction r_2 and a probability p_2. The enemy's resources initially are x at A and y at B. Then the problem is to determine what policy of attack should be followed in order to maximise the total expected damage to the enemy resources.

Let $f_N(x, y)$ be the expected damage if the optimal policy is followed when a maximum number of N raids is possible.

Then $$f_1(x, y) = \text{Max}\,[p_1 r_1 x,\ p_2 r_2 y]$$

Now suppose an N-stage process is begun by attacking A. Then the total expected damage is :

$$a = p_1(r_1 x + f_{N-1}[x(1 - r_1), y])$$

using the Principle of Optimality. Similarly, if B is attacked first the total expected damage is :

$$b = p_2(r_2 y + f_{N-1}[x, y(1 - r_2)])$$

and $$f_N(x, y) = \text{Max}\,(a, b)$$

If the number of raids is to be unlimited (except in so far as the aircraft may be shot down on any raid), the function $f_N(x, y)$ is replaced by the function $f(x, y)$ and the above equation becomes :

$$f(x, y) = \text{Max}\begin{matrix} p_1(r_1 x + f[x(1 - r_1), y]) \\ p_2(r_2 y + f[x, y(1 - r_2)]) \end{matrix}$$

* This is a shortened and rephrased version of Bellman's " Gold Mining " problem[1] (Chapter II of his book).

Note : It can be shown that if p_1, p_2, r_1 and r_2 are all less than 1, the equation for $f(x,y)$ has a unique continuous solution, bounded in any rectangle (x,y).

Now consider the case of unlimited raids.

For given values of p_i and r_i if the resources at each target are increased in the same ratio, the damage resulting from any policy is also increased in this ratio. The choice of policy is, therefore, not affected by a change of resources from x, y to kx, ky and depends only on the ratio $x:y$. It is clear that if, at any stage, when the resources are x, y, the best policy is to attack A, then it will also be best to attack A when the ratio of resources at A to those at B exceeds $x:y$. There must therefore be a value λ such that for $x:y > \lambda$, the best policy is to attack A, while for $x:y < \lambda$, B should be attacked.

To determine this value of λ, consider the condition when $x:y = \lambda$. When this applies either target may be attacked initially.

However, if A is attacked initially the point $[x(1-r_1), y]$, which is necessarily in the B region, is reached, so that the next attack will be on target B. Similarly, if B is attacked first the A region is reached and A must then be chosen for the next attack.

In each case the final point is $[x(1-r_1), y(1-r_2)]$, provided that the aircraft has not been destroyed. As the results of these two procedures must be equal, this gives the equation (considering the unbounded process) :

$$p_1r_1x + p_1p_2r_2y + p_1p_2f[x(1-r_1), y(1-r_2)] = p_2r_2y + p_1p_2r_1x + p_1p_2f[x(1-r_1), y(1-r_2)]$$

i.e. $$x/y = p_2r_2(1-p_1)/p_1r_1(1-p_2) = \lambda$$

Thus a rule is established which helps to decide which target to attack first, and though the final pay-off function has not been evaluated, all the information required at any stage in the process is available.

In this problem, therefore, a policy has been determined which can be used for making decisions, but no knowledge of the pay-off to be achieved from using this policy has been obtained. The implications of this will be considered in the next section.

Returning to the limited case, it may also be shown that

for a finite series of attacks the plane is again divided into two regions by a line $x = \lambda_n y$ where the value of λ_n is dependent on the number of stages (n). λ_n tends to λ as n tends to infinity and, in fact, for any p_i, r_i, there exists a number n_0 such that

$$\lambda_n = \lambda \quad \text{for } n > n_0$$

So the policy may be used in the infinite process as an approximation to that in the finite one.

The Mathematical Formulation of Dynamic Programming Processes

Deterministic Processes

Suppose a system whose ith state is defined by a vector $P_i = (p_{i_1}, \ldots p_{i_n})$. For example, in the wheat problem, the state variables (p_{ij}) are the quantity of wheat available, and the total income to date.

Let T_j be a transformation (i.e. the result of a decision on, say, how much wheat to sow) such that $T_j(P_i)$ giving a new state P_j, represents the effect of the decision on the system.

Then in an N-stage process there are a series of N decisions such that if P_0 is the initial state of the system—

$$\begin{aligned} P_1 &= T_1(P_0) \\ P_2 &= T_2(P_1) \\ &\vdots \\ P_N &= T_N(P_{N-1}) \end{aligned}$$

and the object is to optimise some function of the final state vector $R(P_N)$.

Defining $$f_N(P_0) = \operatorname*{Max}_P R(P_N)$$

where P represents all possible sequences $T_1 \ldots T_N$.

Then, in terms of these variables, Bellman's principle of optimality becomes

$$f_N(P) = \operatorname{Max} f_{N-1}[T(P)] \quad \text{for } N \geqslant 2$$

and $$f_1(P) = \operatorname{Max} R[T(P)]$$

the maxima are taken over all possible T.

This should be obvious from the previous statements of the principle of optimality, since clearly in order to optimise over

N stages, the first stage must be chosen so as to give a new state which, when optimised over an $N-1$ stage process, itself gives an optimal final state. This equation is the basic recurrence relation for processes of this type.

If the process is unbounded (i.e. if N tends to infinity) then the above equation may be replaced by the functional equation

$$f(p) = \operatorname*{Max}_{q} f([T_q(p)].$$

Stochastic Processes

Suppose the process under consideration is not deterministic, but the result of a decision is dependent upon a probability distribution function of the state vector. Then the result of a decision "q" on an initial state "p" gives a result in $(z, z+dz)$ with probability $dG_q(p, z)$. The distribution $G_q(p, z)$ of the transformed state is dependent on both "p" and "q".

Then to optimise the expected return, as a result of the decision "q" on the initial state "p" (this being considered as the first stage of an N-stage process),

$$f_N(p) = \operatorname*{Max}_{q} \int_z f_{N-1}(z)\, dG_q(p, z)$$

with

$$f_1(p) = \operatorname*{Max}_{q} \int_z R(z)\, dG_q(p, z)$$

where the expected value of $R(P_n)$ is maximised.

Again this can be expressed if the process is unbounded, in which case

$$f(p) = \operatorname*{Max}_{q} \int_z f(z)\, dG_q(p, z)$$

The Duality in Dynamic Programming Processes

The mathematical formulations above may make it more apparent that there are always, in these problems, two alternative ways of approaching the solution.

For while $f(p)$, (or $f_N(p)$) is a pay-off function, since it represents the optimal return, there is also the decision at any time $q = q(p)$ as a policy function.

Thus there is a continuous duality throughout the process.

(*a*) Finding the pay-off obtained from using the optimal policy.

(*b*) Determining the policy which gives the optimal pay-off.

This ability to use either space to obtain the best solution proves useful when solving the functional equations given above.

The Relationship of Dynamic Programming to Linear Programming

Consider the following example, which is very similar to the classic " nut mix " problem solved by the Simplex method of linear programming.

Each year the walnut crop consists of different grades, say, $G_1, G_2, \ldots G_k$, in quantities $Q_1, Q_2, \ldots Q_k$. Using various quantities of each grade, assortments of walnuts are put together for commercial sale at different prices. Assuming there are fixed demands d_i for the ith assortment ($i = 1, \ldots n$) and that each assortment mixes walnuts of different grades in its own fixed ratios, how many packets of each assortment should be made in order to maximise the total profit?*

Suppose the proportion of grade j in the ith assortment is p_{ij}. Then the total amount of grade j used is $\sum_{i=1}^{n} p_{ij}w_i$ where w_i is the weight of the ith assortment produced.

Then $w_i \leqslant d_i$ for all i, and $\sum_{i=1}^{n} p_{ij}w_i \leqslant Q_j$ for all j, and the object is to maximise $R = \sum_{i=1}^{n} r_i w_i$ where r_i is the selling price per unit weight of the ith assortment.

Now approach the problem using dynamic programming. If the first assortment only is produced, with quantities q_1, $q_2, \ldots q_k$ of each grade available, the maximum revenue under this condition is given by

$$R_1(q_1 \ldots q_k) = r_1 w_1$$

where $w_1 = \text{Min}\,(d_1, q_1/p_{11}, q_2/p_{12}, \ldots q_k/p_{1k})$. (i)

Introducing the second assortment, this gives, if a weight w_2 of this second assortment is produced,

Revenue $= r_2 w_2 +$ Revenue from producing assortment 1 with the remaining walnuts, $(q_1 - p_{21}w_2, \ldots q_k - p_{2k}w_2)$

and if $R_2(q_1 \ldots q_k)$ is the *maximum* revenue from producing

* This problem is quoted in [1] (Chapter IV).

only the first two assortments, then, using the principle of optimality,

$$R_2(q_1 \ldots q_k) = \underset{w_2}{\text{Max}} \{r_2 w_2 + R_1(q_1 - p_{21}w_2, \ldots, q_k - p_{2k}w_2)\}$$

where $0 \leqslant w_2 \leqslant \text{Min}\,(d_2, q_1/p_{21}, q_2/p_{22}, \ldots q_k/p_{2k})$

More generally

$$R_s(q_1 \ldots q_k) = \underset{w_s}{\text{Max}} \{r_s w_s + R_{s-1}(q_1 - p_{s1}w_s, \ldots q_k - p_{sk}w_s)\} \quad \text{(ii)}$$

where $0 \leqslant w_s \leqslant \text{Min}\,(d_s, q_1/p_{s1}, \ldots q_k/p_{sk})$.

Equations (i) and (ii) suffice to compute R_1, R_2. . . . R_n for any quantities $(q_1 \ldots q_k)$, and in particular

$$R_n(Q_1, Q_2, \ldots Q_k).$$

Thus the example can be solved by dynamic programming, although the Simplex technique will in this instance be much more direct. It is clear that any linear programming problem can be formulated in dynamic programming terms, though the linear programming technique will generally be more direct. However, the dynamic programming formulation of such problems is not bound by the linearity of the functions involved.

Considering the above example, the revenue might be a non-linear function of the amount produced. In this case, it is merely necessary to replace $r_i w_i$ by $r_i(w_i)$ in equations (i) and (ii), and the method of solution remains unchanged. Furthermore, the demands d_i might not be known exactly, but known to conform to some probability distribution. In this case, the problem could still be tackled by dynamic programming, using, for example, the expected revenue as that to be maximised. These various complications do not essentially affect the basic property of this type of approach, in that the problem is still reduced to a series of single variable maximisation problems.

In this ability to handle functions of a general nature lies the advantage of using dynamic programming in this type of problem, though it should be repeated that, where applicable, the linear programming techniques will generally yield the solution more readily.

An Optimal Stock Equation

Another class of problems to which dynamic programming may be applied is that of stock. For example, suppose the

demand for an item is given by

$\phi(s)\,ds$ = the probability of a demand within the range

$$(s,\ s + ds)$$

Consider the situation in which the stock is made up from the existing level x to a level y at fixed intervals. Suppose that the cost of ordering z items is $k(z)$ and that the cost of a stock-out of z items is $p(z)$.

Then if only one period is involved the minimum expected cost over that period is given by

$$f_1(x) = \underset{y \geqslant x}{\text{Min}} \left\{ k(y - x) + \int_y^\infty p(s - y)\phi(s)\,ds \right\}$$

while for n periods,

$$f_n(x) = \underset{y \geqslant x}{\text{Min}} \left\{ k(y - x) + \int_y^\infty p(s - y)\phi(s)\,ds + f_{n-1}(0) \int_y^\infty \phi(s)\,ds + \int_0^y f_{n-1}(y - s)\phi(s)\,ds \right\}$$

This simple equation yields recurrence relations sufficient to solve the problem as formulated for a finite stage process. Similar equations may be derived for cases when an unbounded process is considered and when time-lags occur in the delivery of items. For an account of the formulation of these equations the reader is referred to Reference [1] (Chapter V).

Some Practical Applications

To conclude this chapter, three examples of the application of dynamic programming to industrial and commercial problems are discussed.

(a) A Bidding Problem

Suppose a man is bidding for items of known value $v_1, v_2, \ldots v_n$, against a group of bidders each submitting a single bid for each or some of the items simultaneously.

From the records of previous auctions, suppose the distribution of winning bids w_i about the value v_i of the items is known.

Let $$\int_0^{w_i} \phi(w_i, v_i)\,dw$$

be the probability of winning v_i with a bid w_i.

If there is a limit C to the amount that the man can bid on all the items then

$$\sum_{i=1}^{n} w_i \leqslant C$$

where w_i is the amount we bid for v_i. The object is to obtain the maximum expected return $R_n(C)$ so that—

$$R_n(C) = \text{Max} \sum_{i=1}^{n} \left\{ (v_i - w_i) \int_0^{w_i} \phi(w, v_i)\ dw \right\}$$

Though for certain forms of ϕ it may be possible to solve this problem analytically, in general this will not be possible. However, a solution can always be obtained using the dynamic programming approach.

Considering one item only—

$$R_1(y) = \underset{w_1 < y}{\text{Max}} \left\{ (v_1 - w_1) \int_0^{w_1} \phi(w, v_1)\ dw \right\}$$

and it is easy to maximise this one-dimensional problem. Hence $R_1(y)$ is obtained.

Introducing the second item and employing the principle of optimality gives

$$R_2(y) = \underset{w_2 < y}{\text{Max}} \left[(v_2 - w_2) \int_0^{w_2} \phi(w, v_2)\ dw + R_1(y - w_2) \right]$$

and on the introduction of the ith item this gives the equation—

$$R_i(y) = \underset{w_i < y}{\text{Max}} \left[(v_i - w_i) \int_0^{w_i} \phi(w, v_i),\ dw + R_{i-1}(y - w_i) \right]$$

Thus it is possible to proceed step by step, computing $R_1(y)$, $R_2(y)$ and finally $R_n(y)$ and particularly $R_n(C)$.

(b) A Replacement Problem

*Problem.** In the manufacture of rubber tyres, a machine is used which simultaneously produces a tyre on each of two bladders. Sometimes a bladder fails while a tyre is being produced on it and then the tyre must be scrapped and the bladder replaced immediately.

In order to replace a bladder it is necessary to strip down the machine, and it may then pay to replace the other bladder

* This problem was initially posed by Sasieni.[2] The solution is due to Dreyfus.[3]

also while the machine is stripped down. In some circumstances it may even be worth while replacing both bladders before either of them bursts.

Thus there are four categories of cost :

c_1 the cost of replacing a bladder
c_2 the cost of scrapping a faulty tyre
c_3 the labour cost of stripping down the machine—this is the same whether one or both bladders are replaced
c_4 the cost of lost production time when a bladder fails during use.

The older a bladder, that is the greater the number of tyres which have been produced on it, the greater the chance of it failing. Thus the problem is to determine the best policy for replacing bladders.

Define

$f_N(i, j)$ = the total expected cost using an optimal policy of producing N additional tyres when bladders 1 and 2 have already produced i and j tyres respectively.

p_i = probability of successfully producing a tyre on a bladder which has already made i tyres. It is assumed that these p_i are known from records.

At each stage there are four choices—

1. Replace both bladders
2. Replace bladder 1
3. Replace bladder 2
4. Continue to produce without replacing either bladder.

Consider these four choices in turn :

1. If both bladders are replaced, the total expected cost will be the cost of replacing two bladders, $2c_1$, plus the cost of stripping down the machine, c_3, plus the total expected cost of producing N tyres, starting with two new bladders and using an optimal policy, $f_N(0, 0)$.

2. If bladder 1 is replaced, the total expected cost will be the cost of replacing one bladder, c_1, plus the cost of stripping the machine, c_3, plus the total expected cost of producing N tyres starting with one new bladder and one bladder which has already produced j tyres and using an optimal policy, $f_N(0, j)$.

3. If bladder 2 is replaced, the same costs as in 2 above are increased, but $f_N(i, d)$ replaces $f_N(0, j)$.

4. If production is continued, four occurrences are possible:

(a) a tyre may be produced on each bladder
(b) bladder 1 may produce a tyre and bladder 2 fail
(c) bladder 2 may produce a tyre and bladder 1 fail
(d) both bladders may fail

So, to find the total expected cost of continuing to produce, one has to calculate the expected cost for each possible occurrence and multiply by its probability, obtaining :

$$
\begin{aligned}
& p_i p_j f_{N-2}(i+1, j+1) \\
& + p_i(1-p_j) \operatorname{Min}[c_1 + c_2 + c_3 + c_4 f_{N-1}(i+1, 0), \\
& \qquad 2c_1 + c_2 + c_3 + c_4 + f_{N-1}(0, 0)] \\
& + p_j(1-p_i) \operatorname{Min}[c_1 + c_2 + c_3 + c_4 f_{N-1}(0, j+1), \\
& \qquad 2c_1 + c_2 + c_3 + c_4 + f_{N-1}(0, 0)] \\
& + (1-p_i)(1-p_j) \times [2c_1 + 2c_2 + c_3 + c_4 + f_N(0, 0)]
\end{aligned}
$$

This completes the analysis of the four possible decisions at each stage, and the optimal decision is that for which the total expected cost is least. It applies for all $N \geqslant 1$.

To compute this for any particular N first note that $f_0(i, j) = 0$.

So that $f_1(i, j) = \min$:

replace both $[2c_1 + c_3 + f_1(0, 0)]$
replace bladder 1 $[c_1 + c_3 + f_1(0, j)]$
replace bladder 2 $[c_1 + c_3 + f_1(\mathrm{i}, 0)]$
produce : $(1-p_i)(1-p_j)[2c_1 + 2c_2 + c_3 + c_4 + f_1(0, 0)]$
$+ [p_i(1-p_j) + p_j(1-p_i)] \times (c_1 + c_3)$*

Hence calculate $f_1(i, j)$, and then $f_2(i, j)$ using $f_1(i, j)$, then $f_3(i, j)$ using $f_1(i, j)$ and $f_2(i, j)$ and so on.

Using the values :

cost of bladder,	$c_1 = 10$
cost of scrapping tyre,	$c_2 = 1$
labour cost of stripping m/c,	$c_3 = 2$
cost of lost production time,	$c_4 = 3$

* Actually, there should also be a term $p_i p_j c_5$, where c_5 is the cost of producing a tyre which is not immediately required.

Number produced x :	0	1	2	3	4	5	6
Probability of success p_x :	1	0·9	0·8	0·6	0·4	0·2	0

Using the above recurrence relation the policy shown in the following table is obtained.

Policy :

$j\backslash i$	0	1	2	3	4	5	6
0	*P*–1	*P*–1	*P*–2	*P*–2	*R*–1	*R*–1	*R*–1
1	*P*–1	*P*–1	*P*–2	*P*–2	*R*–1	*R*–1	*R*–1
2	*P*–2	*P*–2	*P*–2	*P*–3	*P*–3	*R*–1	*R*–1
3	*P*–2	*P*–2	*P*–3	*P*–3	*P*–3	*P*–3	*R*–2
4	*R*–1	*R*–1	*P*–3	*P*–3	*P*–3	*R*–2	*R*–2
5	*R*–1	*R*–1	*R*–1	*P*–3	*R*–2	*R*–2	*R*–2
6	*R*–1	*R*–1	*R*–1	*R*–2	*R*–2	*R*–2	*R*–2

where *P*–1 means produce, and if a bladder fails replace that bladder only

P–2 means produce, and if the newer bladder fails replace both bladders, if the older bladder fails replace only it

P–3 means produce, and if either bladder fails replace both

R–1 means replace the older bladder immediately

R–2 means replace both bladders immediately.

The Standardisation Problem

This problem may arise in various guises, each amenable to a similar approach. The application of dynamic programming to this class of problems is demonstrated by the following example.

Suppose a buyer for an organisation has to provision for a certain type of equipment over a range of sizes. A demand for a given size may be satisfied by supplying a size greater than or equal to that demanded. The buyer is able, if he wishes, to take advantage of bulk-buying discounts, by buying large quantities of a few sizes. Given the demand for each size, and the discount rates, how many and what sizes should be purchased in order to minimise the cost of satisfying the demand ?

Suppose $f(w)$ is the demand for all items of size less than

or equal to w and that $g(m, w)$ is the total cost of purchasing m items of size w.

Assume (i) that $g(m, w)$ increases with w (i.e. that the cost per item increases with size) and with m.

(ii) that $\frac{1}{m}g(m, w)$ decreases with increasing m (i.e. that the cost per item decreases as the number purchased increases).

These two assumptions will generally hold in practice, and simplify the formulation of the problem. They are not necessary, however, for this type of approach.

The problem, then, is to balance the discount for bulk buying against the expense of providing a larger item than demanded.

If only one size of equipment were purchased, to satisfy demand up to size w, this would have to be the maximum size demanded, and the cost would be given by—

$c_1(w) = g[f(w), w]$ for all w in the range of demand $(0, w_{\max})$, say.

Suppose, now, two sizes are bought to satisfy demand for sizes up to w. If these sizes are w and w'

$$\text{cost} = g[f(w) - f(w'), w] + c_1(w')$$

Let $c_2(w)$ = minimum cost of provisioning for all items in the size range 0 to w, using two sizes.

Then $c_2(w) = \underset{w' < w}{\text{Min.}} \{g[f(w) - f(w'), w] + c_1(w')\}$

If $c_2(w_{\max}) < c_1(w_{\max})$, it is worth while buying two sizes rather than one.

In general, if $c_n(w)$ = minimum cost of buying n sizes to satisfy demand in the size range 0 to w, then, for all $n > 1$,

$$c_n(w) = \underset{w' < w}{\text{Min.}} \{g[f(w) - f(w'), w] + c_{n-1}(w')\}. \quad \text{(i)}$$

Thus it is possible to compute $c_1(w)$, $c_2(w)$, . . . , etc., and in particular, $c_1(w_{\max})$, $c_2(w_{\max})$, . . . , etc.

Now if $g(0, w) = 0$, this gives

$c_1 > c_2 > c_3 \ldots c_{N-1} > c_N = c_{N+1} = c_{N+2} \ldots$ for some N,

while if $g(0, w) > 0$, then

$c_1 > c_2 > c_3 \ldots > c_{N-1} > c_N \leqslant c_{N+1} < c_{N+2} < \ldots$ for some N

In either case, N sizes are required, the actual sizes being obtained in sequence from c_N, c_{N-1}, . . . , etc.

This problem is rather different from those considered previously, in that it is treated as a multi-stage decision problem, although the number of stages is not known initially.

As an example, take

$$f(w) = \begin{cases} w \text{ for } w \leqslant w_0 \\ w_0 \text{ for } w > w_0 \end{cases}$$

and $$g(m, w) = w(m + k)$$

Then, if only one size is being bought, $m = f(w) = w$, so that

$$c_1(w) = w(w + k)$$

and from (i), putting $n = 2$, gives

$$c_2(w) = \begin{cases} w(w+k) - \frac{1}{4}(w-k)^2 & \text{for } w > k, \quad \text{and } w^1 = \frac{1}{2}(w-k) \\ c_1(w) & \text{for } w \leqslant k, \end{cases}$$

Similarly, putting $n = 3$ into (i), and minimising, gives

$$c_3(w) = \begin{cases} w(w + k) - \frac{1}{3}(w^2 - 3wk + 3k^2) & \text{for } w > 3k, \\ & \text{and } w^1 = \dfrac{2w}{3} - k. \\ c_2(w) & \text{for } 3k \leqslant w \end{cases}$$

In general, for all $n \geqslant 1$,

$$c_n(w) = \begin{cases} \dfrac{n+1}{2n} w(w + nk) - n\dfrac{(n^2-1)}{24}k^2 \\ \qquad \text{for } w > \dfrac{n(n-1)}{2}k \text{ and} \\ \qquad w' = (n-1)\left\{\dfrac{w}{n} - \dfrac{k}{2}\right\} \\ c_{n-1}(w) \qquad \text{for } n\dfrac{(n-1)}{2}k \geqslant w \end{cases} \quad . \quad . \quad \text{(ii)}$$

Thus it is seen that the number of sizes required depends on the values of k, and w_0.

If $k = w_0/20$, then, from (ii)

$$n\frac{(n-1)}{2}k \geqslant w_0 \qquad \text{for } n \geqslant 7$$

so that $$c_n(w_0) = c_{n-1}(w_0) \quad \text{for } n \geqslant 7$$

Thus $c_6(w_0)$ gives a minimum cost, and 6 sizes are required. The largest is clearly w_0. The next largest is, from (ii), and putting $n = 6$,

$$w_1 = 17w_0/24$$

Putting $n = 5$ in (ii), and $w = 17w_0/24$, gives for the next size down,

$$w_2 \text{ (say)} = 7w_0/15$$

Similarly, $n = 4$, 3 and 2 give sizes $11w_0/46$, $2w_0/15$ and $w_0/24$.

Thus sizes are standardised as

$$w_0,\ 17w_0/24,\ 7w_0/15,\ 11w_0/40,\ 2w_0/15 \text{ and } w_0/24$$

REFERENCES

[1] BELLMAN, R. (1957). *Dynamic Programming* (Princeton University Press ; London, Oxford University Press).

[2] SASIENI, M. W. (1956). " A Markov Chain Process in Industrial Replacement ", *Operational Research Quarterly*, Vol. 7, No. 4, December.

[3] DREYFUS, S. E. (1957). " A Note on an Industrial Replacement Process ", *Operational Research Quarterly*, Vol. 8, No. 4, December (being a reconsideration of the problem posed by M. W. Sasieni (see above)).

Chapter XIII

SOME HUMAN PROBLEMS

Introduction

THE aim of this chapter is to outline and to analyse briefly human problems which are encountered by Operational Research investigators during the course of their work, and to suggest an approach towards their solution. Some of these problems should be considered as an inherent part of the whole situation which is to be investigated by the standard methods of Operational Research and as being one of the causal factors of the inefficiencies which O.R. is intended to remedy.

One of the main difficulties which arises in connection with O.R. is a definite and striking difference between the attitude of mind of a research scientist and that of a professional manager or administrator. This problem is very generally felt and a certain amount of work has recently been done on it. The results of this work seem to indicate an approach towards a solution of the problem. This point will be discussed in greater detail in the concluding section of this chapter.

For the purpose of the present analysis, human problems of O.R. can be divided into three categories : those existing before the investigation started ; those which arise during the investigation ; those arising from the completed investigation. Many of the problems belonging to the third group are closely related to those discussed in Chapter XV.

Human Problems Existing before any Investigation

It is well known that technical difficulties and inefficiency give rise to unsatisfactory relations and friction between persons and groups. When frustrated, people generally become unfriendly and unpleasant to each other. It is relatively easy to recognise the existence of inefficient transport arrangements, unsatisfactory scheduling of work, difficulties in flow, queueing and congestion. It is usually believed that when these are alleviated or remedied, improved morale will follow.

This is, however, not necessarily so. Cause and effect often tend to be confused. Although it is true that technical inefficiency usually causes bad morale and friction, in many instances the process is reversed. Unsatisfactory individual relationships and personal friction amongst senior management will invariably result in inefficient technical arrangements. Factors such as unreasonable rivalry between departments or individuals, unwillingness to delegate—or indeed a tendency to over-delegation, reluctance to accept two-way communication between different levels of management and labour ; all these will affect adversely the situation in matters which may appear to be purely technical.

If friction is caused entirely by technical defects such as an unsatisfactory transport system or deficient layout it may disappear completely when these defects have been removed. Should the unsatisfactory technical position persist for too long the morale may be damaged so badly that the mere attention to technical shortcomings will not be enough to improve it.

Take as an example the case of a small shipping office which worked quite efficiently until the volume of work suddenly increased. Each order was kept in a separate file located in a central filing system. In regard to each order a number of clerical assistants had specialised functions : one checked prices, another dealt with insurance, a third looked after shipping arrangements, and other duties were similarly distributed. Correspondence was sorted centrally and distributed to the clerical assistants, who took out the files when required.

The relations were quite amiable and the system worked quite well so long as the load of correspondence and other work was such that each file spent most of its time in a filing cabinet. Due to external circumstances, however, there was a sudden increase in the volume of business, and it came about that quite frequently two or more clerical assistants required the same file with considerable urgency. As things stood before the sudden increase in business, there was no need to keep a record of files removed from their cabinets. As the pressure became greater, more clerical assistants were employed ; this only increased the number of people requiring files at any one time and the bottleneck created by the filing

system was proving worse and worse from day to day. There developed a habit amongst clerical assistants of making a round of their colleagues' desks in search of files urgently needed. The irritation caused by constant interruptions in work became so great that people began to hoard files they needed and hide them under their chairs and in their drawers. In reaction to that, those of the clerical assistants who had more urgent queries on their hands developed the tactics of waiting until certain of their colleagues were absent from their desks in order to remove the files without their permission. In these circumstances, the increase in commercial opportunities resulted in the decline in efficiency of the office, and human relations deteriorated so badly that open quarrels over files became a daily occurrence.

To remedy the situation, a small central filing department was created under the supervision of a senior clerk. Files had to be booked out and in. A strict ruling was made that no document was to be passed between clerical assistants. All files had to be returned to the filing department for re-registration, and no member of the staff was to approach the desk of anybody else in his absence.

This, together with a number of subsidiary matters, resulted immediately in a remarkable improvement in the volume of work. The morale, however, was damaged so badly during the preceding period that the reinstatement of good personal relations in the office was much more difficult. The hostility and bitterness which had developed in the previous situation did not dissipate immediately and it now expressed itself as a strong opposition to the new system—in spite of the fact that it obviously worked. The new rulings were bitterly resented on the grounds that it took much more time to go to the filing department and have the files re-registered than to pass them over to a colleague who might be sitting at a neighbouring desk in the same room. Although this argument appeared to be rational, it was heavily emotionally charged. It did not seem very helpful to point out that the new arrangements resulted in improved efficiency. The clerical assistants who, during the preceding period, quarrelled amongst themselves over the files had now developed a common front against the head of the filing department, who was empowered to enforce the new

rules. Similarly, the consultant who had instituted the new procedure came under heavy and emotional criticism.

Fortunately, the man to be put in charge of the filing department was specially selected for his gifts of firmness, patience and tolerance. As time went on the hostility and bitterness generated during the previous period abated slowly—as there was nothing further for them to feed on. The new system enabled the office to take its advantage of the commercial opportunity and paved the way for further and more fundamental reorganisation. Good personal relations were gradually resumed and technical changes resulted in improved efficiency. If, however, the man put in charge of the new filing department had been less firm or tactful and less prepared to become a target for accumulated hostility, the new system might have had to be abandoned at the expense of a return to the previous chaos.

This brief analysis brings into relief the further fact that whenever an investigator or consultant enters a situation of accumulated hostility, he will become its focus. Similar strong feelings will centre on all individuals appointed to new positions created on the consultant's recommendations and on any person whose responsibility or power has been increased in consequence of the consultant's work. Only if this effect is fully anticipated and allowed for will the recommendation succeed. It is the duty of the investigator or consultant to ensure that people selected for these exposed positions, as well as himself, will take these emotional after-effects in their stride and by their behaviour will bring about gradual dissipation of hostility.

This example illustrated a case where bad morale was caused by the technical situation of an inadequate filing system. The investigator may often encounter problems which at first sight appear strictly technical, and he may feel that if these technical deficiencies are remedied, human relations will automatically improve. However, it cannot be too greatly stressed that the order of causality is frequently reversed. The following example will help to bring out this point.

In a middle-sized firm, the administrative authority was centred in the company secretary. An old servant of the firm, he was an efficient administrator who knew a lot about

production, general and cost accountancy and legal matters. He was largely responsible for pricing and for the day-to-day administration of the company. He was a very steady and reliable man, always arriving punctually at his desk in the morning ; he often stayed late and took work home, but hardly ever left early ; he rarely took days off. His lunch period was very brief and he often spent it at his desk—having light refreshments brought in.

The managing director of the firm had started work as a salesman, he had good success on the road and he came to be appointed sales manager—according to some people, purely as a result of good luck. Shortly afterwards, the previous managing director died suddenly. The board, composed exclusively of non-resident directors, were very favourably impressed with several intelligent suggestions of the sales manager and as a result he was appointed to be the managing director. He was a man of moods, sometimes sparkling and sometimes depressed, inconsistent and not a logical thinker ; he was always a very good salesman, a good representative and figure-head of the firm. His actions as an administrator were based on his varying and momentary moods. Some of his decisions were brilliant and based on a flash of insight. When he was depressed, however, he would often act without adequate knowledge and understanding of the situation.

Considerable enmity had naturally arisen between these two men. The company secretary felt that a man of lesser ability had been promoted over his head as a result of the casual approach of the board to making appointments. He was jealous of the managing director's social gifts and his flair for entertainment, and was only too keen to show up his shortcomings and mistakes in administration. The managing director in turn was afraid of comparisons being made between his own administrative ability and that of the company secretary and took delight in emphasising the latter's lack of salesmanship, his lack of polish or ability for public relations, and his difficulties in getting on with people.

Although superficially on good terms with one another, each man subtly undermined the other's prestige and interfered with his activities. Whenever the company secretary wished to introduce some new office procedure, it was possible for

objectors to have the procedure modified or rendered completely ineffectual by appealing to the managing director. Equally well, it was always possible to sabotage, with the co-operation of the company secretary, any decision of the managing director—who was not in very close contact with the day-to-day administration of the firm. The board did not intervene, as its members did not even know that this situation existed.

As a result, there were many schemes within the firm, inconsistent with each other, which had never been fully put into operation nor yet wholly withdrawn. On several occasions external consultants were employed at the initiative of either the one or the other of the two men. Their recommendations were never fully implemented, for they were never made acceptable to both parties.

Such a situation is bound to result in both impaired morale and technical inefficiency. The latter may appear, superficially, the primary reason and cause of the former. A consultant or investigator may optimistically embark on an investigation hoping that once the technical difficulties are removed human relations will automatically improve. However, in a human situation similar to the circumstances described in the last example, an Operational Research project cannot succeed, unless the role and causal importance of personal relations is fully recognised. Although no numerical or experimental methods are available, the human situation needs to be analysed and planned for with at least as much attention and intellectual effort as that devoted to any other facet of the investigation.

In the situation described above there are obviously several possible solutions. The removal of one or both protagonists is one possibility. Instances of similar types of friction have been alleviated by redefining responsibilities of the people concerned in such a way that there is no need for them to come into frequent contact ; consequently, opportunities for conflict are reduced.

Apart from feuds and friction between persons and groups, other primary causes of what ultimately appears as technical inefficiency are personal idiosyncrasies and personality problems of isolated individuals. One particularly interesting

situation which frequently occurs is that of inability or failure to delegate responsibility. There is a well-known type of manager who only with reluctance passes to his subordinates the handling of anything of relative importance. In some cases, a person of this type—if under pressure—will contrive to pass a fairly important task to more than one person without mentioning this fact to those involved. He will not explain fully the background or the objective, but will tell each person, separately, to get on with it. He is usually so anxious about the outcome of his actions that he insists on being consulted at every subsequent stage and development and makes certain that nothing can be done without his express authority. Some people of this description are even worse, for every time they are consulted they will do most of the talking, evade giving a decision and conclude by asking that the matter be brought before them again and supported by further documentation. The effect of this type of manager on his department is that an increasing proportion of his own, and of his subordinates', time is spent on these conversations. The work of the department slows down. Conscious of this fact, but unable to see its cause, the manager tries to increase departmental efficiency by increased personal control and intervention. As his conversations grow in frequency and duration, so the queue outside his office lengthens and the number of decisions declines. The process is self-generating and the vicious circle will gather momentum until something happens which brings it to a halt, or until it is slowed down by the production of pressure. This type of manager appears to be very busy whilst, in fact, he acts as a bottleneck in the chain of command and in communication—both vertical and horizontal.

A very similar type of personality problem is of particular concern to Operational Research. Many people are given to hoarding, and the tendency increases under pressure or stress. A factory or department run by a man of this type is usually over-stocked. There will be space difficulties and inefficient storage, leading to general inefficiency. These conditions lead to a feeling of being harassed—and hence to an increased tendency to hoard.

An O.R. recommendation that, in the interest of economy, the level of stocks should be reduced will trigger off or inflame

some of these irrational elements. The threat to the emotional security may be such that the recommendation may be violently and stubbornly opposed on grounds which are logically unsound. The remedy in these situations is not at all easy. One obvious solution may be the removal of a difficult person. Apart from extreme cases, this is not generally advisable ; in most situations, effective management training will provide the answer.

Human Problems encountered during an O.R. Investigation

Many people are afraid of change. Others have more or less vague apprehensions which result in a feeling that any investigation is tantamount to a reflection on their personal competence or efficiency. The very initiation of an O.R. project implies the likelihood of an impending change. A few people may be afraid, with good reason, of being shown up, of having their mistakes and deficiencies exposed and examined. Many experience inarticulate and irrational feelings of guilt or inadequacy, although in fact their competence and thoroughness are quite satisfactory—or even excellent. Although unfounded, these fears are further aroused and stirred by the very prospect of anything which resembles an inspection.

It is therefore important that no member of an O.R. team should behave as a criminal detective or a public prosecutor. Any project will be much more successful if the team is there to be helpful and not to find out shortcomings and to report on them. It will be easier to obtain information and to get the final recommendations accepted if all information asked for is treated as confidential, and assurances are given that no use will be made of it without prior consultation. Should a breach of confidence occur after a situation of trust was established, an O.R. investigation will suffer a considerable setback.

As little gossip as possible should be allowed to arise before, and during, an O.R. project. Gossip cannot be prevented by spurious secrecy—on the contrary, it is fed by it. It can, however, be supplanted by informed conversation based on adequate and truthful information. Since the nature of O.R. work is not generally familiar and involves a great deal of advanced technical knowledge, the onus for the spread of understanding is not with members of an organisation which

is being investigated. It is therefore the duty of an O.R. team to create informed opinion and to establish conditions preventing the spread of unenlightened rumours.

To achieve this end, a great amount of effort is needed which could be summed up briefly under the two headings of persuasion and education. All documents, letters and official reports should be very carefully drafted in order not to arouse the various fears and resistances discussed above. Equally, all lectures and other means by which the project and its separate stages are explained must be carefully prepared, with the same object in view. They must be fitted to the character of the audience and use such visual and other aids as may be necessary to make the basic ideas intelligible without over-simplification, and without giving the impression of a patronising attitude.

A small example will help to illustrate the amount of care needed when introducing an investigation. In one firm an O.R. project was introduced very clumsily : a notice was put up announcing the investigation about to be conducted and asking for all assistance and help to be given to the members of an O.R. team. After a sentence or two—full of long and unintelligible words—the notice contained the phrase " we are determined to avoid redundancy at all costs ". The word " redundancy " was one of the very few words used in the notice which was thoroughly familiar to most readers, and needless to say it was picked out from its context and gave rise to serious agitation. Much clearer English could have been used instead and the word " redundancy " ought to have been avoided altogether. What was necessary was to state, in plain English intelligible to most readers in that particular firm, that the purpose of the investigation was to increase productivity and to eliminate wastage of material, but not to reduce the labour force ; this would have been a true and more accurate statement of what was intended.

Over and above the usual and common fear of change and of being shown up, rational or irrational, there is a further aggravating complication connected with the use of scientific —and particularly mathematical—terminology. Most people are afraid of the unknown and unfamiliar, and they do not like being exposed to anything which may be difficult to grasp.

The vast majority have forgotten such skill as they might have acquired at school of even elementary mathematical reasoning, beyond plain arithmetic. They are naturally afraid of having to discuss, possibly at meetings, to criticise constructively and finally to make a decision on something where they feel that the grounds for acceptance or rejection are beyond their intellectual ability.

Furthermore, many people dislike intensely having to admit that they do not understand. If such an admission is not made, the subject is usually dismissed as of no further concern. It is much more difficult, emotionally, to admit that one does not understand, to ask for explanations and to persist until understanding is reached. These common features of human nature result only too often in a situation in which the investigator meets with aggressive but uninformed criticism. There is usually a defensive insistence on lengthy discussion of some subsidiary and unimportant issue, or even a point of niggling detail, which happens to be one of the few things fully understood and familiar to the insecure person who—on some rational or irrational grounds—is frightened of the results.

It is vitally important that these common human weaknesses should be anticipated and treated with patience, sympathy and understanding, but without a patronising attitude. At least one member of the O.R. team should possess some knowledge of these factors and have the requisite personal traits of tact and persuasiveness. It cannot be stressed too strongly that it is incumbent on the investigators themselves to make themselves acceptable, understood and appreciated. Unfortunately, a member of an O.R. team is no less likely than an average industrialist to be an insecure person. He may, and often does, experience a rational or irrational fear that his knowledge of industry is inadequate. His scientific and inquisitive habits of mind do not make for easy companionship with most other members of an industrial community, who prefer to adhere to their well-established routines, and for whom intellectual effort is hard work rather than relaxation or enjoyment.

As mentioned in an earlier section, a consultant or investigator is in an exposed position, being a natural target and focus for accumulated hostility. If, in addition, he happens

to be an insecure person, for reasons described in the previous paragraph, he is only too likely to adopt a defensive attitude. This all too often consists of the well-known recourse to unnecessary specialist terminology, or completely unjustifiable jargon. The use of unintelligible and impressive language as a protective measure strengthens the barriers and is conducive to attitudes of greater suspicion and animosity rather than to improved co-operation. The use of technical language in speech and writing is one of the aspects of a more general problem of a clash of attitudes between professional research scientists and professional managers or administrators to which further reference will be made at a later stage.

Methods of Introducing Operational Research

Generally speaking, there are two main modes of communication : written and spoken. Those useful for introducing an O.R. project can be subdivided in the following way. Written communication will include confidential reports and other documents ; introductory letters of informative and explanatory content ; notices and articles in works or departmental bulletins ; visual aids to written word, including graphs, isotypes, etc. Spoken communication will consist of such things as face-to-face interviews ; small meetings and discussion groups ; lectures to large audiences (including appropriate visual and other aids).

It is of great importance that the order and manner in which these techniques are used are suited to the particular circumstances of each separate project and to the changing circumstances which will develop as the project proceeds. Of these techniques, group discussion is of great importance and it will be referred to again at a later stage.

Conducting group discussion is a specialist skill which not many people possess, and for which unfortunately no formal training is yet available. To a large extent the same applies to the skill in interviewing, in using visual and other educational aids and the skill in public relations. It is vitally important that at least some members of an O.R. team should possess these skills, but in view of the lack of training facilities this is not easy to achieve.

Skill in interviewing comes to some people naturally, but

it does not necessarily go together with the type of inquisitive and scientific turn of mind most useful in Operational Research. There is in existence a method of training in interviewing (to which further reference will be made later) which can be adapted for the purpose of improving the approach of O.R. investigators to this problem. This is a matter of considerable importance in view of the fact that various antagonisms and anxieties discussed in previous sections, stemming from differences in personal attitudes, are mostly aroused in personal contact with members of an O.R. team and can be allayed by skilful interviewing.

There are in existence various forms of training in letter writing and report writing ; these could be extended further, developed and adapted to the particular needs of Operational Research. The use and preparation of visual aids is rapidly becoming a speciality of its own and an Operational Research team might be well advised to place their needs with such specialists, or to include them within their number.

Resistances to Investigation

As already mentioned, any type of investigation, inspection or examination, whether it be in an industrial setting or not, is apt to arouse in those who are investigated rational or irrational fear of being shown up, of change, fear of the unknown and the unfamiliar and of inadequate ability to adapt to new circumstances. If an investigation is undertaken at the request of some higher authority, but without invitation, wholehearted consent, or participation of those most concerned, it is likely to be considered as an intrusion into the privacy of various individuals and of the group. *Mutatis mutandis*, if an investigation originates at the initiative of middle or junior level management, it will be too restricted in scope and will not have the full sympathy of more senior levels, or carry sufficient weight with them. This is easily explained by the natural and common human tendency, at any level of hierarchy, to restrict and to minimise the scope or the importance of anything which has been initiated and transacted at any lower level.

The latter tendency will be particularly strong amongst some members of senior management who, sometimes with

good reason and sometimes irrationally, lack confidence in their own ability to manage or to keep the loyalty of their subordinates. It is natural that such managers dislike the very idea of their subordinates being interviewed systematically by fully or nearly independent investigators or teams and under conditions of freedom of expression and confidence.

In the conditions of modern civilised society, these various resentments and objections will not be voiced openly, but often hidden under an apparent and superficial desire to be helpful and to co-operate, and behind a façade of conventional politeness. The antagonisms and suspicions will, nevertheless, remain and will subtly sabotage the purpose of an investigation. It is therefore vitally essential that the investigators should be able, by means of skilful interviewing and other techniques, to obtain full and genuine co-operation (at all levels) of those concerned, and that they should be able to distinguish the whole-hearted and genuine willingness to help from the mere lip-service to it.

The unresolved resistances can result in two different types of effect, one of which may superficially seem beneficial, whilst the other is definitely not. The first effect consists in various shortcomings being corrected before they are discovered by the investigator, or immediately he has been observed to note their existence. The life of such improvements is generally only little longer than that of the investigation itself, since the basic cause of the inefficiency would not have been discovered and remedied. The superficial amelioration is due to an effort which cannot be maintained for long, and which is directed only at the symptom and not at its basic roots. When, after the conclusion of the investigation, the external stimulus bringing about the effort disappears the relapse into the previous condition is bound to occur.

Whatever the improvements, however, the investigator does not obtain the full and correct picture of the whole situation and he is led to believe that it is better than—or different from—what it actually is. In order to diagnose the position and to find the real root-causes of the difficulties, it is necessary to allow the technical and administrative machinery to continue functioning as closely as possible to the way it did before the investigation started. Due to fears, reasonable or not, of being

shown up it is very difficult to achieve the state of affairs where, in the presence of an O.R. team, an organisation will continue working exactly as it did before the arrival of investigators. An attempt should be made, however, to approximate to this condition as closely as possible, and this can be achieved only by inspiring confidence and assiduously avoiding even a semblance or suspicion of a breach of confidence.

The second effect of unresolved resistances consists in an attempt, often successful, to hide various shortcomings without making any effort to remove them ; to distort, or falsify, information given to the investigator, or to withhold it altogether. These tendencies can nullify any project. As before, the remedy lies in anticipating these tendencies and utilising the knowledge of human nature to inspire and retain confidence. It is important to remember that the slightest resemblance to threat, or an attempt to penalise in this connection, will only result in increased resistances. Again, the skill in interviewing is a basic prerequisite before much good work can be done.

The Participant-Observer Effect

Both types of effect mentioned in the previous section under the heading of Resistances to Investigation will result in the observer not receiving information, or receiving information which is not correct or not representative. There is also a further factor : some of those interviewed by an investigator may be particularly anxious to please him and to satisfy his wishes. This may be either because pleasing others is a deeply ingrained part of their psychological make-up, or because there are some ulterior motives. For example, the investigator may be seized upon as a potential ally in some conflict. In these circumstances, there will be a tendency to volunteer information which is systematically selected and exaggerated in the direction of what will appear to the interviewee to be the investigator's expectations.

There are a number of similar effects which cannot be discussed here, because of lack of space. They all result in distortion of information, and they can be compared to the effect of a measuring instrument on the system, a parameter of which is being measured. For example, if information about the temperature of a liquid is needed, a thermometer is applied

to it. The reading of the thermometer, however, does not give the temperature as it was before the measuring instrument was applied ; the value indicated by the instrument is the one resulting from the equalisation of temperatures between the thermometer and the liquid. In a similar way, the very presence of an O.R. team introduces changes into the organisation which is being examined. These changes in turn affect the information which is elicited and recorded.

In physical sciences and in engineering there are methods which allow a correction to be calculated and introduced into results obtained from measuring instruments. There are no similar methods in the field of social studies and it is, therefore, not possible to calculate the magnitude of error due to social or psychological factors. These errors cannot be corrected in a manner comparable to the way in which numerical correction is introduced into temperature readings. However, given knowledge of human nature and insight, these factors can be assessed with a fair degree of accuracy, and they ought to be taken into account when information is collected. Whenever possible, prior preventive action should be taken to avoid undue distortion of information, and correction or qualification of recorded data should be introduced when these effects occur.

Difficulties in the Final Stages—Weighing the Feasibility of Recommendations in Regard to the People Concerned

Many O.R. recommendations are based on completely correct information and on sound mathematical calculations. They may be fully practicable and of great value to efficiency. This, however, is not enough, for if these recommendations are not fully acceptable to those concerned they are not going to work. It is therefore important to anticipate objections and difficulties, both rational and emotional, to consider their strength when recommendations are made, and to develop methods of overcoming them.

The difficulties in accepting recommendations stem from essentially similar factors to those enumerated under the heading of Human Problems Encountered during an O.R. Investigation, and in subsequent sections. To summarise, and to state them more briefly and generally, the resistances may be

based on fear of loss of personal or group status, or of the increased status of others ; fear of incapacity to understand or follow new procedures or of inability to undergo a course of training ; fear of the working group being split ; fear of change. The O.R. team should not discount too lightly the influence and the value of custom and tradition.

If a man has invested heavily in developing a product which is not successful, he will not quickly abandon it. The more unsuccessful it becomes, the more likely he is to be stubborn and refuse to acknowledge failure. Similarly, when a group of people have invested time and effort in developing a set of ideas or skills, they react with unanimity and obstinacy to suggestions of change coming from outside the group. If the time and knowledge at the disposal of the O.R. team do not allow for fundamental approaches to altering group attitudes, it may well be necessary to alter the recommendations which are to be made.

It should not be inferred, however, that a recommendation should be submitted only if no resistance is anticipated. It is to be expected that any recommendation may cause some opposition, particularly if it goes to the root of any serious problem. What is important is that social skills be applied to creating opportunities for co-operation and participation and that the strength and nature of opposition be anticipated. When, because of circumstances beyond the control of the O.R. team, opposition is encountered, there are yet ways in which it may be dealt with. In some circumstances it may be justifiable even, neither to modify recommendations, nor to attempt to deal with opposition, but indeed to stall and not to proceed any further. A more opportune moment may arise at a later date.

In concluding this brief analysis of human problems in O.R. it may be appropriate to reiterate the basic point. No approach to improved efficiency can be fully scientifically valid or effective if it excludes a consideration of any factor of such importance as human problems, or if it does not utilise relevant techniques and knowledge. Three techniques are of particular importance here, namely, interviewing, group discussion leading and case-study sessions. These are briefly discussed in the following sections.

Interviewing

Inefficiency in a firm creates, or results from, deterioration in human relations. In such circumstances there exists an atmosphere of general distrust, often covered up by a superficial veneer of remote politeness, by studied informality or by pretence of hearty camaraderie. When undue ambition is fostered, a manager having any problem on his hands finds it exceedingly difficult to discuss it frankly and sincerely with anyone who is directly involved in the competitive situation. Inefficiency is a serious problem. It involves those both above and below the manager; it involves the manager himself; if he is aware of it he cannot discuss it.

It is here that the O.R. investigator or members of an O.R. team can satisfy a need. By virtue of their being outsiders not directly involved in the network of personal relationships within the organisation, they are in a position to obtain confidences and to be trusted not to abuse them. Moreover, they are likely to be less insecure and anxious than those under investigation and so able to listen and to decide what is being said. They can give constructive assistance.

This requires an appropriate attitude of mind and some skill in interviewing. People working in an atmosphere of suspicion and resentment begin to feel unreasonably guilty, even about the ordinary failings which are common to everyone. They react with anxiety and aggression to any idea of investigation, and, far from being welcomed, the investigator must expect initial antagonism. If, at interview, he reacts by exhibiting any trace of the attitude of a public prosecutor, any tendency to cross-examine, any sign of impatience, then he will have failed. At the same time, he must be himself, whilst remembering that the other person should be treated with the maximum sympathy and tact.

The first prerequisite is a natural or cultivated ability to listen. This is much more rare than is generally believed. Many who imagine themselves to be good listeners will find, if they hear their own interviews recorded, that they do most of the talking. This is because most people have an emotional need to impress others, to obtain affection, to extort a laugh, or to assert themselves in some other way. These needs can seriously interfere with one's ability to listen.

Once the ability to listen has been acquired, the manner of listening is important. Defensive blankness or passivity makes people ill at ease. A reaction of surprise, embarrassment or irritation leads to loss of confidence. People base their opinions and attitudes more often on emotional than on intellectual grounds. In interviews which have the purpose of finding out about things there is nothing to be gained by proving the speaker to be wrong in his interpretation or different from the interviewer in his opinions and attitudes. The need is for patience and understanding.

The investigator will bear in mind that subtle details of his own behaviour can influence the interview ; he must not show too keen an interest, he must not select from the information points for too intensive a verbal examination. The fact remains, however, that the course of an investigation can be influenced by his behaviour outside the interview situation. His purpose is to use the information obtained so as to increase the efficiency of the unit being examined. If, during the course of eliciting information, something is discovered which is disturbing, exciting or of great practical importance, not only must he not show this unduly or seize upon it selectively, but he must not use the information too quickly, too rashly, or in a way which will allow the source of the information to become apparent. The interviewer is in relationship with a series of individual interviewees, but they are in relationship with one another and with the events around them. The fact-finding interview situation is a continuing one. It must be based on mutual confidence and respect, on an approach to the understanding of the whole problem under review, and to ultimate remedial action based on that understanding.

Lastly, if the primary purpose of interviewing lies not in fact-finding, but in the sphere of modifying or changing the attitude of the person interviewed, then all that has been said—particularly the injunction that the investigator should do most of the listening—is still strictly relevant. A technique involving only individual interviewing is not, however, a sufficient way of altering attitudes, and it is unfortunate that limitations imposed from without have too often forced the O.R. investigator into this approach.

Preliminary and Private Discussion of Reports

Some or all of the results of an O.R. project will be, sooner or later, presented as a written report. Such a report is likely to contain suggestions involving change, and these are often based on, or follow, a description of an existing practice. To very many people this implies criticism of those responsible for the existing state of affairs. Whether this be a correct or incorrect interpretation, it is important for the investigator to realise that he will be held to be criticising and that the final result may hinge on the way in which he presents his findings and proposals.

If the report is intended to be critical—openly or by implication—or if it is anticipated that it will be construed as such, then before anything is put down in writing the investigator should make every attempt to broach matters personally with the persons or person concerned. It is in this way that on occasions co-operation may be secured, misunderstandings removed, proposals rephrased, and undesirable consequences obviated. Three common undesirable consequences of lack of discussion lie in the person criticised accepting the report superficially—but unwillingly—as a result of being antagonised ; in such a person refusing to accept the report ; and, last but not least, in the investigator making a report which, while technically correct, contains words unnecessarily and unwittingly damaging to the reputation of those concerned.

Criticism and plain speaking are made very difficult by the code of manners prevailing in this country. Both criticising and being criticised are experiences which are embarrassing in the extreme. The tenseness and discomfort associated with such a conversation may lead one or both of the participants to become defensive, or aggressive. It is for this reason that many investigators prefer to evade preliminary personal discussion of their results. They prefer to write a report in the first instance and defend it afterwards. Such an attitude presupposes and even engenders the development of a fight, and as a result final decisions are made in an atmosphere of sudden strife.

Training in Interviewing

In view of the importance for Operational Research of the

skill in interviewing, it will be relevant here to discuss some of the ways in which this skill can be acquired, taught or improved. The term "interviewing" refers here both to the initial gathering of information at the beginning of an investigation and to the preliminary discussion of recommendations and findings at the concluding stages before they are presented in a report form. In fact, "interviewing" will be taken, for the purpose of the present chapter, to denote all face-to-face conversations conducted by an O.R. investigator with a member of the organisation which is being examined and with reference to an O.R. project either in hand or concluded. The term will also be taken to embrace face-to-face conversations in connection with the introduction of a suggested O.R. project, or a conversation whose purpose is the discussion of a possible need for, and the desirability of, an investigation.

The method of training in interviewing to be discussed briefly consists of staging and recording an interview which is being observed by a training group. Members of the group are given a set of documents describing a situation which closely resembles circumstances in which they have already found themselves, which they have observed at first hand, or which they are likely to encounter. Some of these documents describe the organisation, the general and specific problems which arise in it, and they also contain detailed pen-pictures or character-sketches of the people connected with the main problem. Other parts of the dossier given to the members of the training group will contain letters, memoranda and other material similar to what they might have to handle in the course of their daily practice.

The dossier is planned in such a way as to lead to a natural conclusion that an interview is desirable with one of the persons described in it in order to discuss, or to attempt to settle, the main problem. One of the members of the group is selected to conduct this interview and he is requested to deal with the situation in his own way. Before the interview is arranged, there may be a general discussion by the group of the various issues and problems arising from the dossier.

On the appointed day a person will arrive who, preferably, is not personally known to the members of the group, to act as the interviewee. In order to impersonate adequately the

man described in the dossier the volunteer is specially chosen so that he resembles both in personality and status the person portrayed by the character-sketch, and is familiar with the general situation. The interviewee is given his own dossier—different from that received by members of the training group. His brief describes the situation as seen from his point of view and through his eyes, being the sort of person he is and occupying a specific position in the executive hierarchy of the organisation described in the set of documents.

The interviewee does not see, before the session, the dossier given to the members of the training group. This is prepared so as to describe the situation as seen from a completely different point of view from his own. The members of the training group, similarly, are not allowed to see the interviewee's brief before the exercise, as this would give them more information than they would ordinarily have about his attitude and frame of mind and about the real nature of his difficulties or objections. The factual information given in both sets of documents is not necessarily identical. Thus the interviewing situation resembles closely what happens in real life where the two parties to a difficult interview are in possession of different, and often apparently conflicting, factual information and where, due to previous history of the problem, and due to their own background and position, their attitudes to the issue in question are at variance.

The interview is recorded and, after the conclusion of the main part of the exercise, the interviewee leaves the room and the interviewer is questioned about his tactics. Subsequently the interviewee is questioned. The members of the group are asked to read the interviewee's brief and, in the light of what transpires now, the recording of the interview is discussed under the guidance of the instructor by both parties and by the whole group in order to discover to what extent the interviewee's frame of mind, his attitude and his difficulties could have been anticipated and allowed for more fully.

An exercise of this nature throws into relief not only personal characteristics of individuals and the ways in which these affect their interviewing, but also the fundamental and still not fully understood differences in attitudes between various professional groups. These differences, if allowed to continue unresolved,

are bound to lead to clashes which with greater understanding of these matters could be more easily avoided.

Group Discussion

Rather than by individual interviews and discussions a number of important issues centred around Operational Research can be better clarified by being discussed in small groups. Group discussion leadership is a specialist skill which requires considerable natural gifts of intelligence, quick-thinking and patience, and which arises from application and experience. Given this skill in the leader, but not without it, a well-organised series of group discussions can substantially help to resolve the emotional tension and conflicts created by an investigation in progress and to bring about modifications of attitudes required by the impending changes in organisation, mode of work or equipment.

It is important to realise that successful group discussion, in the meaning adopted in this section, is different from all other more generally known forms of meeting. The other familiar types of discussion include the sequence of arguments and counter-arguments, apparently purely rational ; an interplay of proposals and counter-proposals in negotiation ; an exchange of views, limited in time and scope by some agenda, and controlled by a chairman, when decisions are made ; the succession of question and answer after a demonstration or a lecture, or on other occasions when an expert is present.

The main feature of group discussion in the present sense of the term—which is little understood by those well accustomed to other forms of meeting—is that the leader, whilst controlling the procedure to a greater or lesser extent, refrains from dominating it or expressing his own views, and that he encourages the expression of any opinion, however unreasonable it may appear at first. As a result of skilled leadership of this sort, a permissive atmosphere arises within the group. The participants are enabled to examine with relative calm their own and each other's prejudices. When an opinion has been expressed with reference to a point which the leader considers to be of importance he will draw out other members of the group to challenge it. He will not allow the discussion to shift ground until some of the obstinacy or heat generated by

the expression of an apparently unreasonable or dogmatic view has been dissipated enough for a cooler and more open-minded exchange of reasoning to take place. No contributions to discussion are ruled out as irrelevant, but their relevance itself is discussed and the leader decides, subject to the feelings of the group, which issues should be taken up first and which can be postponed to a later stage. The silent members of the group and others who happen to have special knowledge or views of a given subject are drawn out by the leader. There is no need for the group to arrive at any definite decisions or conclusions. The purpose of the discussion is a purely exploratory ventilation of views and attitudes, exchange of information and a calm consideration of various possible arguments and different practical solutions.

In this way, the participants become informed more fully about each other's opinions and attitudes and about the reasons for them. They arrive together at new opinions, attitudes, decisions or conclusions which are not imposed on them, but which arise naturally from new and genuine convictions acquired as a result of the interaction within the group. The effect of these changes in conviction is much more lasting than anything achieved by an instruction or a decree from above or by direct forms of lecturing or persuasion. In the context of Operational Research, group discussion should be employed in the initial stages so as to air the basic attitudes towards Operational Research and to deal with the tensions of a developing investigation. At a later stage, actual problems arising within the organisation may be examined in the same way. During the course of an investigation there are usually requests for factual information ; such information without specific recommendations should be offered by the O.R. team at subsequent meetings and discussed in a similar manner. Further tentative recommendations can be put before discussion groups in order to seek their criticisms and in order to allay their own apprehensions. From this, the discussion groups will naturally proceed to request and to receive definite recommendations.

When these definite recommendations are finally formulated and put before the discussion groups, they are likely to be—as a result of the whole process—completely acceptable. In con-

sequence, the groups will be able to discuss calmly, dispassionately and constructively the obstacles in the way of implementation of these recommendations and the problems which might arise when they have been carried out. Thus at each stage it will be possible to obtain advice and active co-operation of a fairly large number of people who otherwise may not have been very forthcoming, or even positively antagonistic to a project in progress.

The Use of Case-Studies

Discussion of any issue which happens to be a topical problem in a given organisation is bound to proceed in the atmosphere of an emotional fog. The feelings about such issues are strong, the opinions differ and are highly controversial. The prejudices and emotions may be hidden behind one of the many possible façades such as customary politeness, cool and distant aloofness or studied and superficially friendly informality. Any one of these façades interferes with good communication and makes a frank and adequate discussion very difficult, or even impossible. Guarded and diplomatic expression of strong views is apt to be enigmatic, to be misconstrued and misunderstood. In consequence, a stilted discussion of matters which arouse strong feelings is only likely to confuse these issues further.

When a discussion of emotionally charged issues is forced on a reluctant group, there may be a breakdown of defensive façades resulting in unpleasant altercations, or even open quarrels. Members of the group may become unnecessarily harsh and unjustifiably critical of each other. As discussed earlier, the hostility aroused within the group may become focused on the discussion leader. Unless his skill can rise to the occasion, so that he can make good use of these emotional processes and utilise them fruitfully, such a discussion can do more harm than good.

One way of minimising and obviating these dangers is to present for the discussion suitably prepared case-studies rather than the immediate practical issues. The case-studies should be so prepared as to raise problems similar to those existing in the given organisation. The circumstances described in these case-studies should not be identical with those under

which the members of the group work, but, while being different, these conditions should be sufficiently close to strike them as being familiar.

The advantage of using such case-studies, carefully selected or specially prepared for these occasions, is that it arouses the interest and stirs the emotions sufficiently to demonstrate their existence and to make the discussion worthwhile. At the same time, the emotional pitch is not quite as high as when real issues are discussed and therefore the dangers briefly discussed in the preceding paragraph are not equally imminent.

It must be stressed at this point that both preparing such case-studies and conducting discussion of them is a specialist skill of its own. Without previous experience or appropriate training in this field, no attempt should be made to use these techniques in a delicate situation where inexpert handling might quite easily damage the prospects of an O.R. investigation. On the other hand, with sufficient knowledge and skill, these techniques may be directly instrumental in steering a difficult project to a successful end.

Suitably prepared case-studies lend themselves to be used as a background for practical exercises in interviewing as described in the previous section. Such an exercise, when preceding the group discussion of a case-study, adds considerably to its interest and liveliness. These techniques, therefore, when used conjointly have a much greater practical effect than any one of them on its own is likely to do. When the main purpose of the procedure is to clarify a management problem, the staging of an interview should not be construed as aiming at training in interviewing. The *raison d'être* of the practical exercise is to provide a live demonstration of what is likely to happen under the circumstances described in the case-study. The general experience is that such a practical demonstration drives home many more points than a discussion without it—or any amount of straight lecturing—will do.

Scientists versus Managers

Reference has already been made to certain differences in the attitudes of mind of research scientists as compared with professional managers. These are very generally felt and produce a certain amount of friction. A research scientist—and

this term includes here an O.R. investigator—must have an absorbed and almost dedicated intellectual interest in the problems he is dealing with, for their own sake. He must be enthusiastic about the intellectual rather than the commercial aspects of his work, and he must by nature experience his own satisfaction in the solution of the problems he uncovers as one of the principal rewards of his toils. This is not to say that he need not, or must not, be interested in his own status, the amount of his remuneration, or the prestige and standing of the organisation he is working for. However, if he allows any one of these to overshadow his basic intellectual enthusiasm and interest, his value as an original research scientist will have gone.

The same statement applies in reverse to a businessman or a manager. To be successful, he must have an absorbed and powerful interest in the commercial success of the venture he is working for. This does not mean that he should not have any other interest nor that the intellectual aspects of his work should be of no concern to him. On the contrary, a greater intellectual understanding of his work may essentially contribute to his efficiency at it. If, however, he allows the intellectual interest of his occupation to overshadow his sense of and his desire for commercial achievement, his value as a businessman or as a manager will come to an end.

However, it must be remembered that there are many positions whose occupants—in order to be fully successful—must be able to combine both attitudes to a considerable extent. For example, the manager of a research department in a large firm needs to be both a research scientist and a businessman. The balanced outlook is, nevertheless, difficult to achieve and most people come on one side or other of the dividing line.

In certain circumstances, the differences and conflicts become particularly pronounced. Young people straight from university often worship the idea of pure or fundamental research, and although they may be paying lip-service to the realities of a commercial situation, their actions show hardly any understanding of them. Similarly, people hardened by years of business experience may be able superficially to acknowledge the need for more thorough and more fundamental examination of the assumptions and traditions on which

difficulties arising from the differences in outlook, and their discussion needs expert handling. How fruitful discussion of such case-studies may be was abundantly shown in practical application of these methods. References to literature describing this work in further detail are listed below.

BIBLIOGRAPHICAL NOTE

A considerable amount has been written about problems of human relations in management, and a certain amount about the problems arising from the differences between the outlook of a research scientist and that of a general manager. There is also a growing literature on management training. However, very little detailed work has been published so far on any methods of practical training directly relevant to the problems discussed here.

The method, referred to above, of resolving or alleviating some of the differences of attitude between research scientists and administrators, in its application to the British Civil Service, has been described in fair detail in *Public Administration*.[1] The same article contains also a full example of a case study and a consideration of the technique of conducting group discussion, both used for this purpose, as well as a theoretical analysis and background of some of the problems involved. Considerations of the use of case studies and of the methods of increasing their impact have been presented by Avice V. Turnbull.[2] Further description of these methods of training can be found in the references [3], [4] and [6]. In particular the technique of conducting case study discussion has been considered in [1] and [2], and the technique of training in interviewing in [5] and [7].

Theoretical analysis of the type of human problems relevant here has been developed more fully in [8] and [9], in a way especially adapted to audiences trained in natural sciences or technology, with whom this analysis can be used as a theoretical background to practical training.

REFERENCES

[1] TARKOWSKI, Z. M. T., and TURNBULL, AVICE V. (1959). "Scientists Versus Administrators: An approach towards Achieving Greater Understanding", *Public Administration*, Vol. 37.

[2] TURNBULL, AVICE V. (1960). "The Use of Case Studies in the British Civil Service, and Further Developments", *Case Study Practice*. London, British Institute of Management.

[3] TARKOWSKI, Z. M. T. (1958). "Training in Human Relations—1", *BACIE Journal* (British Association for Commercial and Industrial Education), Vol. 12, No. 1 and No. 2.

[4] TURNBULL, AVICE V. (1957). "The Use of Case Study in the British Civil Service", *Public Administration*, Vol. 35.
[5] TURNBULL, AVICE V. (1958). "Practice Interviewing by Role-Playing", *BACIE Journal* (British Association for Commercial and Industrial Education), Vol. 12, No. 2.
[6] TURNBULL, AVICE V. (1958). "Techniques of Training in Human Relations", *BACIE Journal* (British Association for Commercial and Industrial Education), Vol. 12, No. 3.
[7] TARKOWSKI, Z. M. T. (1958). "Training in Interviewing", *H.M. Treasury Training and Education Division News Letter*, No. 4.
[8] TARKOWSKI, Z. M. T. (1958). "Understanding the Behaviour of People at Work: Introductory Notes for Engineers and Executives"; reprinted from *Engineering*, January–June.
[9] TARKOWSKI, Z. M. T. (1958). "Models in Physics and Psychology", *The Times Science Review*, No. 28.

Chapter XIV

PRESENTATION AND IMPLEMENTATION OF THE RESULTS OF OPERATIONAL RESEARCH

General

IMPLEMENTATION of Operational Research starts at least as soon as formulation of the problem. Anything later is likely to be unsuccessful, and indeed it is quite probable that a decision to undertake Operational Research is engendered by a desire to implement recommendations for improvement. In that case the first preliminary to implementation must have been taken before a decision to undertake any Operational Research. Conversely there is little charm in undertaking work which is unlikely to be implemented. Human problems (see Chapter XIII) are important in this connection. Permission to carry out a study is only of value if it implies an interest in results.

It is therefore necessary to carry this question of implementation continuously in mind during the whole course of an investigation. In the early stages of formulating a problem for solution the question of implementation must always be in mind. This is particularly important when the problem is a specific one.

It has been stated elsewhere that Operational Research investigations may be undertaken on a vast scale, not because of any specific symptoms which appear to require study for cause and cure, but merely on the grounds that such a study is bound to throw up interesting and useful information for future action. This is rather on the principle that the over-45's should have their eyesight tested every two years, and that regular visits to the dentist are necessary and useful, and that an occasional check-up of the system by a competent practitioner is desirable. In these cases it may be that there is nothing to implement, although such experience as has been gathered to date suggests that this is somewhat unlikely.

In the case of the specific problem, implementation may in

fact be more difficult than obtaining a mathematical solution. A remedy which is difficult to apply may be worse than no remedy. A blunderbuss remedy also may be undesirable, and a remedy which applies only in abnormal circumstances is not good.

It is also frequently true that a problem may have no single answer; in mathematical parlance the answer to a problem may be a function of time or of demand, or of both.

There may also be solutions depending on which of a number of different policy decisions are taken. This may seem at variance with the repeated view that Operational Research is concerned with policy decisions. However, some policy matters cannot readily be expressed in numerical form. A policy decision, for instance, to reduce a range of products by simplification may lead to quite different decisions on necessary stocks, from the results of a decision to increase a product range. A policy decision to try and attract more business by giving more rapid delivery will certainly be followed by an increase in the recommended stock-holding.

It is generally undesirable to give rigid solutions, quite apart from considerations that there may be no single answer. There seem to be few circumstances where a clear-cut answer can be given—unless the problem has been originally specified in such broad terms as to be virtually free from restrictions. Some problems set in mathematical form in the field of applied statistics can lead to single simple answers. These are, however, rather trivial cases in the policy field. In the field of design, for instance, it is very rare indeed to be able to meet a requirement exactly, and all answers are subject to limitations, but it may be that different answers are subject to different limitations and, in these circumstances, the recipient of the results of the investigation must be offered alternative solutions from which a selection can be made.

There are other circumstances also in which alternatives should be given; a solution, for instance, which yields an optimum in the form of an awkward numerical value is never very satisfactory. Fortunately, there is usually a nearby round number which, calculation suggests, is pretty well as good. Indeed, it is very dangerous to suggest that a number of four different integers is ever a good solution. It suggests that the

answer is as good as the calculation, whereas, of course, the answer is only as good as the original data.

A suggestion that the optimum quantity of material to order is in lots of 2·7183 tons, is liable to excite derision. A suggestion that something in the region of 3 tons appears to be correct, however, is likely to be received quite well, especially if coupled with the remark that it does not much matter whether $2\frac{1}{2}$ or 3 tons are ordered at a time, but that anything outside this range will be more expensive, can be readily appreciated.

There are other circumstances, also, in which great care has to be taken in recommending action. It may well be that two solutions can be offered, one of which can be readily implemented and the other, although somewhat better, may be very difficult indeed to implement. It may, for instance, require a major organisational change. It may involve an element of risk which, although small, could lead to disaster. In either of these cases, the remedy most readily implemented would appear to be the first choice. This does not, of course, suggest that recommendations for action should not be given if they are liable to cause upset. After all, if existing policies or organisation have resulted in an undesirable state of affairs, it is not unreasonable to suggest changing either or both of them, and the fact that the only remedy proposed may require some major surgery is unfortunate.

Similarly, if two courses of action respectively require little or large capital expenditure (with the latter the more attractive proposition) a decision may be taken on grounds of calls on available capital *vis-à-vis* other schemes rather than on the attractions of the two schemes postulated.

Report Writing

In the course of carrying out an investigation, not only must it be borne in mind that the results must be capable of implementation, but they must be presented in a form which makes implementation practicable. For this purpose both the spoken and written word are necessary. It is not unknown for there to be a desire to implement results which have not been proven, and, therefore, some care must be exercised in the use of the spoken word. Nevertheless, discussions at all times—of the

nature of the problem, the validity of the data and hypotheses as to calculations—are essential. It is not, however, judicious to give the results of an investigation before they have been checked. When they have been obtained it is quite good to propose them on a tentative basis with, if possible, readily understandable reasons, which are not likely to contain abstruse mathematical calculations. These are required for record purposes and should be given in writing as part of the report.

The method of presentation of the report will, of course, vary with the nature of the problem. If the results are quite clear-cut, then it is usually sufficient to start the report with a brief statement of the problem and the results achieved, then go on and give the data, the methods and the calculations in detail, but relegate to appendices any difficult mathematics and any explorations off the main line of research.

Very few writers can produce a draft of a technical document of any size and guarantee it to be free from errors, ambiguities, innuendoes (unintended, of course), or phrases which could have been better put. It is strongly recommended that a draft report be circulated to all interested parties asking for their assistance in eliminating all these faults. If, at the same time, the interested parties care to comment on the suggestions, this offers a first-class opportunity for bringing them into a form which can, and will, be accepted. It is much better when recommending a line of policy to persuade someone to go half-way in the right direction than to ask for the whole and get nothing.

The circulation of a draft has other benefits also. If the recommendations are obviously reasonable and sensible, it may be that by the time the draft report has been put into its final form and published, the greater part of the recommendations will have been implemented. Good advice cannot be put into effect too soon. Minor changes that are sensible (and some major ones too) hardly need the sanction of a formally presented final report before action is taken. The question may then be asked as to why in these circumstances there is any need to publish a final report. There seem to be three reasons for this : the first is that an agreed and finished document is a record of the proceedings which can be readily

referred to when any question of future action arises. It may be a question of action in a field not specifically dealt with in detail, but which nevertheless has been studied to such an extent as to form a valuable starting point for fresh work.

A second reason is historic. It is not likely that the results of an investigation will be valid indefinitely. It is, on the contrary, very likely indeed that the work will either have to be repeated or extended in view of new circumstances arising in the activity studied. A most valuable starting point for such a fresh study is the completed version of the previous study. The draft will not do. One of the major purposes of circulating a draft is to free it from error, mis-statement, etc., and to start a fresh investigation from a draft may well take more time than starting again from scratch. The finished document, on the other hand, should be a satisfactory starting point.

The third reason for a finished document is educational. It is to the advantage of the authors of the document to prepare a finished draft, in the sense that it is in many ways a completion of their efforts as authors, and the document, if of good quality, is likely to be useful as a guide to others engaged in similar activities even if its virtues are mainly in the field of " don't do it this way ". But the approach to a problem given by one set of authors is always of interest to the other workers in the same field. Also, in some circumstances, such a document may receive wide circulation in the firm concerned for the general information of members of the staff, who are not directly concerned with the problems in hand.

The emphasis just placed on drafting and completing a report does not imply that presentation of a draft or even a final report completes a project.

Acceptance of Recommendations

Presentation of recommendations can, and should, start from the draft report. This will no doubt be a basis for meetings at which the draft is discussed in detail and in overall content. Even before that, there will have been discussions between the project personnel and interested executives. Recommending action on these preliminaries, however, involves risk of sub-optimisation or even change of view as more information is analysed.

Two case examples are useful. The first leads to a readily implemented conclusion, the second to a less clear cut situation.

The case readily implemented was a study of three factories making the same basic product (or rather range of products) in slightly different mixtures and at different costs. The study in first draft showed clearly why the costs were different. The explanation was simple—differences of a minor nature in the processes, different scales of staffing and amenities and differences in overall utilisation of facilities. The study went on to consider making the products in a lesser number of factories, i.e. could any one or two of the three factories be, with advantage, closed, with a corresponding increase in output at the factory or factories retained? The draft showed that after allowing for savings and expenditures to bring all factories to the same level of process effectiveness, staffing and amenities, there would be savings if any one of the factories was closed and the others expanded. One factory, however, would be difficult to expand for space reasons and, in any event, expansion of output would involve heavy capital expenditure. Its current output (smallest of the three) could be easily absorbed between the other two with little capital expenditure. Accordingly it seemed correct to recommend closing this smallest factory, despite its lowest operating costs (it was the middle factory on total profit).

This draft report was discussed at great length with the director concerned. Supplementary questions—labour absorption and saleability of site—were satisfactorily answered. A condensed statement was then prepared for the company board and, in due course, action was agreed.

The confidential nature of this recommendation prevented normal discussions other than between the writer of the report and his team, the writer and the director, and both with an external consultant.

No difficulties were in fact encountered on implementation. Labour was absorbed without difficulty and other appointments found for staff. Many, however, decided to find new posts rather than move home.

The second case was much more difficult. The recommendations were partly changes in policy, partly changes in administrative methods and partly engineering proposals.

Furthermore, the amount of data accumulated and analysed could not be readily compressed into a brief document. Indeed, the full record of the work was a volume comparable in size with this book. Fortunately it was possible to issue a series of documents, each being a factual account of a particular part of the plant being studied. These documents were passed round to all parties with an explanation that they were drafts and, although every effort had been made to make them accurate, help in removing mistakes and filling in omissions would be gratefully accepted. This help was readily given by all concerned. When the very bulky factual account had been assembled and all the points raised, and after comment had been met and incorporated, the whole was printed as a work of reference. The material was then used to prepare as short an account as possible of the deductions that could be made from the data and suggestions for eliminating the cause of concern, which had led to the investigation being started.

The most important conclusions were that the balance of plant was not satisfactory, i.e. it was impossible to make proper utilisation of certain machines without working other machines very long hours or installing extra capacity. It was also noted that various different products passing through the plant required quite different times in total and that, in particular, it was quite unreasonable to assume the cost of putting all materials through a particular process was the same. This led to a recommendation for modifications in the cost accounting system, and as these were accepted readily by all concerned no formal implementation was necessary—a most satisfactory state of affairs.

It is always desirable for a recommendation to be accepted by the person to whom it is addressed as a sensible course of action, and, if it is within the competence of the person concerned to take that action, then he should do so without being further instructed. This sort of amicable arrangement between sensible people, who all have the same object in view (to improve whatever operations are under their control), is a very desirable method of implementation.

The policy recommendation to remove certain classes of raw materials from the plant involved discontinuing particular products in one case, processing elsewhere in another case and,

in a third case, the products had to be retained pending transfer to a plant which was in course of construction. Here it was agreed, as a matter of policy, that small quantities of a particular product were most undesirable in a plant equipped for handling large-scale production and, in the case of some small-scale products, the remedy adopted was to try to bring the production up to a level of worthwhile throughput, but to agree that if this could not be done, the product should be discontinued (this was not likely to cause customer concern because none of the products was the only option for its particular end use).

So far recommendations discussed in open meeting were dealt with in the process of going through the details of the report at a series of meetings with interested parties. Recommendations in the form of other administrative changes were not put in the form of a ready-made solution, but in suggestions to improve particular aspects of administration. Such proposals are easy to accept—and in fact were accepted—but they may necessitate a fresh study of prolonged duration. It is a philosophical point to decide whether the accepting of a recommendation is the implementation or whether implementation actually occurs when the further studies have themselves been carried out. If the latter is true a situation can arise that, in a sense, a recommendation is never implemented in full, but a series of continuing improvements occur.

Recommendations for modifications and design of plant are simple to accept and sometimes simple to carry out. In any event they are best left in the hands of the appropriate technologists. Proposals in broad outline should always be preferred to detailed recommendations from a general study to a particular study. For instance, a suggestion that " The throughput of a particular machine must be increased : efforts should be made to run it at higher speed " is sensible. It is not, however, in general sensible to suggest that a (say) 300-h.p. motor be replaced by a (say) 500-h.p. motor. If the proposal for greater speeds is sensible, then—as a corollary—a study of bearings, gear drives and motors will follow. This should always be left in the hands of the appropriate designer. Indeed, as a general point it is better to recommend a course of action by specifying an objective than to recommend a

detailed course of action with insufficient knowledge or detailed investigation.

Oral Presentation

It may be practicable, particularly in the military field and in concerns which tend to have daily, or at any rate, frequent meetings for briefing purposes, to make use of such meetings to present the results of Operational Research. This is a limited technique, for in order to ensure that an oral presentation is successful, it must be brief and to the point. The problem, therefore, must be one which is readily enunciated and appreciated ; the solution must be simple and not only must it be correct in a mathematical sense, but it must be reasonable in a management sense. The conclusions reached must be in a form suitable for action to be decided at a conference. One case is where Operational Research has shown that the existing policies or methods are sufficiently good to make the upset caused by change likely to be more costly than the likely benefits. The decision to be taken is " no change ". Other instances are those where Operational Research results in a proposal to change the number of a particular type of facility ; a simple case is a recommendation to change two loading bays to three loading bays, or to show that the number of locomotives on a siding, or the number of wagons for storage purposes, should be altered. These are matters which could be presented with proper visual aids in the form of wall diagrams or slides, or perhaps, an epidiascope. There may be other occasions when it is desired to exhibit evidence in an attractive form— the use of films here is a possibility. These are particularly suitable for demonstrating certain types of congestion, both of human and wheeled traffic. Even in these cases, however, it is essential that a proper record of the work be kept to permit reference back on future occasions.

Implementation

So far in this chapter implementation has been used in the sense of agreeing to accept the results of Operational Research and to put the recommendations of research into practice. It, also, rather implies that once the recommendations have been made the problem is put into a neat pigeonhole and forgotten

by all except the man who is taking the advice prescribed. Such circumstances do, of course, occur. If the Operational Research project has been on a much larger scale and results in recommendations for major changes in organisation, it is very unlikely that these will be worked out to the last detail by an Operational Research team. It is much more likely that a separate implementing team will have to be set up to plan the method in which the changes are to be made, rather than make the actual changes. This team will need as a member a representative of the Operational Research team—and it may need more than one representative for, in the course of detailed planning, it may be necessary to do additional minor Operational Research studies. As an example, consider a stock-holding investigation which is faced with the situation of quite excessive stocks of a commodity. The Operational Research investigation may well show that stocks should be allowed to fluctuate between two much lower figures and only very rarely pass beyond these limits. Making a reduction in stocks of raw materials is comparatively easy. This is done by reducing purchasing orders unless, of course, purchasing contracts are such that alteration is difficult. In these circumstances implementation of recommendations for reduction of stocks of raw materials will have to wait until necessary changes in terms of purchase are made. After that, smaller quantities can be purchased until stocks are run down. These manœuvres may necessitate the supplier running up his stocks or cutting back his production.

Running down stocks of finished products—or indeed building them up—as a result of an Operational Research investigation can be a very difficult process, particularly when it is combined with studies of optimum campaign length or optimum purchasing quantity. In an expanding market, it seems that minimum dislocation will be obtained by holding production steady and letting the increase in business take care of excessive stocks. In a stable market, an approach is to try and sell excess stocks, if necessary at cut prices, outside the company's normal customers. Selling them to the company's normal customers is obviously not advantageous, for it is bound to end eventually in a cut-back in production. Dumped exports is a form of cut-price selling.

Increase in stocks to a desired level may be even more difficult. The obvious way of obtaining increased stocks is to increase output without, at the same time, increasing sales. This involves an increase either in output per hour, which may be very difficult, or in hours of work, which is impossible for continuously operating plant, or in campaign length with consequent saving in either set-up time or cleaning and other preparation time. The most feasible of these is to increase the campaign length, but this is only possible if there is some stock of some commodity available.

Implementation in the sense of these examples can be much more difficult than the calculations leading to recommendations as to what levels stocks ought to be. They are nevertheless considerations that Operational Research workers and users of Operational Research must bear in mind from an early stage in a project.

After all, they are sufficiently obvious for someone in the firm to have noticed them ! What is not so obvious is the cause. One of the most difficult things in any Operational Research is to check back from symptoms to dig out underlying causes and try and get some idea as to the cause of the situation which is arousing disquiet. That is one part of the problem.

The next part is, having got these ideas, to go into the field of mathematics. This involves picking out the important parameters ; in other words, having got some ideas in general terms, try and be more specific and isolate the various parts of a problem which can be isolated ; after that, start putting numbers in. This may sound rather nebulous. The suggestion is, however, that if somebody can give a nice, neat, tidy problem that can be solved at once by the application of a little mathematics or statistics, it is in the highest degree unlikely that it is a case for Operational Research ; rather it is a case for statistics or a little arithmetic. No criticism of statistics is implied here—far from it, but the basic point about Operational Research is that it is research, and therefore involves finding out things that people do not know about and one of the things that people do not in general know is exactly what form of problem they are faced with. Operational Research has to find this out.

Formulation when the Problem has not yet been recognised

If O.R. is introduced into a firm where there are no obvious problems, a quite different method of approach to formulation will be needed. It may be that someone is disturbed by the sight of a warehouse full of raw materials that stand no chance whatever of being used ; he may say : " We would like something done about these stocks, they must be got down." On the face of it, that sounds like a very innocent problem which could be solved by the application of a little arithmetic or statistics. It is not so simple, however : one has to find out how those stocks got there, what levels of stocks might be desirable, and, eventually, it will be necessary to develop a stocking policy to try and ensure that, come what may, the dangers of accumulating such stocks can be avoided in future.

On the other hand, it may be decided to make a survey of the activities of a firm with a view to finding out what sort of

problems there are in the firm. A general survey of a plant or activity should consider the policies implicit in its organisation and predict the consequences of changes. Such a survey may be done with little more intention than trying out the methods of Operational Research in a fresh field. There may of course be excellent reasons for initiating a survey, the simplest case being a policy decision of a few years past having led, when implemented, to somewhat disappointing results.

A very useful starting point in a general survey is to ask a few questions as to just how much money is invested in various parts of the firm. For instance, it is as well to find out roughly what the buildings are worth, what the capital value of the plant is, and it is advisable to make some enquiries about methods of writing off. It is a thoroughly good idea to enquire how those values compare with the working capital, how much money is locked up in raw-material stocks, how much in work in progress, how much in goods awaiting delivery. Incidentally it is advisable to differentiate here between goods awaiting delivery and goods that have been made and are awaiting sale.

Having got that sort of information, it is not very difficult to find out whether any of those particular things is wrapping up a very large part of the firm's assets. If they are, one can hardly go wrong in the first place in making some tentative proposals to see what can be done to improve the utilisation of the most valuable assets of the firm. (One may perhaps get into certain difficulties if it is discovered that the most valuable assets are brains, which cannot be priced.) Having dealt with this tentative field of enquiry, attention can be given to the major heads of expense. If it appears that the main head of expense is the purchase of raw materials, then a detailed study of raw materials would seem to be indicated ; if there is a lot of money wrapped up in finished stocks, then these can be looked at in more detail ; if a major heading of expense is found to be loss from obsolete goods in warehouses, then clearly further studies should be directed here. It may be, of course, that in certain types of process the major heads of expense are labour, while in other circumstances perhaps services or maintenance costs are the most expensive item. In any case, enquiries of this sort will always show a field where investigation is likely to be profitable. There are

a number of other approaches, of course, and one virtually infallible guide, in the stock-holding field, is that if there is a policy to stock all goods by the same rule, such as : " Carry three months' estimated consumption " and there are a variety of goods with different rates of consumption, then it is safe to assume that the fixed rule is wrong. (There is a possible case where it might not be all that wrong and that is when work is on an assembly line, but even then it is pretty certain to be wrong.) Unless all commodities are being used at the same rate, with the same fluctuations, it is most unlikely that the same level of stocks will be valid. The same type of question can be applied to finished goods too, where relevant. If the aim is to have so many months' estimated sales in stock, then—except in the very unlikely circumstances of all selling at the same rate with the same type of fluctuation—that stocking policy is wrong and can be investigated.

The most interesting type of problem, however, is that of trying to balance between, say, four heads of expense which are more or less equal. If, in a particular factory, where depreciation charges, maintenance charges, labour costs and heat, steam—or power in one form or another—are all about equal, someone has set his mind on being extremely economical with one of these four things, then it is more than likely that these economies are being secured at the cost of wastages in the other resources. This should not be construed as a condemnation of economy ; the point is that it is not unknown, in the field of Operational Research particularly, to discover a situation where one department is quite unintentionally, quite fortuitously, making itself efficient at the cost of the firm's efficiency. For example, there is a factory where each process in the factory—and raw material goes through a string of processes—is carried out quite reasonably economically until the final stages of production are reached, and here machines are being grossly underworked. In the search for bottlenecks it was found that the various intermediate machines all work at different rates (which is not surprising), the goods do not flow through every machine, they wander about. In such circumstances it is not really surprising that the throughputs do not balance. It may be that, working back, one discovers perhaps one machine through which nearly everything goes. It may

be that by running that process extremely economically, machines at later processes are starved ; possibly, doubling the capacity of the key machine by putting in a new modern and highly expensive machine may quite successfully double or quadruple that process cost. It may be well worth while to replace an ancient machine on which there is no depreciation or maintenance, since, for the cost of increasing unit process cost at one stage, it is possible to double the flow-through, which will mean better utilisation of the majority of machines.

This is a very typical situation. In most Operational Research work there is a need to try and find a balance between various contending factors, and the contending factors generally have the pattern that both too little and too much are expensive. In a cement plant, for instance, there was one loading machine which could, in the course of the day, put on to the lorries the entire output of the plant, which was very satisfactory. The only snag about it was that the lorries all came at the same time in the morning, so some poor unfortunate lorry-drivers had the dubious privilege of waiting from, say, 8 o'clock to 5 o'clock before being loaded up. Now it would have been much more economical to have had two or three loading bays and got rid of the lorry queue. The whole art, or science, of formulating a problem in the O.R. field is to get the factors isolated, to find out which are the causes of increased expenditure as a factor is increased, or decreased, and the availability of that factor.

It is only fair to say at this stage that there may be quite different boundary conditions in a problem, according as to whether one is considering the interests of a department, of a factory, of a group of factories, or of a group of firms. Whatever the background, in trying to track down cause and effect it is very necessary to make quite sure that the problem is being correctly described.

The People who make an O.R. Study

At this point it would perhaps be as well to consider the people involved in an Operational Research investigation. The type of problems so far mentioned are quite often beyond the capacity of one person to deal with. It is clear that various

types of knowledge will be needed in the different aspects of a survey. It has already been suggested that information about costs will be useful and this calls for some knowledge of accountancy. Again, production data are sure to be needed and this too calls for detailed knowledge. (Incidentally, it is worth bearing in mind that at each level in a firm there will be a quite different standard of knowledge of performance ; the main board, the factory manager, the floor supervisor and the man who actually did the job will all have very different views as to what was done.) Where the problem is a large one it is an unwise economy to let one person try and sort it out for solution. A group of people is needed, preferably people who know different things—in other words, people with different training, different background, different experience.

In any team going deeply into a large subject it is likely that accounting, technological and statistical knowledge will be required. It may be that the nature of the technology is such that, in addition to the technologist's knowledge of the process or plant being studied, engineering, physical or chemical knowledge will also be needed in the team. It is not to be expected that the person responsible for the investigation will be knowledgeable in all these fields, although clearly he must have acquaintance with them.

Much current research work is carried out by teams of men of many talents. Frequently it appears that there is no one person in overall control of day-to-day work. This may work very well in fundamental research, but it is not advocated in Operational Research. There are many reasons for this, the most compelling being economy in time and effort, liaison with the organisation studied and presentation of results. Economy in time and effort necessitates rejecting seductive backwaters or enticing leads ; it also means proceeding no further than is required for the purpose in hand. Liaison with the organisation studied is discussed later under communication. Presentation of results (see Chapter XIV—Implementation) needs authority to ensure consistency in style and layout.

These considerations lead to the suggestion of a team with a variety of talents under the guidance of one senior man. At this stage thought has to be given to communication and responsibility. Obviously it is not proposed that all the

personnel in the team—or even a majority of them—should bear direct responsibility to the immediate head of the activity investigated. It is equally unsatisfactory for the man in charge of the investigation to be too closely linked with the activity studied, but if this is unavoidable, as it may be, then it is quite essential that some members of the investigating team at least should be independent, if only to say things which might be injudicious from somebody too near to the work. It is, however, undesirable that the team should all be independent of the activity studied.

A team should be composed partly of staff seconded from the activity considered and partly from an Operational Research department. Seconded staff must be seconded full-time, or they will find that their normal duties continue to occupy most of their time. This secondment is of value for three reasons : it gives ready access to all available data (not always easily accessible to comparative strangers) and it facilitates implementation (see Chapter XIV). Such a team may well work all together for six months and the whole project, from initiation to decision for implementation, take a year.

In the early stages of understanding and problem formulation, only the responsible person is needed. As soon as the problem is agreed in form, technical work can start ; once this is in hand and information becomes available then costing studies can start. This implies a two-stage build-up of a team. When technical work is completed, then report drafting begins. As each individual's contribution reaches a draft stage he can be released part- or full-time to other duties. These may be minor problems of a related or separate nature, or information gathering. Keeping up to date on recent technical developments is not very easy when working at full stretch on a specific and urgent project. Seconded personnel, of course, revert to their normal duties. Eventually there will be only a nucleus left of, say, one or two people who are charged with a final check of the work (see Chapter XIV).

Need for Communication

Communications between members of the investigating team can readily be ensured by arranging that they can all work in close proximity. It is very undesirable that during

the investigation any of them should have other duties, or indeed that they should retain their normal offices. They will also require communication at the appropriate levels to all sources of relevant information in the organisation studied, and to such external sources as may be necessary.

Investigating parties will also need direct contact with the head of the organisation and with the authority responsible for initiating the work. If the work is being done at the suggestion of a member of a board and is studying a factory under the jurisdiction of that board, then whoever is in charge of the investigation must have direct access to the factory manager and to the board concerned.

It can be argued that the suggestions for formulation made so far have been rather too general in nature. Surely all factories are different and each problem needs individual treatment. In a sense this is true, but the underlying relationships between cause and effect are likely to be similar over a wide variety of places, and this is one reason why it is possible to take a team of people who may never have seen the particular firm to be studied and let them loose with great hope of success. It is essential to be absolutely certain as to the object of the exercise ; and it is necessary also to describe the problem on paper in quantitative terms and make sure that all concerned in the supervision and direction of the unit studied agree that the qualitative terms are right.

Cost and Value of an O.R. Study

One important question which must be asked about any problem is whether the solution is likely to be worth while. This in turn is linked with the number of people that one might put on to a particular problem. Taking four very senior personnel out of their jobs for a period of a year might well cost a firm something like £15,000 to £20,000 by the time all the appropriate overheads have been taken into account. If a firm is thinking of setting out an O.R. problem for four people of that calibre to tackle, it is just as well to ensure in advance that there is some prospect of seeing a fair return on the money. Now it is quite true that in some fields Operational Research can produce quite spectacular results : in stock-control problems it is not in the least degree unknown to

offer a change by dividing stocks by a factor of three. Now, supposing it has been decided, after a preliminary survey of a problem, that there is work for about four people for a year, say, a bill of £15,000 to £20,000. What size problem does it have to be in the line of financial rewards to make it worth while? In the field of stock control, that team would be worth putting on to about £½ m. worth of stock—which will give a tenfold return on the research work. If, on the other hand, the stocking problem is such that one not too senior person with the requisite knowledge can solve it in a month (the bill for that might be £200 or so), then something in the order of, say, £10,000 worth of stock sounds like a sufficient proposition. Similarly with queueing problems—a recurrent theme in Operational Research—one has to decide whether possible relieving of the queue is as beneficial as the work is costly. It may be of course that cut-throat methods such as looking at a bottleneck and saying: "Well, if you double that in size, of course you won't have a bottleneck any more" is a solution. If a machine worth, say, £1,000 is causing a bottleneck, it is not very sensible to put £5,000 of work in before deciding that a second machine would be a good idea.

Case-Histories

The three case-histories which follow are nice illustrations of occasions where a rather nebulous problem has been taken and studied in detail by a considerable group of people, and the results are given here briefly.

Study of a Group of Factories

The first was a study of a group of factories[1] where about four people—a works manager, a deputy works manager, a cost accountant and an O.R. man—were put on to the job for about four months; the total bill was about £4,000. Actually, only one of the four people was engaged over the whole four months, the others came and went as required. The problem posed was: "Here is a group of factories which are only just running at a very small profit; the investment in the factories is not worth while. Why? and what is to be done about it?" *

* This problem is discussed, from a different angle, in Chapter XIV.

Now that sounds like a reasonable problem in Operational Research ; it is a very clear statement of what the problem is, but it is not in the least degree a statement of a problem where an O.R. man can go and solve it. In this particular case, the formulation of the problem took the form of saying : " All right, what do the factories make ? What does it cost to make it ? " Of course, the question that was really asked was : " How do the costs in the various factories fluctuate according to the amount of throughput in those factories ? " The situation, by the way, was further complicated by the fact that the capital investments in the factories were very different, partly because the factories varied in size.

Information was collected on the capacity, condition of plant and buildings, utilisation, throughput and profits of each factory. (Results are charted in Table 16.1. The numbers given

TABLE 16.1

Factory	*Capacity*	*Utilisation*	*Plant and Buildings*	*Throughput*	*Profits*
A	Large 3	50%	Good	1	4½%
B	Medium 2	75%	Tolerable	1½	5%
C	Small 1	100%	Bad	1½	15%

are not the actual figures, but they do give a realistic picture of the problem—in other words, the distortion is that of a modern artist and not that of a cartoonist !) This is the sort of basic information that a diagnostician would want in tackling a problem of this nature.

Having reached this stage, knowing what was in fact happening, the obvious question was " What can be done about it ? " Now this is a prime characteristic of good Operational Research. It is no use saying at this stage : " We have discovered the following things, they are all true " if you are not prepared for someone to retort : " How very interesting, what do you suggest we do about it ? " If one is trying to study relations between various parameters that can be expressed in numerical form it is absolutely essential to know what can be done about it. In this particular case, there are three courses of action that one could take in these factories. Clearly, a

factory can be shut down—as there is ample capacity. Another alternative is to do nothing ; that is what is known as null decision. Thirdly, a factory can be expanded. There are other possibilities, but in this question of formulating a problem one is looking for likely courses of action, for the object is to try and solve the problem and this means that proposals should be realistic. The alternatives Close, Expand, Do nothing, gave nine possible courses of action, of which some can speedily be rejected as unfeasible. The short list was then studied in rather more detail. What had to be done in this case was to examine production very closely. All feasible solutions were put down from a technical point of view and the team then went round and eliminated a few more. It transpired that factory B could perfectly well be shut, because there was a gross labour shortage in the area ; shutting factory C was less attractive from this point of view as there was very little alternative employment. One should always remember that in this type of work all calculations are approximations. Because of this, two possible solutions to this problem were said to be indistinguishable. The final recommendations still left room for further discussion, as it was suggested that either factory B or factory C be closed and the other two extended. In fact, factory C was closed a year later. As a result of this study savings of up to £$\frac{1}{3}$ m. were achieved ; some of this represented improvements in production techniques which had been made possible by the inter-factory comparison. The saving was not obtained for many years, because the particular product is being greatly changed in its nature and made by a totally new process, and in fact the ultimate probability is that all factories will be shut and there will be one small plant in a quite different place making this product. Nevertheless, it is very likely that the two factories left will then be turned over to other things. But in any case £$\frac{1}{3}$ m. return on £4,000—the cost of the investigation —seems more than satisfactory.

The Lengthening Order Book

Another study started off from the fact that a particular factory was getting further and further behind with its orders, while the volume of output was not increasing at all. The

question put was : " Is the factory's plea for about a 25% increase in their plant justified ? " It soon became clear that the two basic pieces of data were perfectly correct : the order book was lengthening and the proportion of orders that became overdue was getting higher—and that was not due to over-optimistic sales promises, either, although it might well have been. Investigation showed that the root of the trouble was that the product was getting more and more complicated every day, and in fact it was getting complicated at such a rate that the production director concerned had not been able to keep up with it. The statement on his desk, saying how many varieties there were, was out of date by the time it was printed, new varieties were going in so rapidly. There was also a slight tendency for orders to become more erratic. The answer to that of course is some stock-holding. That is an extremely interesting example of the importance of formulation in an O.R. problem. The fundamental question was : " Why is the factory getting behind ? " and the answer was : " Because the variety of the products is increasing far more rapidly than the productive efficiency of the factory "—which was increasing, by the way, very readily. In other words, the range of raw materials was ever-increasing—and therefore so was the money tied up in raw materials ; the range of finished products was ever-increasing—and therefore so was the money tied up in finished products ; the production problem was getting more and more complicated every day and it was hardly surprising that the factory was falling behind. The remedy is a certain amount of more plant, a certain amount of product simplification and a certain amount of persuasion of the customer that he could be satisfied by a choice from a small standard range of products and did not need something made exactly to his specifications.

Reduction of Waste

The last case is a charming instance. There is a lot to be said for going round a place and asking questions about where the money is being expended to an extent that seems out of proportion, or where there are instances of waste. This factory was effectively one which made a product in a strip that was then cut up. Someone had reason to be dissatisfied with

the realisation of good material from the cut strip, and a study was made of the method of cutting up. The realisation had been about, say, 83% good product out of the good finished work; a change in the method of cutting pushed that up by 10% and led to an annual saving of £25,000. The cost of the study was only about £2,500, since it had merely involved a man and a boy spending three months doing arithmetic, playing games on the cutting-up process.

Establishing O.R. in a Firm

Once Operational Research has proved itself in a firm, it may well be that management wish to see it established on a permanent basis, and some ideas on the requirements of an O.R. department may be useful. Initial staff of an O.R. department must be sufficient to half man a project. No specific size of team has been recommended. Teams of three have been highly successful, one of each of three disciplines with one as a leader. Teams of up to about six at a time are readily manageable, but a project demanding more should be split into sub-projects, each with its own team. If a team of up to six is envisaged, this implies an initial technical staff of three plus secretarial and computing aids. This is a minimum. As work develops it is only necessary to legislate for the number of major projects to be studied at one time and create a department *pro rata* with the addition of one manager (but see Chapter XIV).

The number of major investigations running at one time can be assessed on a turnover basis: a factory with a £1,000,000 annual turnover may require a major study every five years with a certain amount of intermittent help. This assumes a five-man team with two permanent men, and a study lasting in all about a year. The equivalent load is therefore $\frac{2}{5}$ man per £1,000,000 turnover for studies, or, after allowing for odd jobs, say half a man. It would seem, therefore, that one permanent man per £2,000,000 annual turnover is about right.

This may be compared with some American information.[2] There it is suggested that a nine-man group (permanent) is about right for an annual turnover of £100,000,000 per annum plus, but one permanent man for every £3,000,000

to £6,000,000 per annum. This is about half the rate suggested for British industry. The feature of slower growth than *pro rata* to gross income is very interesting. It may well be correct that doubling the cash value of a problem does not double its difficulty. However, there is some agreement that firms with a turnover of about £2,000,000 a year might well find it advantageous to employ a full-time Operational Research man.

In a medium-sized organisation it may be economical to combine Operational Research with some other function for accommodation, administration and services, but the principle of direct access must still be applied. (A combined Operational Research, Organisation and Method and Work Study department may have much to commend it : see Chapter XV.) Small organisations will usually find it desirable to use the services of an Operational Research consultant rather than employ permanent staff, but there is a growing belief that even large firms may find O.R. on an external consultant basis more economical than on an internal basis. The reason is that the extra cost (which is only significant for the first man in a section) is more than recovered by a higher rate of progress obtained by more experienced direction and a higher degree of specialisation. After all, a section of, say, three or four men cannot be expected to deal expeditiously with all types of problem. If these men can be changed at need from a larger pool then prospects are much better.

REFERENCES

[1] HEPBURN, JOHN H. (1958). " Operational Research Study of Manufacture at a Group of Factories ", *Operat. Res. Quart.*, **9**, No. 2, June.

[2] " Operations Research— A Basic Approach ", *American Management Association Special Report*, No. 13, New York, 1956 (p. 12).

INDEX

Abeele, 42
Accountancy, 301
Ackoff, 114, 117, 127, 311
Activity ratio, 19
Activity sampling, 227
Aitken, 235
Allcock, 237
Analogue computers, 119
Analogue simulation, 121 *et seq.*
Analysis of Variance, 141
Andress, 43
Archimedes, 213
Argyle, 43
Arnoff, 114, 117, 127, 311
Ashcroft, 79, 94

Bailey, N. T., J. 74, 80, 81, 94
Bailey, P., 115, 117,
Barrer, 94
Batch size, 96
B.B.S.A.T.R.A., 36, 39
B.C.I.R.A., 31
Beeching, 38, 42
Beer, 33, 42, 158
Bell, 220
Bellman, 239, 257
Benson, 48, 93
Bergman, 43
Bidding, 250
B.I.S.R.A., 22, 49
Black box, 137
Blackett, 29, 220
Bowley, 237
Bowman, 127, 200
British Association, 213
Brockmeyer, 93, 220
Bronowski, 29
Brown, 117
Buffer stock, 96, 102
Building Research Station, 39
Bureau of Labour Statistics, 36
Byrt, 211

Capital equipment, 39
Carnot, 213
Case Institute, 117
C.E.G.B., 162
Charnes, 179
Churchman, 114, 117, 127, 311
C.I.O.F.F.I., 43
Coal, 34
Cobham, 94
Coefficient of correction, 138
Cohen, 94
Cole, 42
Communications engineer, 151
Computers, 15, 21, 23, 24, 88, 102, 149
Conant, 29
Congestion, 44 *et seq.*
Consumer research, 14, 227
Control systems, 128 *et seq.*
Cook, 211
Cooper, 42, 179
Coppinger, 127
Cost accountancy, 39
Cotton industry, 34
Cotton spinning, 31, 39
Count of the yarn, 34
Cox, 77, 81, 93, 94
Crowther, 220
Curvilinear regression, 37
Cybernetics, 14, 15, 130, 150
Cyclic queues, 84
Cyclic system, 47

Data handling, 222 *et seq.*
Data processing, 229
Deterministic processes, 246
Difference equations, 58, 234
Digital computers, 119
Distribution, 180 *et seq.*
Doig, 93
Downton, 94
Dreyfus, 257
Duckworth, 163, 178
Dynamic programming, 235, 239 *et seq.*

Easterfield, 43
Eddison, 29, 93
Efficiency, 30 *et seq.*
E.O.Q., 104

E.P.A., 31
Erlang, 47, 69
Exponential smoothing, 100
Extrapolation, 233

Feasible solution, 161
Feedback, 128, 144, 149
Fetter, 127
Fire power, 213
Fisher, 127
Flying hours, 34
Forecasting, 131 *et seq.*, 144
Fornallaz, 42
Fournier analysis, 233

Gadsby, 221
Gamma distribution, 57
Gardner, 43
Gass, 178, 179, 198, 200
G.E.C., 40
Goode, 221
Gregory's formula, 234
Gregory-Newton, 234
Guest, 210

Haley, 178
Halstrom, 93, 220
Hammersley, 42, 311
Harbury, 40, 43
Hawthorne, 19, 40
Henderson, 178, 179
Hepburn, 117, 325
Hill, 210
Hirsch, 43
Holland, 42
Human problems, 201, 258 *et seq.*
Huxley, 27, 29
Hyperexponential, 64

I.C.W.A., 35
Implementation of results, 289 *et seq.*
Incentive payment, 33
Industrial economies, 41
Industrial engineering, 301
Integral equations, 58
Interfirm comparison, 35
Interpolation, 233
Inventory, 46
I.P.E., 35

Jackson, 93
Jaques, 210
Jensen, 93, 220
Jones, 37, 42, 237
Jury problems, 235

Kahn, 43, 127
Katz, 43
Kay, 163, 178
Kendall, D. G., 74, 93, 94
Kinniburgh, 35, 42
Kimball, 220
Koenigsberg, 84, 93
Koopmans, 178
Kopal, 238
Korn, 125

Labour turnover, 201 *et seq.*
Labour wastage, 201 *et seq.*
Lanchester, 220, 213
Langerburg, 38, 42
Lead time, 96
Least squares, 235
Lindley, 72, 93
Linear difference equations, 234
Linear programming, 159 *et seq.*, 180
Loss coefficient, 47
Lost sales, 47
Luttrell, 41, 43

Machine speeds, 39
Machol, 221
Magee, 117
Marching problems, 235
Mark, 40, 43
Market surveys, 227
Markov Chains, 58
Marshall, 127
Martin, 42
Matthew, 311
McCloskey, 127
Meisling, 93
Metrology, 222
Meyer, 127
Military applications, 212 *et seq.*
Milne, 237
Milne's method, 235
Milne-Thomson, 237
Monge, 213
Monte Carlo, 23, 49
Monte Carlo method, 121 *et seq.*
Morale, 201
Moran, 117

Morse, 70, 93, 219, 220, 221
Morton, 163, 178
Muir, 100, 117
Multiple channels, 65
Multiple correlation, 37

Naor, 77, 94
Negative exponential, 57
Newton-Raphson, 235
Nomograms, 235
North-West corner rule, 166
Numerical analysis, 228
Numerical integration, 234

O and M, 301, 325
O.E.E.C., 32
Optimality principle, 243
Ore handling, 50
Owen, 29, 93
Oxford Institute of Statistics, 39

Palm, 94
Periodograms, 233
Personnel selection, 301
Phipps, 94
Pitt, 220
Planning, 301
Poidevin, 210
Poisson, 58, 61, 83, 101
Posner, 42
Potentiometers, 149
Pre-emptive priority, 82
Presentation of results, 288 *et seq.*
Processing time, 44
Production planning, 33, 144
Production statistics, 224
Productivity, 30 *et seq.*, 39, 40, 41

Quadratic programme, 188
Quality control, 142
Queue discipline, 45
Queueing, 22, 44 *et seq.*, 234
Queueing theory, 154, 214

Radar, 25, 213
R.A.F., 213
Rand Corporation, 127
Ratio delay, 19
Recruiting labour, 209
Redish, 238
Redundancy, 202
Regression Analysis, 20, 21, 41
Reiners, 43
Replacement, 234, 251
Report writing, 290
Rice, 210
Richardson, 237
Ridgion, 237
Riley, 179
Ring spinning, 33
Robinson, 237
Rodgers, 34
Roethlisberger, 29
Rogers, 42
Rosander, 178
Rudd, 43

Saaty, 81, 83, 93, 198, 200
Sasieni, 251, 257
Schlaifer, 178
Service, 44
Service channel, 61
Service time, 57, 59
Servo mechanism, 24
Shoe industry, 33
Shoes, 36
Simplex method, 162
Simpson's rule, 234
Simulation, 118 *et seq.*, 224
Slack variables, 170
Smith, 38, 42
Snap method, 48
Stability, 156
Statistical control, 142
Steady state, 81, 88
Steel, 34
Stephan, 94
Stochastic processes, 247
Stock control, 95 *et seq.*, 214
Stocks, 46
Stock-out costs, 104
Stone, 43
Stringer, 178
Swan, 311
Systems engineering, 14, 216

Tarkowski, 286, 287
Telephone systems, 214
Telephones, 44
Textiles, 48
Thornthwaite, 21, 29
Time study, 35
Tippett, L. H. C., 19, 42
Tocher, 127
Traffic, 80
Transient probabilities, 81

Transition equations, 69
Transition probability, 58
Transportation method, 162, 182
Turnbull, 286, 287

University of Michigan, 40
University of Oxford, 40

Vajda, 178, 179
Vallance, 42
Vincent, 42

Wagner, 178
Warehouse problem, 97
Weapon systems, 215
Weiss, 220
Welch, 113, 117
Whitin, 116
Whiddington, 220
Whitehouse, 237
Whittaker, 237
Wiener, 117, 130, 158
Wiener-Hopf, 73
Wilkinson, 127, 178
Winsten, 74, 93
Work study, 14, 48, 132, 224, 301, 325

Yates, 127